

Evaluation of Information Access Systems in the Generative Era

by

Negar Arabzadehghahyazi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2025

© Negar Arabzadehghahyazi 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Jamie Callan
 Professor, Dept. of Computer Science
 Carnegie Mellon University

Supervisor: Charles L.A. Clarke
 Professor, Dept. of Computer Science
 University of Waterloo

Internal Member: Jimmy Lin
 Professor, Dept. of Computer Science
 University of Waterloo

Internal Member: Mark Smucker
 Professor, Dept. of Computer Science
 University of Waterloo

Internal-External Member: Lukasz Golab
 Professor, Dept. of Management Engineering
 University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This dissertation includes first-authored and peer-reviewed work that has appeared in conference proceedings and journal publications in the fields of information retrieval and natural language processing, including Information Retrieval Journal (IRJ), the ACM SIGIR Conference on Research and Development in Information Retrieval, European Conference on Information Retrieval (ECIR), the European Chapter of the Association for Computational Linguistics (EACL), and the Conference on Empirical Methods in Natural Language Processing (EMNLP).

The following is a list of the papers incorporated into this dissertation.

Portions of Chapter 3

- Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles L.A. Clarke. Shallow Pooling for Sparse Labels. Published in *Information Retrieval Journal*, 2022.
- Negar Arabzadeh and Charles L.A. Clarke. Fréchet Distance for Offline Evaluation of Information Retrieval Systems with Sparse Labels. Published in the *Proceedings of the 2023 European Chapter of the Association for Computational Linguistics (EACL)*.

Portions of Chapter 4

- Marwah Alaofi, Negar Arabzadeh, Charles L.A. Clarke, and Mark Sanderson. Generative Information Retrieval Evaluation. Published as a book chapter in *Information Access in the Era of Generative AI*. Co-lead author.
- Negar Arabzadeh and Charles L.A. Clarke. Benchmarking LLM-Based Relevance Judgment Methods. To appear in *Proceedings of the 2025 ACM SIGIR Conference*.
- Negar Arabzadeh and Charles L.A. Clarke. A Human-AI Comparative Analysis of Prompt Sensitivity in LLM-Based Relevance Judgment. To appear in *Proceedings of the 2025 ACM SIGIR Conference*.

Portions of Chapter 5

- Negar Arabzadeh, Amin Bigdeli, and Charles L.A. Clarke. Adapting Standard Retrieval Benchmarks to Evaluate Generated Answers. Published in the *Proceedings of the 2024 European Conference on Information Retrieval (ECIR)*.
- Negar Arabzadeh, Fernando Diaz, and Junfeng He. Offline Evaluation of Set-Based Text-to-Image Generation. Published in the *Proceedings of the 2024 ACM SIGIR Asia-Pacific Conference*.
- Negar Arabzadeh, Siqing Huo, Nikhil Mehta, Qinqyun Wu, Chi Wang, Ahmed Awadallah, Charles L.A. Clarke, and Julia Kiseleva. Assessing and Verifying Task Utility in LLM-Powered Applications. Published in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

I hereby declare that while I was the lead author of the papers included in this thesis, the research they present would not have been possible without the contributions of my collaborators. I am especially grateful to my colleagues at the University of Waterloo, Google Brain, and Microsoft Research. In particular, I would like to thank Charles L.A. Clarke, Julia Kiseleva, Fernando Diaz, Amin Bigdeli, Xinyi Yan, Alexandra Vtyurina, Junfeng He, Marwah Alaofi, Mark Sanderson, Nikhil Mehta, Qinqyun Wu, Chi Wang, and Ahmed Awadallah for their invaluable support and insights throughout our collaborations.

I also used AI-based tools to improve the grammar and writing style of this thesis. All substantive content, ideas, and analysis are my own.

Abstract

The rapid advancement of information access technologies, including neural retrieval models and generative information-seeking systems, has outpaced traditional evaluation methodologies, exposing fundamental gaps in assessing their effectiveness. Existing evaluation frameworks struggle to adapt, particularly in the presence of sparse relevance labels, limiting their ability to fairly and comprehensively compare retrieval and generation-based systems. The emergence of large language models (LLMs) further complicates evaluation, as they challenge conventional assessment paradigms while offering new opportunities for automated evaluation. To address these issues, it is crucial to first identify flaws in current evaluation methodologies and then develop more robust, efficient, and adaptable assessment strategies.

This thesis begins by demonstrating that evaluation based on sparse labeling introduces substantial biases and inconsistencies in system rankings, often failing to recognize genuine improvements in retrieval effectiveness. We show that in traditional IR benchmarks, stronger models may retrieve highly relevant but unjudged documents, leading to underestimation of their performance. To mitigate this, we propose an alternative evaluation approach based on distribution of retrieved results and labeled data using Fréchet Distance. This method not only improves robustness in the presence of sparse labels but also facilitates direct comparison between retrieval-based and generative models on a common evaluation scale.

We then investigate how LLMs can be leveraged to evaluate IR systems, distinguishing between their use for evaluating retrieval-based methods and generative IR systems. A key focus of this work is the role of LLMs in automated relevance judgments. We systematically compare different LLM-based relevance assessment methodologies, highlighting the lack of standardization in evaluating these approaches. To address this gap, we propose a structured framework that evaluates relevance judgment methods based on their alignment with human labels and their impact on system rankings. Furthermore, we examine the effect of prompt formulation on LLM-based evaluation, demonstrating how prompt variations can significantly influence the consistency and reliability of assessment outcomes.

Finally, we extend our study beyond retrieval-based evaluation to assessing generated content across multiple applications. We explore retrieval-assisted methods for evaluating generative textual content, IR-inspired approaches for assessing text-to-image generation models, and a broader framework for evaluating LLM-powered applications. These contributions lay the foundation for a new generation of evaluation methodologies that keep pace with evolving information access technologies, ensuring that improvements in retrieval and generative AI systems can be accurately and meaningfully assessed.

Acknowledgements

I am deeply grateful to Professor Charles L. A. Clarke, my supervisor, for teaching me to be brave—to think boldly rather than incrementally, and never shy away from controversial ideas. His unwavering support for every decision I made, including the internships I pursued, has been extraordinary. His generous introductions to leading experts at conferences have shaped not only my research but also my confidence as a scholar. Our conversations have been a constant source of inspiration and joy.

My heartfelt thanks also go to Professor Ebrahim Bagheri, who supervised my master’s thesis, but who has been far more than a graduate supervisor. His mentorship has shown me that true supervision extends beyond research guidance—it encompasses genuine friendship, encouragement, and support in every aspect of life. I am truly honored to be his mentee.

I would like to acknowledge my colleagues in the Data Systems Group at the University of Waterloo—especially Xinyi Yan, Mahsa Seifikar, Amin Bigdeli and Siqing Huo—and the members of the LS3 Lab at Toronto Metropolitan University. Your team spirit, insightful discussions, and readiness to help have enriched this journey immeasurably.

During my internships, I was fortunate to learn from exceptional mentors. I am particularly thankful to Fernando Diaz, Julia Kiseleva, and Bhaskar Mitra for their guidance, encouragement, and example. In fact, I am indebted to my collaborators all around the world from the University of Amsterdam including but not limited to Maarten de Rijke and Chuan Meng, as well as to my teammates from Meta, MIT, Microsoft, and AIRI during organization of NeurIPS IGLU competition. Working with you has broadened my horizons and strengthened this research.

This thesis was definitely made possible by the generous support of the David Cheriton Graduate Scholarship and the Ontario Graduate Scholarship (OGS). I am deeply grateful for this financial assistance.

To everyone who contributed—through feedback, technical help, or simply believing in me—thank you.

Dedication

To my mother, Afi—the strongest woman I know—whose boundless support and unconditional love have shaped every part of who I am.

To my lifelong friends of eighteen-plus years—Parisa Mahvelatishamsabadi, Paria Kargar Samani, and Pegah Rahmani—who have stood beside me from elementary school through high school, university, and the long road of immigration, never letting my spirits flag and always cheering the loudest.

To the lifelong friends I was fortunate to meet—Mahbod Mehrvarz and his wonderful family—thank you for welcoming me during my hardest moments and filling my new home with warmth.

And to my cousins Aida Zolfaghari, Hanieh Zolfaghari, and Mohammad Hazrati: though continents separate us, your daily calls and constant presence made sure I never felt alone.

To all my family members and friends who laughed with me and cried with me—thank you for celebrating my accomplishments and standing by me through the dark hours. Your presence has meant more than words can express.

This work is for all of you.

Table of Contents

Examining Committee	ii
Author's Declaration	iii
Statement of Contributions	iv
Abstract	vi
Acknowledgements	vii
Dedication	viii
List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Challenges and opportunities	4
1.2 Contributions	7
1.3 Thesis Organization	8
2 Background	9
2.1 Evaluation of Information Retrieval Systems	9

2.1.1	Historical Evolution of IR Outputs	9
2.1.2	Test Collection Topics and Queries	10
2.2	Relevance Judgments	13
2.2.1	Pointwise vs. Pairwise Relevance Judgments	14
2.2.2	LLM-Based Relevance Judgments	16
2.3	LLM as a Judge	21
3	Evaluation of Information Retrieval Systems with Sparse Labels	24
3.1	Shallow Pooling for Sparse Labels	25
3.1.1	Challenges in MS MARCO Relevance Labeling	27
3.1.2	Better Than Perfect	28
3.1.3	Shallow pooling	35
3.1.4	Leaderboard impact	39
3.1.5	Takeaway	43
3.2	Fréchet Distance for Evaluation of IR systems with Sparse Labels	44
3.2.1	Fréchet Distance	45
3.2.2	Experimental Setup	47
3.2.3	Assessment with Sparse labels	49
3.2.4	Assessing with Comprehensive labels	50
3.2.5	Assessing Unlabeled Retrieved Results	54
3.2.6	FID Correlation with IR Evaluation Metrics	56
3.2.7	Impact of Document Representation	57
3.2.8	Takeaway	59
4	Evaluation of Information Retrieval Systems with Large Language Models	60
4.1	Generative Information Retrieval Evaluation	60
4.2	LLM-based Relevance Judgment for IR Evaluation	64

4.2.1	Assessment of Relevance Judgment	67
4.2.2	Relevance Judgment Methods	69
4.2.3	Alignment with Human Labels	75
4.2.4	Agreement with System Rankings	77
4.2.5	Experiments and Results	78
4.2.6	Takeaway	85
4.3	Prompt Sensitivity in LLM-Based Relevance Judgment	86
4.3.1	Prompt Creation	87
4.3.2	Experimental Methodology	93
4.3.3	Results and Findings	95
4.3.4	Takeaway	97
5	Evaluation of Generative Models	100
5.1	Adapting Retrieval Benchmarks for Generative Information Evaluation . .	101
5.1.1	Similarity to Labels: An Anchor for Evaluation	102
5.1.2	Experimental Setup	103
5.1.3	Validation through Cross-Grade Relevance Similarities	108
5.1.4	Assessing Generated Responses with Relevance Judgments	108
5.1.5	Assessing Generated Responses without Relevance Judgments . . .	110
5.1.6	Takeaway	113
5.2	Offline Set-based Evaluation of Text-to-image Generation	114
5.2.1	Text to Image Generation Evaluation	116
5.2.2	Problem Definition	118
5.2.3	Method	119
5.2.4	Experimental Setup	122
5.2.5	Results and Validation	128
5.2.6	Takeaway	133
5.3	Evaluation of LLM-based Applications	135

5.3.1	AgentEval for Task Utility	137
5.3.2	Experimental Setup	139
5.4	AgentEval Results	144
5.4.1	AgentEval for Math Problems	144
5.5	Criteria Diversity	149
5.6	Verification	150
5.6.1	Criteria Stability	152
5.6.2	Quantification Robustness	153
5.6.3	Discriminative Power	155
5.6.4	<i>VerifierAgent</i>	156
5.6.5	Limitations	157
5.6.6	Takeaway	158
6	Concluding Remarks	159
6.1	Summary of Contributions	159
6.2	Broader Implications	161
6.3	Limitations	162
6.4	Future Research Directions	163
6.5	Closing Remarks	165
	References	167
	APPENDICES	215
A	URLs for Reproducibility	216

List of Figures

3.1	Screenshot of the data annotation process.	31
3.2	Examples of Preference Judgments from Section 3.1.2.	32
3.3	Results of Passage Comparison for Better than Perfect Experiment	33
3.4	Size of preference judgment pools for all queries.	36
3.5	Size of preference judgment pools for selected queries.	36
3.6	MRR@10 on Evaluation set vs Development set with Official qrels	39
3.7	Impact of Shallow Pooling and Preference Judgments on System Ranking Evaluation.	42
3.8	Bootstrap sampling of MRR@10 vs FD@10 on MS MARCO Dev set.	51
3.9	nDCG vs FD on TREC runs	52
4.1	A GenIR system as a synthetical search engine	61
4.2	RAG Architecture overview	62
4.3	Human alignment with labels for different relevance judgments	80
4.4	Agreement with system ranking for most aligned relevance judgments	81
4.5	Alignment with human labels vs agreement with system ranking	84
4.6	Diversity of words across human and LLM-generated prompts.	93
4.7	Agreement of LLM-based relevance judgments with human annotations	96
4.8	Krippendorff’s inter-agreement rate between prompts	98
5.1	Prompts used in different settings for generating the answers	105

5.2	Distribution of inter-similarities between relevance judgements	107
5.3	The similarity of the responses of the generated models	109
5.4	Comparison between submitted runs vs generated responses	111
5.5	Performance of the generated models based on similarity to top retrieved documents	112
5.6	Overview of our proposed approach	121
5.7	Diversity vs Saliency on four image grids	126
5.8	Annotation results of TTI systems	127
5.9	An overview of the AgentEval framework	136
5.10	The taxonomy of tasks assessment	138
5.11	An example of Math Problem solving	141
5.12	An example of ALFWorld Householding task	142
5.13	AgentEval assessment of math problems solutions	145
5.14	AgentEval assessment of ALFWorld solutions	148
5.15	Task-based vs solution-based criteria	149
5.16	Distribution of <i>QuantifierAgent</i> output	152
5.17	Criteria stability analysis for math problem solving	154
5.18	Quantifier Robustness on Math Problem Solving task	155
5.19	Discriminative power analysis on math problem solving task	156

List of Tables

3.1	Agreement between qrels and top results from neural ranker.	29
3.2	Results of Pairwise Passage Comparison by Query Category.	29
3.3	MS MARCO Passage Ranking Runs Performance across Different qrels. . .	34
3.4	Win Ratios in Direct Comparisons of Top Documents from Different Runs.	38
3.5	MRR@10 vs FD@10 of different retrievers	49
3.6	Impact of label sparsification on nDCG@10 vs FD	53
3.7	Performance of FD on evaluation of unlabeled retrieved results	55
3.8	Correlation between FD and Traditional IR metrics	56
3.9	Impact of Document Representation.	58
4.1	Scoring functions for Nugget evaluation	74
4.2	Statistics of datasets used in our experiments.	79
4.3	Agreement with system ranking	82
4.4	LLMs used for prompt sensitivity experiments.	93
4.5	Statistics of agreement of LLM-based relevance judgment with human . . .	95
5.1	Details of the retrieval models used for experiments	104
5.2	Details of generative models used in the study	105
5.3	Example of generated results with different LLMs	106
5.4	Examples of prompts from different datasets	122
5.5	Different variations of our proposed metrics	124

5.6	Agreement rate between the three annotators	129
5.7	Agreement rate of the metrics with human annotations	130
5.8	Comparison between Preferred and Not-Preferred set of grids	132
5.9	Verification Criteria for MathProblems	144
5.10	Verification Criteria for ALFWorld Housholding Tasks	147
5.11	Example pairs of similar criteria	150

Chapter 1

Introduction

Information access stands as one of the most important pillars of modern society’s digital infrastructure. Every day, billions of users rely on search engines and information retrieval (IR) systems to satisfy curiosity, make decisions, and acquire knowledge [418]. From students seeking educational resources to doctors consulting medical literature, efficient access to relevant information is a critical enabler across domains. In fact, people depend heavily on systems that support information seeking and comprehension [150, 422], whether for staying informed about global events or making critical decisions in health, education, politics, and daily life [166, 156]. Ensuring that people can swiftly find accurate and useful information is not only a technical endeavor—it is also a societal imperative [353, 271]. As the volume of digital content continues to grow and global connectivity expands, the challenge of connecting users with the information they need has become increasingly complex. Consequently, the need for robust evaluation of information access systems has never been more urgent [462, 16].

Information retrieval (IR) and broader information-seeking systems play a central role in this landscape, acting as the bridge between users’ questions and relevant knowledge. At their best, these systems empower individuals by enabling access to diverse perspectives, up-to-date facts, and new discoveries. In an increasingly complex and dynamic world, the ability to find the right information at the right time is not just a matter of convenience—it is essential for informed decision-making and meaningful democratic participation [356, 274, 261].

However, simply accessing information is not enough. The quality and accuracy of that information matter deeply. If people are exposed to incorrect, misleading, or incomplete information, the consequences can be serious. Misinformation can erode trust, polarize

societies, and lead to harmful decisions [296, 218, 370]. Worse, when information-seeking systems amplify only certain viewpoints or reinforce preexisting biases—creating so-called filter bubbles or echo chambers—users may never encounter alternative perspectives or challenge their assumptions [137, 209]. In this sense, the effectiveness of information-seeking systems is not only a technical issue but also a societal one. Ensuring that these systems return relevant, diverse, accurate, and trustworthy results is critical for individual agency and collective well-being.

Therefore, it is vital to rigorously evaluate the accuracy and quality of the information presented to users. As Lord Kelvin famously said, *“If you cannot measure it, you cannot improve it”*. Evaluation lies at the heart of progress: without it, we cannot determine whether one system is better than another, whether a new technique improves search results, or whether a generative model provides more useful answers. Evaluation frameworks guide innovation, inform deployment, and establish trust in the systems we use every day.

Yet in recent years, the landscape of information access has changed dramatically [34, 238]. Modern information access systems – ranging from search engines to conversational agents – are undergoing a paradigm shift driven by advances in neural retrieval and generative Artificial Intelligence (AI) [224]. Over the past decade, information retrieval (IR) has evolved from traditional term-based methods to sophisticated neural ranking models that have dramatically improved the core retrieval task. Unlike earlier “bag-of-words” approaches, neural models (such as deep learning to rank frameworks and transformer-based re-rankers) can capture semantic meaning and context, allowing users to find relevant information even when query wording and document language differ [272, 144, 273]. With the rise of neural retrieval models, information-seeking systems have made tremendous strides in performance, often surpassing previous benchmarks by large margins [157, 386]. However, this rapid progress has exposed cracks in our long-standing evaluation methodologies. These neural approaches while significantly reshaping the Information Retrieval (IR) landscape, they introduced new challenges such as reliance on large training data and a tendency to produce answers. For instance, traditional evaluations that rely on sparse relevance labeling—where only a small subset of documents is judged for relevance—often fail to accurately reflect the effectiveness of modern neural systems [23, 320]. These models frequently retrieve relevant documents that were never judged, leading to underestimation of their true performance and inconsistencies in system rankings. As a result, our current evaluation tools may no longer be able to distinguish between meaningful improvements and mere artifacts of incomplete labels.

The rise of Large Language Model (LLM)s has significantly reshaped not only the retrieval component of information access systems but the entire user interaction paradigm [9]. With their impressive capabilities in language understanding, generation, and reason-

ing, LLMs have prompted researchers to integrate them into various stages of information pipelines [146, 134, 470]. These models can generate fluent and seemingly plausible answers, though they are not always accurate [180]. A pivotal moment in this transformation came with the public release of ChatGPT in late 2022—a true “game-changer” that drew global attention and accelerated the shift toward generative AI applications. In this new era, the boundaries between information retrieval and generation have become increasingly blurred [224]. Users now engage with conversational agents that not only retrieve documents but also synthesize answers, generate summaries, and provide real-time recommendations [430, 460]. This has led to the growing adoption of the conversational search paradigm, offering more natural and interactive information-seeking experiences [266].

Major search engines have begun to embrace this shift by incorporating LLM-driven features—for instance, generating synthesized answers from web results or enabling multi-turn dialogues with users. Yet, despite these advancements, traditional retrieval technologies remain foundational. Generative systems often rely on retrieved evidence (e.g., in retrieval-augmented generation) to ground their responses and enhance factuality [220, 170]. For users, the promise of Generative Information Retrieval (GenIR) is a more natural and efficient search experience – instead of sifting through documents, one can get concise answers or have a conversation to refine needs [423]. For the IR research community, however, GenIR raises new evaluation questions [9]. What does it mean for an answer to be relevant and useful in this context? A generated answer might combine information from multiple sources, present it with novel wording, or include the system’s own reasoning. Traditional evaluation, which would check if the top documents contain the answer to the query, no longer suffices. We must assess the quality of the AI’s output itself. This paradigm shift forces us to rethink long-standing definitions and metrics: relevance now intersects with notions of correctness, completeness, and usability of the answer provided. Evaluating generative outputs requires new criteria, shifting the focus from document ranking performance to the utility of the information experience provided to the user. As system capabilities, result formats, and user expectations evolve, so too must our evaluation methodologies—otherwise, we risk misjudging what constitutes real progress.

We are living in an incredibly exciting era for artificial intelligence. The rise of generative models, while opens up powerful new possibilities, it also introduces significant challenges for evaluation. As systems become more flexible and open-ended, our evaluation frameworks must evolve to match. We must now grapple with questions such as: How do we measure the quality of a generated response? How do we assess factuality, completeness, or diversity? And how can we fairly compare fundamentally different systems—retrievers, generators, and hybrids—on a unified basis? These challenges are central to this thesis, which seeks to develop evaluation strategies that reflect the true utility of

modern information access systems for real users. As such, in this thesis, I focus on evaluating information-seeking systems through the lens of quality, accuracy, and utility. As these systems become increasingly intelligent and dynamic, it is critical to ensure that they reliably surface correct, relevant, and useful information to end users. Beyond traditional notions of retrieval effectiveness, I am particularly concerned with whether accurate information is truly accessible, available, and—perhaps most importantly—detectable. In other words, does the system make trustworthy information easy to find and discover when users need it most?

With this broader vision in mind, this series of research explores how well modern information access systems align with real user needs. Specifically, I aim to evaluate the true utility of these systems—how effectively they serve their purpose in helping users find meaningful answers. As our tools grow more complex, we must be equally rigorous in measuring whether they are actually helpful, not just technically impressive. It is not enough to generate text or retrieve documents—we must be able to evaluate whether these outputs satisfy the user’s intent, reflect accurate knowledge, and foster trust in the system. These challenges are at the heart of this thesis. Rather than treating them as obstacles, I view them as opportunities to rethink how we measure progress. If we can develop robust, adaptive, and meaningful evaluation strategies for the generative era, we not only keep our research accountable—we also help shape more reliable, responsible, and user-centered information access systems for the future.

1.1 Challenges and opportunities

From an evaluation standpoint, the rapid evolution of GenIR systems introduces a mix of concrete and speculative challenges and opportunities.

- **Challenges** arise primarily from evaluating the *form and quality* of GenIR outputs, which are often synthesized textual responses rather than ranked lists of documents. These outputs may exhibit issues such as hallucinations, lack of factual grounding, or incomplete answers, complicating traditional evaluation based on document relevance.
- **Opportunities** emerge from the ability of generative models to simulate human behavior and judgment. This opens the door for automating components of the evaluation pipeline, such as estimating relevance, simulating user satisfaction, or conducting rapid feasibility assessments prior to costly user studies.

The rapid development of neural and generative information-seeking systems has outpaced the evaluation methodologies that information retrieval IR researchers have relied on for decades [136, 307]. Traditional IR evaluation is rooted in the Cranfield paradigm, which uses static test collections consisting of queries, documents, and human relevance judgments [408, 407]. Typically, systems are compared using metrics such as Mean Average Precision (MAP) or Normalized Discounted Cumulative Gain (NDCG) over these benchmark datasets. While this paradigm has been the backbone of reproducible IR research, it faces well-documented challenges when applied to modern systems [341, 406].

A major issue is the sparseness of relevance labels in most test collections: only a subset of documents (typically those retrieved by top systems) are judged for relevance, while the rest remain unjudged [372]. It is commonly assumed during evaluation that unjudged documents can be treated as non-relevant. In practice, this assumption often does not hold—many unjudged documents may in fact be relevant [341]. As a result, evaluation metrics may systematically underestimate the performance of strong retrieval models that surface relevant documents outside the judgment pool [23]. This bias can distort system rankings and lead to incorrect conclusions about which system performs better. Ultimately, when relevance judgments are sparse, evaluation becomes noisy and unreliable, sometimes failing to capture genuine improvements in retrieval effectiveness. In this thesis, we begin by identifying such flaws and use these observations to design evaluation methodologies that are more robust to label sparsity and better aligned with the capabilities of modern retrieval systems.

Addressing the limitations of current evaluation requires innovative strategies that align with the nature of today’s systems. One direction is to mitigate the problem of sparse labels by generating additional relevance judgments automatically. Large Language Models themselves offer a promising tool for this. Recent work suggests that LLMs can be used as surrogate assessors, labeling the relevance of documents to a given query in lieu of human judges . By carefully prompting an LLM with guidelines and examples, it can assign fine-grained relevance scores that correlate well with human judgments. Such an approach could “patch up” missing relevance judgments in test collections by filling in the gaps for unjudged documents. If successful, this would greatly reduce evaluation bias: systems retrieving genuinely relevant information would no longer be penalized simply due to lack of human labels. However, employing LLMs for evaluation is not straightforward – there is variability in how different LLMs and prompting techniques perform, and no standardized protocol yet exists for their use in IR evaluation. This motivates a systematic study of LLM-based relevance judgment methods (explored in this thesis), to compare their reliability and to develop best practices for automated, LLM-assisted evaluation. Therefore, we investigate how LLMs can be used to have more accurate evaluation of IR systems.

Another challenge is that traditional ad hoc retrieval benchmarks were designed to evaluate ranked lists of documents. But how should we evaluate a system that directly generates an answer or summary, possibly drawing from multiple documents? If an LLM-based system provides a fluent answer, using standard metrics is not straightforward – one cannot simply mark the answer as “relevant” or “not relevant” in the same way as a retrieved document. Generative models can produce content that is partially correct and partially incorrect, or omit key details, challenging the binary or graded relevance criteria that classic evaluations use. Evaluating the quality of generated text (e.g., for accuracy, completeness, or coherence) usually requires intensive human assessment or specialized metrics, and these may not align neatly with traditional IR measures. The result is a growing mismatch between what we want to measure and what our tools can capture. Current evaluation frameworks, built around static relevance labels and predetermined answers, often fail to fairly compare a search engine that retrieves documents against an AI agent that writes an answer. In summary, existing methods struggle to comprehensively assess modern retrieval and generative systems: they neither account for the incomplete ground-truth in large collections nor capture the nuanced success criteria of generative outputs. This gap in evaluation methodology forms the core problem that this thesis tackles. To ensure that real improvements are recognized and that different classes of systems can be compared on equal footing, we must rethink and extend our evaluation strategies.

We interpret evaluation in the generative era along three complementary directions:

1. **Identifying limitations of traditional evaluation approaches:** Existing evaluation frameworks for information retrieval often depend on manually curated relevance judgments, which are costly to obtain, inherently sparse, and frequently incomplete. These limitations can result in unreliable or biased assessments. Recognizing these challenges is a critical first step toward effectively leveraging large language models to address and overcome them.
2. **LLM-assisted evaluation of IR systems:** Large language models offer new opportunities to enhance traditional IR evaluation—for instance, by generating relevance judgments, simulating annotator behavior, or enabling scalable, low-cost assessments. These methods help bridge gaps caused by sparse or incomplete ground truth.
3. **Evaluation of generative systems and their utility:** As generative information-seeking systems (e.g., Retrieval Augmented Retrieval (RAG)-based systems) become prevalent, we must also evaluate the quality and utility of their outputs. This involves assessing dimensions such as fluency, factuality, completeness, and useful-

ness—particularly in open-domain scenarios where ground truth is ambiguous and/or success is not well defined.

1.2 Contributions

Building on the above motivations and approaches, this dissertation makes the following key contributions:

- **Empirical diagnosis of sparse-label bias.** We present the first large-scale study quantifying how incomplete relevance pools distort system rankings, showing that state-of-the-art neural retrievers are systematically undervalued when they surface unjudged—but truly relevant—documents.
- **Frechet-Distance evaluation for retrieval *and* generation.** Adapting distributional comparison techniques from generative-model research, we introduce an FD-based metric that (i) remains stable under severe label sparsity and (ii) places retrieval systems and answer-generating systems on a common scale.
- **Standardised LLM-assisted judging toolkit.** We design a reproducible pipeline that prompts LLMs to deliver graded, TREC-style relevance labels, release a benchmark comparing models & prompting patterns, and provide calibration guidelines to align LLM scores with human assessments.
- **Prompt-robustness analysis for automated judging.** Using our toolkit, we systematically characterise how wording, example ordering, and instruction granularity influence LLM judgments and downstream leaderboard stability—establishing best practices for practitioners replacing or augmenting human judges.
- **Unified evaluation for generative information-access tasks.** Beyond document retrieval, we extend IR-inspired metrics to (i) retrieval-augmented answer generation, (ii) offline set-based text-to-image evaluation, and (iii) AgentEval, a task-utility framework for complex LLM applications. These studies demonstrate how user-oriented criteria (factuality, diversity, task success) can be operationalised without exhaustive human annotation.

Through the above contributions, this thesis significantly advances the state of the art in evaluation methodology for information retrieval and access systems. The outcomes

help ensure that as our systems become more powerful and complex – combining retrieval with generation – our evaluation techniques remain rigorous and meaningful. Ultimately, the new evaluation tools and insights presented here aim to keep evaluation in lockstep with innovation, so that we can accurately measure improvements and drive progress in the field.

1.3 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 reviews relevant literature and background, covering both traditional information retrieval evaluation methodologies and recent efforts to evaluate neural and generative systems. Chapter 3 introduces the problem of sparse relevance labels by analyzing biases in conventional evaluation approaches and proposes a Fréchet Distance-based method designed to mitigate these challenges. Chapter 4 explores the use of large language models (LLMs) for evaluation, presenting the design of an LLM-based relevance assessment framework, comparing its variants, and examining how prompt sensitivity affects judgment quality. Chapter 5 extends evaluation beyond document retrieval to generative applications, demonstrating how IR techniques can be adapted to assess answers generated by LLMs. Further, we broaden the scope of evaluation to more complex systems such as text-to-image generation, proposing a general framework for evaluating LLM-powered applications. Finally, Chapter 6 concludes the thesis by summarizing the main findings, discussing their implications, and outlining promising directions for future research.

Chapter 2

Background

2.1 Evaluation of Information Retrieval Systems

The evaluation of Information Retrieval (IR) systems has traditionally followed the Cranfield paradigm, which employs standardized test collections to ensure reproducible and fair comparisons. A typical test collection consists of a document corpus, a set of user queries (or topics), and relevance judgments that indicate which documents are relevant to each query. When given a user query, an IR system produces a ranked list of documents, and its effectiveness is assessed by comparing its rankings against human relevance judgments. Standard evaluation metrics—such as Precision, Recall, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG)—quantify how well the system ranks relevant documents toward the top of the list.

Offline evaluation using test collections has been the predominant approach for assessing IR system performance due to its reproducibility and relatively low cost once the judgments are in place. This methodology has underpinned IR research for decades, enabling consistent benchmarking across systems. However, constructing these collections—particularly obtaining reliable relevance judgments—has historically been labor-intensive with lots of challenges.

2.1.1 Historical Evolution of IR Outputs

Early IR systems primarily returned a ranked list of documents, as reviewed by Harman in pre-web retrieval systems [160]. Over time, the presentation of search results evolved

to enhance user experience and retrieval effectiveness. The introduction of query-sensitive snippets [390]—summaries extracted from the body of the document that highlight relevant content—marked a significant advancement in search result presentation [99].

Commercial search engines further extended result pages beyond ranked lists, incorporating direct answers [432], quick links [73, 155], entity cards [54, 286], and query suggestions [64]. These additions transformed search results into richer, more interactive experiences, shifting the focus from ranking individual documents to optimizing the entire search results page.

This evolution led to research on whole-page relevance [36], which considers the collective effectiveness of various components on a results page. However, despite its potential, this approach did not gain widespread adoption in academic IR research, likely due to the high costs associated with implementing and evaluating it.

In this thesis, we explore a broad spectrum of output formats produced by modern information retrieval systems, ranging from traditional ranked lists of documents (Sections ?? to more conversational single response format of generative models (Section 5.1).

2.1.2 Test Collection Topics and Queries

A standard test collection consists of a set of search topics (expressed as queries), a corpus of documents, and relevance judgments, often referred to as *qrels*, that indicate which documents are relevant to which topics. To evaluate an IR system, queries from the collection are processed through the system, and its effectiveness in retrieving relevant documents is assessed. This automated evaluation approach enables cost-effective system optimization. However, creating high-quality test collections remains an expensive and labor-intensive process.

Several simplifying assumptions in offline evaluation practices are highlighted in [408]:

- Relevance is approximated by topical similarity, implying that all relevant documents are equally desirable, the relevance of one document is independent of another, and the user’s information need is static.
- A single set of judgments per topic is representative of the user population.
- (Essentially) all relevant documents for a topic are known.
- Each information need is represented by a single query.

These assumptions help manage the cost of constructing test collections, but the process remains prohibitively expensive for individual researchers. Consequently, large-scale initiatives such as Text Retrieval Conference (TREC) [403], Cross Language Evaluation Forum (CLEF) [312], and NII Testbeds and Community for Information Access Research (NTCIR) [294] were established to distribute the workload and share resources.

Early test collections contained only a few hundred to a few thousand documents¹. In these early efforts, human assessors manually reviewed documents for relevance, as seen in Cranfield II and LISA [347]. However, the cost of manual relevance assessment constrained the scalability of test collections. To address this, Spärck Jones and Van Rijsbergen [373] introduced document pooling, where documents retrieved by multiple participating systems (runs) are sampled for relevance judgments. While this approach facilitates large-scale evaluation, it risks omitting relevant documents [473], which raises concerns about the reusability of collections [409, 58]. Nonetheless, it enabled the creation of test collections using collections such as ClueWeb09², ClueWeb12³, ClueWeb22 [299], which contain millions of documents.

In traditional test collections, each search topic is typically represented by a *single* query. Queries are either manually generated based on information need statements or extracted from query logs. Expanding beyond one query per topic to represent a diverse range of user expressions has historically been considered unnecessary and costly. However, studies have shown that users formulate many variations of queries to express the same information need, with over fifty variants observed per topic [37, 255]. Factors such as device type [95, 163], domain expertise [282, 424], age [392, 48], and language proficiency [94] influence query formulation, impacting retrieval effectiveness. Culpepper et al. [119] found that query variation has a greater impact on system effectiveness than topic selection or ranking models. Despite this, test collections still predominantly rely on single-query evaluations, leaving a significant gap in assessing system robustness across diverse user formulations.

LLMS for Test Collections

Recent research explores LLMs to address this challenge. Unlike relevance judgments, where LLMs have been validated as a substitute for human labels, their application for

¹Historical test collections are archived by the University of Glasgow: https://ir.dcs.gla.ac.uk/resources/test_collections/

²<https://lemurproject.org/clueweb09.php/>

³<https://lemurproject.org/clueweb12.php/>

query variant generation is in its infancy. Penha et al. [310] demonstrated that artificially generated query variants can significantly degrade neural and transformer-based retrieval models. Similarly, Alao et al. [8] found inconsistencies in retrieval results across query variants in commercial search engines and inverted indexes, highlighting the need for robust evaluations. Other studies explore using LLMs to simulate query variants interactively [130] or to generate queries for training better rankers [50], query expansions [256], and document retrievability improvements [311]. A critical open question is whether LLMs can align query generation with real-world user behaviors, reflecting the diverse properties identified in query formulation studies.

In this thesis, in Chapter 4, we investigate how LLMs can be leveraged to automate the construction of test collections, reducing the cost, time, and labor involved in creating high-quality relevance judgments.

MS MARCO

The MS MARCO⁴ benchmark focuses on deep learning for search, comprising a corpus of over 8.8 million passages in V1 and over 138 million passages in MS MARCO V2 collection [287]. It includes training, development (validation), and evaluation (test) query sets, with most queries associated with a single known relevant document (qrel). Training and development qrels are publicly available, while evaluation qrels remain private. To participate in the leaderboard, research teams train and validate rankers on the public data, generate ranked lists for evaluation queries, and submit results to the MS MARCO team. In this thesis, most of the experiments are done on this collection.

The passage retrieval leaderboard, which is the focus here, is based on queries sampled from the Bing search engine. The training set for MS MARCO V1 contains over 500K queries, the development set 6,980 queries, and the evaluation set 6,837 queries. In the development set, 94% of queries have a single qrel, and no query has more than four qrels. While both development and evaluation sets track retrieval improvements, Mean Reciprocal Rank (MRR) among top-10 retrieved documents on the evaluation set serves as the official measure.

TREC Data

The TREC Deep Learning (Deep Learning (DL)) Tracks [112, 113, 117, 116] examined the impact of extended relevance judgments on evaluation. Unlike MS MARCO’s dev

⁴<https://microsoft.github.io/msmarco/>

set, which contains thousands of queries with mostly single qrels, TREC DL focuses on a smaller set (around 50 queries) with multi-level judgments (graded 0-4) as follows:

- 3 Perfectly relevant: Dedicated to the query, worthy of being a top search result.
- 2 Highly relevant: Provides substantial information on the query.
- 1 Relevant: Contains some useful information, albeit minimal.
- 0 Non-relevant: Provides no useful information on the query.

The DL-19 and DL-20 tracks used the MS MARCO v1 corpus (8.8 million passages), while DL-21, DL-22 and DL-23 employed MS MARCO v2 (138 million passages). The official evaluation metric for these datasets is nDCG@10, reflecting the richer judgment granularity.

2.2 Relevance Judgments

Relevance judgments serve as the foundation of Information Retrieval (IR) evaluation, providing the ground truth for determining which documents satisfy a given information need. Traditionally, these judgments are obtained from human assessors who examine query-document pairs and assign relevance labels, either as binary relevance (relevant or not) or on a graded scale. Early IR experiments could afford to exhaustively judge all documents for a query when collections were small. However, as document collections expanded, it became impractical to judge every document for every query.

To address this scalability issue, the community adopted pooling, introduced by Spärck Jones and van Rijsbergen [190, 371], as a strategy to select a subset of documents for assessment. In pooling, the top-ranked results from multiple retrieval systems are merged into a pool, and only these documents are judged by humans. This approach became the standard in large-scale evaluations like TREC and remains widely used today. While pooling significantly reduces the annotation burden, it introduces the issue of label sparsity: many relevant documents outside the pooled set remain unjudged. By design, unjudged documents are treated as non-relevant in evaluation, which can unfairly penalize systems that retrieve truly relevant documents that were not included in the pool. As ranking technology has advanced—particularly with neural rankers—systems now retrieve relevant documents beyond those originally judged. If these newly surfaced documents remain

unjudged, evaluation metrics may underestimate the true performance of modern ranking models.

Another challenge with human relevance judgments is their consistency and cost. Human assessors may disagree on what is relevant, especially on graded scales or nuanced information needs, leading to variability. Traditional labeling is also expensive and time-consuming, particularly when expert assessors are needed. These issues have motivated researchers to explore alternative judgment approaches, including pairwise (or preference-based) judgments.

2.2.1 Pointwise vs. Pairwise Relevance Judgments

The two primary approaches to collecting relevance judgments—pointwise and pairwise—differ in how they elicit assessments and their implications for evaluation.

Pointwise judgments are the conventional method in IR evaluation. In this approach, assessors independently label each document in isolation, assigning it a relevance score (e.g., binary relevance or graded scales such as 0–4). This method is widely used due to its straightforwardness and alignment with traditional evaluation metrics like MAP and NDCG. While pointwise judgments enable efficient pooling-based evaluation, they can be inconsistent due to assessor subjectivity and fail to capture the relative preference between documents, which is crucial in ranking tasks.

Pairwise (or preference-based) judgments offer an alternative approach where assessors compare two or more documents and indicate which is more relevant to the query. Instead of assigning absolute scores, assessors express relative preferences between documents. This approach is motivated by the idea that ranking effectiveness depends more on relative ordering rather than absolute relevance levels. Pairwise judgments can be more intuitive for assessors and may mitigate some inconsistencies in pointwise ratings. However, they require a greater number of comparisons. While pointwise judgments require only N assessments for N documents, pairwise comparisons theoretically require up to N^2 comparisons. Assuming transitivity (if A is preferred to B and B is preferred to C , then A is preferred to C) can reduce the number of required judgments to approximately $N \log N$, but in practice, transitivity does not always hold due to assessor disagreements.

Despite the potential benefits of pairwise judgments, their adoption has been limited. Studies have produced mixed findings: for example, Sakai and Zeng [342] crowdsourced over 100,000 pairwise judgments for an NTCIR evaluation but found no clear advantage over traditional pointwise judgments. Conversely, Yang et al. [446] showed that pairwise preference judgments can be cheap, effective, and as reliable as pointwise judgments. Work

by Carterette and colleagues [70, 469, 67, 69] suggests that pairwise judgments can improve assessment speed and accuracy.

One notable concern preventing widespread adoption is the computational overhead. While pointwise judgments require only N judgments for N documents, pairwise comparisons theoretically require a quadratic number of judgments. However, alternative designs—such as tournament-based comparisons or selecting the single best document—can reduce this burden. Recent work from our group [102] has explored tournament-style pairwise evaluations to efficiently identify top- k documents while minimizing required comparisons.

Pairwise approaches are particularly useful in ongoing leaderboard-based evaluations, such as MS MARCO. The qrels in MS MARCO implicitly define certain passages as the “right answer” and evaluate ranking models based on their ability to place these answers at the top. As retrieval models improve, better answers may emerge, requiring qrels to be updated. This motivates the need for a dynamic, comparative assessment approach: when a previously unjudged document is surfaced at rank 1, it should be compared against the current best-known answer, updating the qrels if it proves superior.

This best-known answer evaluation aligns naturally with comparative assessments rather than traditional pointwise judgments. Comparative methods have been explored in other domains, such as image search, where Shao et al. [360] demonstrated that side-by-side comparisons led to judgments that better reflected user satisfaction than pointwise ratings. Similarly, Xie et al. [434] and Maddalena et al. [257] explored multi-document comparisons to refine assessments in fine-grained ranking tasks.

Both pointwise and pairwise judgments aim to provide relevance labels that enable fair and accurate evaluation of IR systems. While pointwise judgments remain the dominant approach due to their efficiency and established metrics, pairwise judgments offer potential benefits by focusing on relative ordering and mitigating annotation inconsistencies. However, concerns regarding judgment scalability and metric validation have slowed their adoption. In cases where rankings are continuously updated—such as leaderboards—comparative assessments can provide a more dynamic and efficient evaluation framework.

In this thesis, we address how do point-wise and pair-wise relevance judgment formats compare when used with large language models? While both formats are commonly employed, a systematic comparison of their behavior and agreement has been lacking. In Chapter 4, we propose a framework for directly comparing these two types of LLM-based relevance judgments, examining not only how closely they align with human annotations, but also how sensitive each is to prompt variations.

2.2.2 LLM-Based Relevance Judgments

The automation of relevance judgments has a long history in Information Retrieval. Long before the emergence of large language models, researchers explored ways to replace or augment human assessments using traditional machine learning techniques. For example, in early work by Büttcher et al. [61], the authors proposed training classifiers such as Support Vector Machines and Kullback-Leibler divergence-based models to predict the relevance of unjudged documents using existing judged documents as training data. Their study showed that with a sufficiently rich pool, these simple models could extend relevance judgments to unjudged documents and mitigate evaluation bias, particularly in settings with incomplete or system-biased pools. Similar attempts using pseudo-relevance feedback and regression-based models also appear in foundational IR work [450, 31, 384]. These early experiments laid the groundwork for today’s LLM-based approaches.

Recent studies have demonstrated that LLMs can be used to generate relevance judgments, offering an alternative to human assessors. In May 2023, researchers at Microsoft Bing announced the use of GPT-4 to produce relevance labels, later detailed in [388]. The LLM-generated labels were found to be as accurate as those created by crowdsourced workers and were integrated into the training process of Bing’s production system. Around the same time, Faggioli et al. [131] reported promising results using LLMs for relevance labeling. Although these findings require further validation and may have limitations, they suggest a potential shift away from document pooling, which was originally introduced to mitigate the costs of human annotation. With LLMs, it may now be feasible to generate comprehensive relevance judgments at scale or at least create deeper pools at a significantly lower cost.

The use of LLMs to reduce the cost of relevance judgments parallels a historical shift in the value of aluminum. When the Washington Monument’s aluminum cap was placed in 1884, aluminum was as rare and valuable as silver. However, within two years, Paul Héroult and Charles Hall independently developed an electrolysis-based extraction process that made aluminum inexpensive and widely available. Similarly, IR evaluation may be experiencing its own Hall-Héroult moment: once a scarce and costly resource, human relevance judgments can now be produced on demand by LLMs at a fraction of the cost, opening new possibilities for evaluation.

One such opportunity is the ability to refine the definition of relevance by incorporating additional dimensions of information utility tailored to different users. Voorhees et al. [409] highlighted the issue of score saturation in the TREC Deep Learning Track (2021), where many systems could retrieve ten relevant documents per query from large corpora, necessitating *“different metrics or a more focused definition of relevance”*. Traditionally,

relevance has been assessed based primarily on topicality; however, other dimensions of relevance have also been explored [351]. The use of LLMs presents an opportunity to create richer, more specific relevance labels that better reflect real-world information needs, leading to more meaningful system evaluations.

Another advantage of LLMs is their consistency in generating relevance labels. Unlike human assessors, LLMs do not experience fatigue, nor are they influenced by prior judgments. Studies have shown that human relevance assessments exhibit significant inconsistency [352, 348], whether due to forgetting earlier decisions, recalibrating relevance criteria based on previously seen documents, or making errors, sometimes leading to contradicting assessments for similar documents [45, 352]. By leveraging deterministic LLMs with controlled parameters, evaluations could achieve greater reproducibility and standardization.

While dynamic, on-demand relevance labeling using LLMs may reduce the need for static test collections, it also challenges the principle of shared, reusable resources in IR evaluation. Test collections provide a common benchmarking foundation, enabling fair and repeatable comparisons across systems. Without such standardized resources, dynamic evaluation may introduce inconsistencies in how different models are assessed. Thus, while LLMs present exciting new opportunities, their role in IR evaluation should be carefully integrated to balance cost-effectiveness with the reliability and stability of traditional test collections.

While there have been several recent attempts to use LLMs for automated relevance judgment, this thesis makes two key contributions that, to our knowledge, have not been explored before. First, in Section 4.2 we systematically benchmark different approaches to LLM-based relevance judgment and introduce a unified methodology to compare their effectiveness, agreement with human annotations, and consistency across tasks. Second, we study the prompt sensitivity of each method in Section 4.3, analyzing how variations in prompt phrasing affect judgment quality.

Grounding Simulations: Gold is Still Precious

Evaluation outcomes of systems using test collections reflect ‘anticipated’ real-world performance. Although these test collections appear concrete, featuring human queries and relevance judgments, they are fundamentally abstract and considerably simplified *simulations* of real-world search scenarios. Use of so-called *offline evaluation* imagines a simplified process of a searcher browsing the sorted list, top to bottom, identifying relevant pages and at some point stopping[279, 277, 25, 367, 414]. The extent to which evaluation outcomes

reflect actual user satisfaction is crucial. Turpin and Scholer [394] first demonstrated that test collections may poorly reflect reality and then this observation was further investigated by Al-Maskari and Sanderson [7].

The use of LLMs to simulate users in creating test collections raises questions about the validity of this simulation and necessitates further exploration of how well LLMs are aligned with real users. Before going further in simulation and drawing conclusions about how well systems perform, we need to first substantiate the validity of our user simulations. This requires datasets, tools, metrics, and procedures.

User relevance judgments and queries are abundantly available through numerous iterations of shared tasks. Consequently, the approximation of queries and relevance labels to human generated ones can be examined. However, if personalized relevance labels are to be simulated, for example, taking into account other dimensions of information utility, then we have almost no way to validate their performance since such ground truth data is not widely available. For instance, in a context where we would like to evaluate how well a system performs in response to an expert user as opposed to a non-expert, such data is not readily available. Similarly, when simulating query variants issued by multiple users, very few sources of data are available for validation, and demographic data is often missing. Real human data that fits the definition of gold [35], where both the query and relevance assessments are produced by a diverse set of humans operating in different contexts and demographic data is collected are highly needed in order to facilitate the research of simulation validation.

In terms of measuring the accuracy of simulations, that is how closely the LLM aligns with human searchers, one can consider if the simulated data exhibits similar properties to human-generated data or leads to comparable conclusions [39], as exact matches may not be feasible in tasks involving language, where queries can be formulated in various ways. Statistical properties of queries, such as length and complexity, can serve as indicators. Other metrics may assess the impact of simulated data compared to human-generated data. For instance, do the generated queries demonstrate similar effectiveness to human queries and/or produce similar pools of documents? Do the relevance labels result in the same system rankings as if those produced by humans are used?

Slow Search for Evaluation

In 2023, researchers proposed replacing human relevance assessments with LLM assessments [388, 131]. A common objection to these proposals recognizes their circularity. Using automated methods to assess other automated methods is not without its dangers.

If, as is common in this chapter, one looks at historical precedence for current events. One could look at the way in which automated relevance assessments were attempted earlier in the history of IR. A classic example is pseudo relevance feedback [118]. This is a technique that assumes a query from a user will be sufficiently accurate that one can make the assumption that top ranked documents returned by that initial query are themselves likely to be relevant. The text of those documents can then be used in an internal reformulation of the query to produce better results e.g., using pseudo relevance feedback [440].

If LLM assessment is sufficient to replace a human assessment, then why not treat the LLM as a ranker, ranking items according to their LLM assessed relevance? If an LLM-based evaluation is generating the labels for evaluation, ranking by those labels always produces an ideal result.

One way to avoid this circularity is to consider the difference in time and resources needed by a production GenIR system vs. the time and resources required for LLM-based evaluation. For evaluation purposes, we can take all the time we need to find the best response, and then use that response to evaluate the efficiency vs. effectiveness trade-off between, for example, a production system that responds in 100ms and one that responds in 500ms. From the standpoint of an efficiency vs. effectiveness trade-off, for the purposes of evaluation we can essentially ignore efficiency.

The trade-off between retrieval efficiency and effectiveness has long been a subject of academic research [60, 100, 29, 463, 21] and a key consideration for commercial search engines, which aim for an average query latency in the hundreds of milliseconds [264, 29]. However, in the past we have had relatively few methods for tuning the trade-off between efficiency and effectiveness beyond a narrow range. Efficiency vs. effectiveness trade-offs might be measured in terms of tiny percentages of effectiveness improvements at the cost of milliseconds of query latency, but we could never improve effectiveness enough to justify a latency of seconds or longer

Teevan et al. [385] in advocating for “Slow Search” write, “*With even just a little extra time to invest, search engines can relax existing restrictions to improve search result quality. For example, complex query processing can be done to identify key concepts in the query, and multiple queries derived from the initial query can be issued to broaden the set of candidate documents to cover different aspects of the query.*”

Unfortunately, it was never fully demonstrated that investing more time would ever achieve these goals. We had no way to operationalize the proposal of Teevan et al. [385]. If a search engine is fast, the searcher can quickly see if the results are not relevant and immediately reformulate their query [335]. If a query is missing a key concept, the searcher can add it. Low latency is an important feature of search engines, since it facilitates rapid

interaction. We can only justify higher latency if rapid interaction is not required.

We have now entered an era where deriving multiple queries and other complex query processing might genuinely improve the results in more than a trivial way. With more time, our GenIR system might prompt an LLM to make relevance judgments, determine what aspects of a document make it relevant, and automatically refine queries in light of these determinations. A Gen-IR system might compare one item against another, until it identifies the best overall result. In some cases, it might be worth the time of the searcher to wait for this result, but if not, it can still be used to evaluate the faster result actually returned to the searcher.

In some sense, evaluation has always been slow search with a human in the loop. In a traditional TREC ad hoc task, we build a pool and humans assess items in the pool, creating an ideal response. Now we can use an LLM to replace these humans. However, unless we determine that taking all the time we need always produces the best possible response, we still need a way to evaluate the results of slow search. If the quality of LLM assessment can reach the level of traditional human assessment, do we consider this as our peak achievement? Or do we recognize that there is still room for improvement by involving humans to perhaps monitor LLMs or revisit our ideal definition of relevance.

We might ask what replaces these principles in a GenIR system. One idea is provided by the work of Rajput et al. [325]. They propose *nuggets* as a basis for evaluation, where we might think of nuggets as an atomic unit of relevance, e.g., some fact, relationship, or concept that a perfectly relevant document would contain [343, 103]. Rajput et al. [325] propose to build a reusable test collection in a two-phase process. In the first phase, human assessors would identify and extract nuggets from relevant documents. In the second phase, these nuggets would be automatically matched against unjudged documents to measure relevance, providing a reusable test collection that does not depend on a fixed corpus with relevance labels for individual items.

While they provide experimental support demonstrating both the feasibility and benefits of this approach, it was not widely adopted for either academic or industry assessment. Possible reasons include the need for reliable and trained assessors to identify nuggets, as well as the need to automatically match the nuggets against documents. In 2012, they could only suggest a surface-level, lexical approach to matching, and of course humans are expensive. Crowdsourcing might reduce the cost, but might increase noise and decrease reliability.

In 2024, an LLM might be expected to reliably and cheaply extract nuggets and match them against documents. All it takes is a few calls to an API, costing fractions of a cent per call. It’s now almost trivial to realize the vision of Rajput et al. [325], and this proposal

is just one of many such proposals in the literature. All the proposals for IR evaluation in terms of diversity, novelty, fairness, completeness, conciseness, effort, or whatever are now both cheap and straightforward to implement.

We can already see nugget-based evaluation emerging as a basis for GenIR evaluation. For example, the new TREC 2024 RAG track ⁵ takes a nugget-based approach. To formulate a general principle we turn to Zhai et al. [454]. They propose *subtopic evaluation*, which is closely related to nugget-based evaluation. Evaluation with subtopics is “based on *dependent relevance*, instead of *independent relevance*, as has been assumed in most traditional retrieval methods. The subtopic retrieval problem has to do with finding documents that cover as many different subtopics as possible”. To extend this idea to GenIR, we might articulate a principle that the system’s response should cover as many nuggets or subtopics as possible.

A competing definition of “Generative Information Retrieval” systems describes a GenIR system as one that does not generate the answer to the searcher’s query, instead it replaces a traditional search engine by using a neural model to directly generate the identifiers of documents that answer the query [445, 83, 317, 377, 84]. While these systems are “generative” in the sense that a neural model is directly generating document identifiers, from an evaluation perspective they are no different from any other retrieval component that returns document identifiers, except in one respect. Since the document identifiers are generated, it is conceivable for such systems to “hallucinate” document identifiers that do not exist. Nonetheless, for the purpose of our discussion, we view them as a type of retrieval component.

2.3 LLM as a Judge

Users interacting with various systems often operate under specific utility functions [226, 33, 6]. Traditionally, user satisfaction metrics were derived from large-scale behavioral signals [204], tailored to applications such as intelligent assistants [205, 206], web search engines [427, 428, 426, 24], dialogue systems [355], multi-turn conversations [227, 280], and personal assistants [203]. Recent work highlights the need for multiple evaluation metrics to accurately assess user satisfaction [28].

A growing body of research explores the role of LLMs as evaluators [88, 142, 9, 19, 181], including in qualitative research [42] and summarization tasks. Jain et al. [182] examined

⁵trec-rag.github.io

the effectiveness of few-shot prompted LLM evaluators in assessing LLM-generated summaries. Similarly, Wang et al. [411] investigated whether ChatGPT could be directly used as an evaluator by prompting it to score texts. Other studies explore how LLMs serve as proxies for human behavior [389, 241, 88, 267], or assist human evaluators, as demonstrated by CoEval [223]. Pan et al. [303] further illustrate how LLM-based evaluations can improve downstream task performance. However, concerns regarding the reliability of single-LLM evaluations have led to approaches such as multi-step calibration [415].

Recent research investigates multi-agent LLM evaluation frameworks to address single-agent limitations. Chan et al. [75] introduced ChatEval, where multiple LLM agents engage in discussions to evaluate responses, improving alignment with human preferences. Similarly, Chern et al. [87] proposed a debate-assisted meta-evaluation framework for enhancing evaluation consistency.

Building on this, we propose an automated multi-agent assessment framework that evaluates the utility of LLM-powered applications. Our approach provides in-depth insights for developers, identifying flaws in applications and facilitating iterative improvements as the application flow evolves.

LLMs are increasingly utilized for evaluation, particularly in assessing open-ended generation tasks. Chan et al. [75] developed ChatEval, a multi-agent referee team that debates and evaluates responses. Similarly, Chern et al. [87] proposed a meta-evaluation framework leveraging multiple debating agents to enhance annotation consistency. Wang et al. [411] explored ChatGPT’s capability as an evaluator by prompting it to score texts, while Chiang and Lee [88] analyzed LLM-based text assessment against human ratings. Other studies have identified weaknesses in LLM-based evaluation and proposed enhancements, such as multi-step calibration [415]. In contrast, we advocate for a structured multi-agent evaluation framework where each LLM assumes a defined role, improving assessment quality.

Despite advancements in Natural Language Generation (NLG), accurately evaluating generated content remains an ongoing challenge [340, 44, 176, 246]. Prior work [154, 472, 78, 229, 26] has examined various evaluation dimensions, including ethical considerations [374], alignment with human preferences [168, 213], robustness [412, 357], and reasoning capabilities [46]. LLM evaluation has also extended to specialized domains such as medicine [187], multi-modal tasks [269, 41], and interactive agent-based environments [244].

In the field of Information Retrieval (IR), LLMs have enabled significant advancements in document ranking [461, 378, 319], query generation [413, 375], and system evaluation [9, 323, 344]. Their ability to generate relevance judgments that closely align with human assessments has enhanced IR evaluation methodologies [388, 221, 471].

Traditional IR evaluation relies on ranking effectiveness metrics based on human-annotated query-document relevance labels [302, 301]. While initiatives such as TREC, NTCIR, CLEF, and FIRE have established reusable test collections [4, 194, 161], human annotation remains resource-intensive, costly, and susceptible to inconsistencies [396].

The rapid evolution of AI demands continuous updates to evaluation benchmarks [466, 154]. As models achieve near-ceiling performance on existing benchmarks [265], the emergence of novel tasks necessitates automated evaluation to complement human efforts [78]. Automated evaluation methods accelerate benchmarking processes and enable more efficient system development [87, 27].

While the use of LLMs as relevance judges has been widely explored, prior work has mainly treated them as static evaluators—given a fixed prompt and set of criteria—without considering how evaluation itself can be adapted or verified. What remains underexplored is the idea of a closed-loop evaluation process, where an LLM not only performs judgments but also dynamically proposes the criteria, quantifies system performance based on those criteria, and verifies the soundness of its own evaluation. This kind of feedback-driven evaluation is essential for building practical LLM-based applications, where developers need reliable and adaptive tools to assess system behavior as they iterate. In this thesis, we address this gap in Section 5.3.1 by introducing a framework that operationalizes this closed-loop evaluation process—automatically generating evaluation criteria, applying them, and verifying the resulting assessments.

Chapter 3

Evaluation of Information Retrieval Systems with Sparse Labels

The evaluation of IR systems fundamentally relies on gold standard labels e.g., relevance judgments, which serve as the ground truth against which system performance is assessed. These labels, typically obtained through human annotation, playing as an anchor in determining the quality and reliability of retrieval models. The higher a retrieval system ranks the relevant judged items, the more effective the system is [278, 276]. However, acquiring such labels is resource-intensive, expensive, and inherently limited, leading to a persistent challenge: label sparsity. In large-scale retrieval tasks, only a fraction of the retrieved documents can be judged, resulting in an incomplete ground truth that affects system evaluation [106].

Evaluation with incomplete labels presents certain challenges and limitations. First, it introduces biases in system comparisons—a model that retrieves an unjudged relevant document may be penalized unfairly, while another model that retrieves a known judged document may appear superior despite comparable retrieval effectiveness [61, 57]. The presence of unjudged relevant documents means that systems may be optimizing for known relevance rather than actual relevance, leading to a false ceiling effect where further improvements in retrieval effectiveness are not captured due to missing ground truth labels. Second, label sparsity reduces the reliability of traditional evaluation metrics such as nDCG, as they assume a relatively complete set of relevance judgments [341].

In this chapter, two core challenges associated with sparse labeling in evaluation are addressed. First, the effect of sparsity in relevance labels in evaluating retrieval systems is examined. We highlight how modern neural rankers can surpass the performance of

documents labeled as relevant within evaluation with sparse annotation, raising concerns that traditional sparse relevance labels are insufficient for assessing the performance of more advanced retrieval models [275]. As retrieval systems advance and retrieve relevant documents beyond traditional pools, the inability to identify truly superior results due to incomplete judgments distorts performance comparisons. In this chapter, this issue is analyzed in depth by investigating the limitations of current pooling strategies and their implications for evaluation.

Second, we propose an alternative evaluation strategy that is relatively more robust to label sparsity. Inspired by Fréchet Inception Distance (FID) in computer vision [380], we explore the application of Fréchet Distance as a means to evaluate IR systems in settings where labeled data is scarce. Unlike conventional evaluation metrics that rely heavily on complete ground truth annotations, this approach measures the distributional similarity between retrieved and judged documents, offering a more resilient framework for assessing retrieval effectiveness in sparse-label environments.

By addressing these two aspects, this chapter provides a comprehensive examination of the impact of sparse labels on evaluation of IR systems when relevance judgments are sparse.

3.1 Shallow Pooling for Sparse Labels

For well over a decade, non-neural learning-to-rank methods have become firmly established in contexts where many non-content features are available, such as web search [59, 188]. Unfortunately, attempts to extend these methods to content-based ranking have had a mixed record unless substantial volumes of data training are available [379, 79]. However, the advent of transformer-based rankers has markedly advanced the state-of-the-art in many core information retrieval tasks [402]. Prior to 2018, a state-of-the-art ranker for these core document and passage ranking tasks might employ BM25 followed by a pseudo-relevance feedback method such as RM3, as typified by the open-source Anserini system¹ from the University of Waterloo [444]. By 2020, a state-of-the-art ranker might employ a dense retriever [437, 174, 201] followed by one or more transformer-based re-rankers [288, 290, 158, 320]².

¹<https://github.com/castorini/anserini>

²At the time of these experiments, LLM-based re-rankers were not yet available. Therefore, the following experiments are done with the state-of-the-art neural-based rankers of 2021.

Over this period, the IR community has tracked this progress on a number of leaderboards, most notably the MS MARCO³ leaderboards [236, 114]. Each test collection is based on a corpus of passages or documents, and comprises a set of training queries, a set of development (i.e., validation) queries, and a set of evaluation (i.e., test) queries. It is common among these test collections that with a few exceptions, each query has one known relevant item in the associated corpus. We call these labeled relevant items “qrels” for simplicity⁴. For the training and development sets these qrels are public. For the evaluation sets, the qrels are private. To establish a place on a leaderboard, research groups train and validate rankers with the training and development sets, run the evaluation queries to produce a ranked list for each query, and submit this run to the MS MARCO team.

Since there is often only one qrel per query, these leaderboards use mean reciprocal rank (MRR) as their tracking measure. In effect, a qrel represents the “right answer”, and rankers are evaluated on their ability to place this answer as close to the top as possible. This approach stands in contrast to the approach used for many traditional information retrieval evaluations, such as those conducted as part of the long running TREC evaluation exercise⁵. In a typical TREC experiment, top-ranked items from each submission are pooled for relevance judgments, so that measures such as NDCG may be applied. Unfortunately, even relatively shallow pooling (e.g., the top three items from each submission to compute NDCG@3) often requires dozens of judgments per query, limiting the number of queries for each experiment. In this section, we will discuss further that the method used to identify MS MARCO qrels supports our view that these qrels do not reflect relevance in the traditional sense, but instead only represent an attempt to identify an answer, and not necessarily the best answer.

Whithin our designed experiments, we observed that the top passages returned by neural rankers in the passage retrieval leaderboard of MS MARCO often appeared as good as, or even superior to, the qrels. We are not the only researchers to make this observation. The creators of RocketQA [320] write:

“...we manually examine the top-ranked passages (retrieved by our retriever) that were not labeled as positives in the original MS MARCO dataset, and we find that 70% of them are actually positives or highly relevant.”

³<https://microsoft.github.io/msmarco/>

⁴In established jargon, the word “qrel” describes any judgment, which could be relevant or non-relevant, graded or binary, multi-faceted, etc. Here, we use the term exclusively to mean an item judged *relevant* on a binary scale. MS MARCO collections do not contain explicitly non-relevant items, so that MS MARCO qrels always indicate a relevant item

⁵<https://trec.nist.gov>

Based on this observation, they trained a model with a dual encoder architecture that is able to perform cross-batch negative sampling, consequently decreasing the probability of selecting false negatives in the training step. Compared to other state-of-the-art dense retrievers, their proposed approach focuses on the importance of selecting negative samples in the training step. In part, their success depends on explicitly recognizing that an official “right answer” may not be the best answer.

3.1.1 Challenges in MS MARCO Relevance Labeling

Craswell et al. [114] provides a complete exposition on the background and status of the MS MARCO project addressing many common concerns regarding the MS MARCO datasets, including questions of internal validity, robust usefulness, and the reliability of statistical tests. In particular, Craswell et al. [114] address concerns raised by Ferrante et al. [135] who apply measurement theory to draw attention to important shortcomings of established evaluation measures, such as MRR. Many of these measures are not interval scaled, and therefore many common statistical tests are not permissible, and properly these measures should not even be averaged. These concerns are further addressed in a related paper by the same authors [236]. Here, we focus solely on the process employed to select the qrel for each topic. Craswell et al. [114] address this external validity concern as well, writing:

“...there could be quirks of the MS MARCO sparse labeling that pretrained transformer models can learn, giving good performance on MS MARCO sparse labels in the test set, but the improvements would vanish if we relabeled the data with slightly different judging scheme. In that case, the results would be specific to the setup of our study, lacking external validity. We could only claim a real improvement if we think real users have exactly the same quirks as the MS MARCO labels”

They argue that this concern is addressed by experiments conducted for the TREC Deep Learning Track [112, 111], which follows the typical methodology of a TREC evaluation exercise. In a traditional TREC-style evaluation exercise, the top items from each submission are pooled to a fixed depth, perhaps ten or more, and these items are individually judged by assessors with respect to a defined relevance scale. On the other hand, in MS MARCO queries for the collection were extracted from the logs of the Bing search engine. Top documents were retrieved by Bing and 10 candidate passages were automatically extracted from these documents. These passages were shown to an assessor as a set and the assessor identified a passage containing an answer. In some cases, two or more passages were selected. These passages became the qrels.

While nothing indicates that these assessors made comparisons between passages when selecting answers, nothing prevented these comparisons. Moreover, assessors were not required or encouraged to identify all passages containing an answer, nor does it appear that they were encouraged to identify the passage containing the best answer. As the authors indicate:

“there are likely passages in the collection that contain the answer to a question but have not been annotated as [such].”

As a result, the labels for MS MARCO datasets can not be treated as traditional relevance labels, which are assessed independently, nor are they complete.

Since it does not appear that the assessors were encouraged to identify the best answer, an important implication for MRR is left unrecognized. Rankers are rewarded for placing qrels as high as possible in the ranking, but if there are better answers in the collection, it becomes possible for a ranker to outperform the qrels by placing better passages above the qrels. In Section 3.1.2, we experimentally test this possibility. As the state-of-the-art improves and better answers are surfaced, we wish to first investigate the quality of these retrieved answers compared to the annotated qrels. Further, we tend to identify these improved answers and replace the qrels with them. At any point in time, a leaderboard should reflect the ability of rankers to place the best known answer at the top of the ranking. Apart from some evaluation exercises on named item finding and similar tasks (e.g., Metzler et al. [268]) evaluation by the best known answer is relatively rare. Navigational queries provide a major exception, where returning the desired URL in response to a navigational query remains a core function of Web search engines [55]. We note that since we are treating a qrel as a best known answer, we can employ MRR as our evaluation measure. Recent work from our research group has proposed other measures for preference judgments [104, 251] but for a single best answer, MRR remains appropriate [114].

3.1.2 Better Than Perfect

For each query, MS MARCO measures the performance of a ranker according to the rank at which it places a single known answer, or qrel, or one of small set of them. The higher the better. MRR@10 is then used to average over a set of queries.

In our experiments, we make side-by-side comparison between the top passages returned by a representative neural ranker and the qrels used to evaluate experimental runs. If the qrels were the best answers, we would expect a preference for qrels over other passages.

	Number of queries	% queries
Category A: qrel at first rank	1868	26.76%
Category B: non-qrel at first rank	5112	73.24%
Total	6980	100.00%

Table 3.1: Agreement between qrels and top results from our neural ranker. A hypothetical perfect ranker would always place the qrel first.

	Category A	Category B
qrel preferred	1,228	2,116
comparison passage preferred	640	2,996
Total	1,868	5,112

Table 3.2: Results of passage comparison. For Category A queries, the comparison passage is the second passage returned by the neural ranker. For Category B queries, the comparison passage is the top passage returned by the neural ranker.

Since both the qrels and our subsequent preference judgments depend on error-prone human assessments, this preference will not be universal, but we would certainly expect qrels to be preferred the majority of the time.

As a representative neural ranker, we employ the ranking stack described by Nogueira et al. [290] as provided by their GitHub repository. This ranking stack utilizes Anserini as a sparse first-stage retriever, which is followed by two BERT-based re-rankers, which they call MonoBERT and DuoBERT. The former was introduced by Nogueira and Cho [288] as a pointwise, second-stage re-ranker. The re-ranked list produced by MonoBERT feeds the third-stage DuoBERT pairwise re-ranker to generate the final ranking. We choose this ranking stack for our experiments for no other reason that we happened to be working with it when we observed the apparent superiority of the top passages it returned.

For our experiments, we split the queries in the development set as follows:

1. **Category A:** Queries for which the top passage and the qrel are the same.
2. **Category B:** Queries for which the top passage and the qrel are different.

While 94% of the queries in the development set have a single qrel, some have up to four. For these queries, we arbitrarily selected the first qrel in the file as the qrel used for this comparison. Alternatively, for these queries we could have selected the highest ranked

qrel returned by the ranker or conducted multiple comparisons. Our approach keeps the selection of the qrel independent of the ranker and avoids multiple comparisons on the same query.

Table 3.1 shows the split between the categories for the 6,980 queries in the development set. For the Category A queries, where the ranker and qrels agree, we would expect this passage to be generally preferred over other passages. For comparison purposes, we choose the second passage returned by our ranker. Even though this passage is highly ranked, we would still expect the top passage to be preferred when the two passages are placed side-by-side. For the Category B queries, we compare the top passage with the qrel. Altogether we have a pair for each query, which can be compared side-by-side by a human assessor to determine which represents the better result.

For assessment, we crowdsourced comparisons on the Amazon Mechanical Turk platform. To build tasks for crowdsourced workers, we grouped comparisons into sets of ten, which were combined with three manually constructed test comparisons, which helped to assure quality. These test comparisons were always between a known relevant passage (i.e., a qrel) and an obviously off-topic and non-relevant passage. Data from workers failing one or more of these test comparisons was excluded from our experiment and the task was repeated by a different worker. For each task we randomized the ordering of the pairs, as well as the right-to-left ordering of passages in the pairs. A full task comprises these thirteen pairs, which are presented one at a time to a worker along with the associated query. For each pair, the worker was asked to indicate the passage that “best answers the question.” Figure 3.1 illustrate a screenshot of the data annotation process for this experiment.

Method

The Mechanical Turk platform allows tasks to be restricted to workers that satisfy specified criteria. Since the MS MARCO passages and queries are taken from an EN-US environment, we restricted tasks to workers who were located in the US, with more than ten thousand approved tasks and an approval rate of at least 97%. In addition, we excluded workers who had previously failed the test comparisons for one our tasks. For each task containing 13 pairs, we paid \$2.00, plus a fee of \$0.40 to Amazon. Total cost was \$1,720, include pilot tests, debugging, and other minor costs.



Figure 3.1: Screenshot of the data annotation process.

This data labeling process was reviewed and approved by our institutional review board. After accepting a task, workers were shown a consent form and were required to provide consent before undertaking the task. Workers were given an option to exit the task at any point. Workers exiting before completing the task were paid a prorated amount. As required by our review board, the rate of pay for a task was estimated to provide compensation consistent with our local minimum wage. As part of the ethics review process we provided an example task. This example task was completed by the reviewer, who specified a minimum amount we could pay for the task. Assessed pairs (without identifying information) were approved for release after completion of the experiment. Examples of questions, qrels, passages, and preferences are provided in Figure 3.2.

Results

Figure 3.3 and Table 3.2 show the results of the crowdsourced passage comparison. For Category A queries, where the neural ranker returns the qrel as the top passage, the assessors agreed that this passage was preferable to the passage ranked second by the neural ranker roughly two-thirds of the time (18% vs. 9%). Given that this second-place passage is itself not unlikely to provide a reasonable response to the query, this outcome provides some assurance regarding both the crowdsourcing process and the quality of the ranker. If we assume that the qrel/top passage is in fact the best passage in the collection overall, this result also provides a rough estimate of crowdsourcing error at 33%.

i) **query 423878:** *is skin cancer genetic*

preferred qrel passage 7590792: This was a genetic study looking to identify new genetic origins of melanoma. Melanoma is the most serious type of skin cancer. Some of the risk of developing it may be influenced by the genes that you inherit, or your family history. The researchers estimate that around 10% of people with melanoma have one first-degree relative previously diagnosed.

non-qrel passage 7590794: New genetic clues about skin cancer. Skin cancer: Genetic mutations 'warn of risk'.

ii) **query 573954:** *what are the steps to a waltz dance*

preferred non-qrel passage 7403850: The basic step for waltz is a box step. It's named after a pattern it creates on the floor (box or square) and forms the foundation of the dance. A box step can be divided into two parts - a forward half box and a backward half box. Each half box has three steps - a step forward or backward, a step to the side, and a step to close the feet together.

qrel passage 7403851: Here is the basic waltz steps diagram for the leader. 1 Step forward with the left foot. 2 Right foot step sideways to the right. 3 Bring your left foot next to your right foot. Step back with the right 1 foot. Step back sideways with the left foot. Bring your right foot next to your left foot.

Figure 3.2: Two arbitrarily selected examples of preference judgments for the experiment reported in Section 3.1.2. These examples were selected from our logs without prior knowledge of the questions, passages, or outcome. In each case, we list the preferred passage first. In the top example, the preferred passage appears more complete than the other passage. In the bottom example, the preferred passage stands on its own, while the alternative refers to an unseen diagram and also only provides steps for the dance leader. All four passages could reasonably be judged relevant under traditional pointwise assessment.

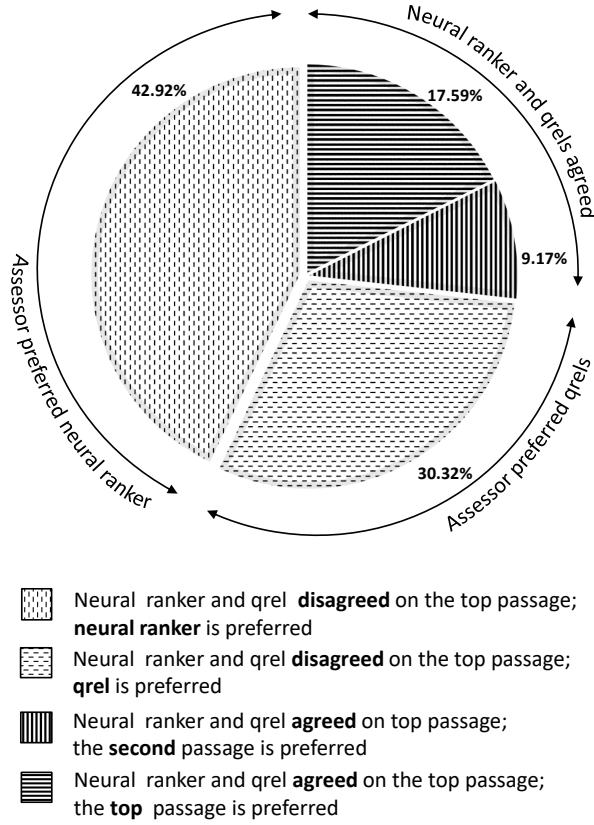


Figure 3.3: Results of passage comparison. When the neural ranker and the qrels agree on the top passage, crowdsourced assessors prefer it over the second passage from the neural ranker nearly 65.7% of the time. When the neural ranker and the qrels disagree, the crowdsourced assessors prefer the top passage from the neural ranker over the qrel for 58.6% of the pairs.

For the Category B queries, the top passage from the neural ranker was preferred over the qrel for nearly 59% of the queries (43% vs. 30%). Since the result on the Category A queries provides us with some confidence in the crowdsourcing process, this result suggests that the neural ranker is out-performing a hypothetical perfect ranker that always returned the qrel in the top rank. It is in this sense we claim that current neural rankers may already be “better than perfect” on the MS MARCO leaderboards. In the following sections we consider the potential impact of this result on our understanding of the the state-of-the-art.

Run Name	Leaderboard Rank	MRR@10				Reference
		Evaluation Set All queries	Development Set Original qrels All Queries	Development Set Original qrels Selected Queries	Development Set Preference qrels Selected Queries	
Perfect ranking		1	1	1	0.3320	
A	3	0.4190	0.4205	0.4013	0.5069	<i>not available</i>
B	4	0.4080	0.4213	0.3953	0.5638	Pradeep et al. [316]
C	5	0.4070	0.4210	0.3881	0.5827	Han et al. [158]
D	9	0.4010	0.4118	0.3863	0.5036	<i>not available</i>
E	10	0.4000	0.4074	0.3722	0.4955	<i>not available</i>
F	11	0.3990	0.4079	0.3697	0.5703	Hofstätter et al. [174]
G	12	0.3950	0.4046	0.3658	0.5199	Han et al. [158]
H	13	0.3940	0.3998	0.3606	0.5617	<i>not available</i>
I	30	0.3790	0.3904	0.3541	0.5464	Nogueira et al. [290]
J	34	0.3760	0.3880	0.3442	0.4985	Liu et al. [247]
K	81	0.3090	0.3181	0.2984	0.3663	Hofstätter et al. [173]
L	86	0.2940	0.3039	0.2779	0.4666	Zhan et al. [455]
M	92	0.2770	0.2901	0.2681	0.3251	Hofstätter et al. [173]
N	93	0.2720	0.2768	0.2550	0.3257	Nogueira et al. [289]
O	110	0.2180	0.2216	0.1955	0.2403	Nogueira et al. [291]
P	115	0.1860	0.1873	0.1536	0.2033	Yang et al. [443]

Table 3.3: MS MARCO passage ranking development runs used in our experiments, ordered according to the MS MARCO leaderboard at the time of writing. These were gathered from a variety of sources, as described in Section 3.1.3. The table lists MRR@10 values on several sets of qrels. The “Selected Queries” are described in Section 3.1.3. The “Preference qrels” are described in Section 3.1.4.

3.1.3 Shallow pooling

In the previous section, we compared the top documents returned by a single neural ranker to the qrels. When they disagreed, the top passage from the neural ranker was preferred more often. In this section we extend this comparison to encompass a set of top submissions from the leaderboard. For each query, we create a pool of candidate best answers by taking the top passage from each submission, plus the qrel. The passages in each pool are paired and judged by crowdsourced workers. We then directly compare runs in terms of their top documents.

Method

We base our experiments on the runs listed in Table 3.3. Like the previous experiment, this experiment uses the development set because qrels for evaluation queries are not publicly released. The leaderboard rank indicates position of a run as it was in 2021, when the research effort reported in this paper was first undertaken. The table lists official MRR@10 scores on both the development queries and evaluation queries, and the table is ordered by MRR@10 on evaluation set. The table also includes a preference MRR@10 using the qrels created by the process described in Section 3.1.4.

Since runs are not provided through the MS MARCO leaderboard, we created or obtained these runs from a variety of sources. In some cases, as referenced, the runs could be re-generated from code in github repositories and checked against the leaderboard. In other cases, the runs were provided by to us by the MS MARCO team for the purposes of this paper. We sincerely thank the MS MARCO team for responding positively to our “cold call” and providing those runs on the development set that were easily available to them. We focused our efforts to recover runs on the top of the leaderboard, and were able to recover four of the top ten and eight of the top 20.

We pooled the top passage from each run, plus the qrel if it did not appear among these top passages. For this experiment, we continue to use a single qrel per query, as described in the previous section. Figure 3.4 shows the size of these pools. For 117 queries the pool size is one, i.e., all rankers returned the qrel as the top document. The average pool size is 6.23 passages, with a median pool size of 6 passages. If we paired the documents in these pools and judged each pair once, it would require 141,887 comparisons and cost over \$34,000. Since this experiment is exploratory, rather than a complete evaluation of the runs themselves, we reduced our costs in two ways. First, we judged a random subset of 500 queries, rather than the full set of queries, as explained below. Second, we restricted judging to the available runs at the top of the leaderboard, i.e, runs A to K, which are more

likely to surface documents that outperform the qrels. These runs all have an evaluation MRR@10 between 0.367 and 0.419, while the next-highest run has an MRR@10 of 0.309. In addition, we included run L, which exhibited an interesting property on the pairs from the previous experiment. For this run, when both passages in a judged pair appeared in its top-10 ranking, they were ordered consistently with the preferences more often than any other run, with 63.07% agreement.

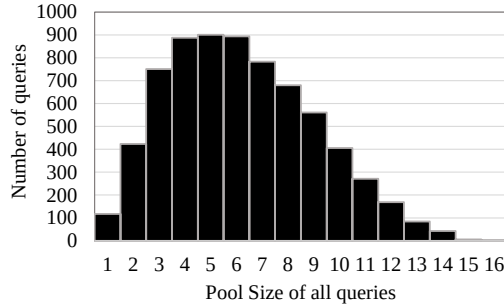


Figure 3.4: Sizes of preference judgment pools for all 6,980 queries in the MS MARCO passage retrieval development set pooled over all 16 available runs.

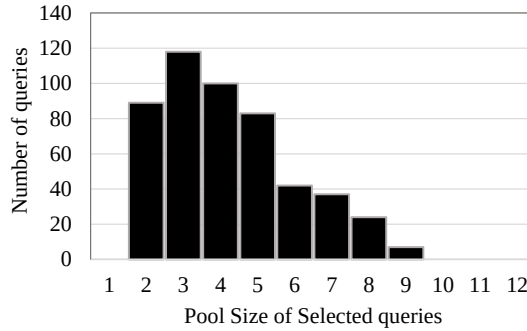


Figure 3.5: Sizes of preference judgment pools for 500 selected queries pooled over 12 selected runs.

We build new pools for this restricted set of runs, and then selected 500 queries at random for which the pool size was at least two. Figure 3.5 shows the size of these pools. The average pool size is 6.32 passages, with a median pool size of 4 passages. Sampling down to these 500 queries from the development set has no impact on the leaderboard ranking, as shown in Figure 3.6. This figure plots the MRR@10 on the evaluation set,

as shown on the leaderboard, against the MRR@10 on the 500 sampled queries from the development set. The values are shown in Table 3.3 in columns 3 and 5. The rankings are identical.

Judging each pair once requires 4,210 comparisons. We crowdsourced these comparisons on Mechanical Turk following the procedure described in the previous section, for an actual cost of \$1,022, including pilot tests, debugging, and other minor costs. Assessed pairs (without identifying information) were approved by our institutional review board for release after completion of the experiment. However, since these pairs include passages surfaced by runs provided to us by the MS MARCO team strictly for this experiment, their general release is not possible.

Results

Just as we did in Section 3.1.2, we directly compare pairs of runs with the preference judgments. For each pair of runs, we compute a *win ratio* for the top passages from each run. When the top documents differ, the win ratios indicates how often one run is preferred over another.

The results are shown in Table 3.4. Win ratios compare the columns to the rows, so that a value above 50% indicates that the run in column beats the run in row more often than not. Bolded numbers indicate significant differences under a binomial test with $\alpha = 0.05$, after a Bonferroni correction. Even under the conservative Bonferroni correction three runs significantly outperform a perfect ranking under the official qrels. The last row indicates the number of wins for the run in that column. Run H is particularly interesting since it wins against all other runs. Unfortunately, we know little about this run, since there is no associated reference.

	Perfect	A	B	C	D	E	F	G	H	I	J	L
Perfect		49.3%	60.8%	58.6%	51.4%	49.7%	59.0%	55.8%	58.7%	55.6%	51.8%	51.0%
A	50.7%		55.2%	58.0%	48.4%	45.7%	53.1%	53.8%	55.3%	52.7%	48.6%	45.6%
B	39.2%	44.8%		51.5%	44.9%	44.4%	49.0%	48.1%	52.0%	51.2%	45.1%	41.1%
C	41.4%	42.0%	48.5%		42.4%	42.5%	49.2%	34.5%	51.9%	47.9%	39.8%	40.5%
D	48.6%	51.6%	55.1%	57.6%		46.4%	54.6%	53.1%	55.8%	52.1%	50.6%	46.7%
E	50.3%	54.3%	55.6%	57.5%	53.6%		54.2%	51.5%	54.8%	51.1%	47.6%	48.1%
F	41.0%	46.9%	51.0%	50.8%	45.4%	45.8%		44.3%	55.6%	49.8%	42.5%	42.0%
G	44.2%	46.2%	51.9%	65.5%	46.9%	48.5%	55.7%		56.3%	55.8%	44.1%	43.3%
H	41.3%	44.7%	48.0%	48.1%	44.2%	45.2%	44.4%	43.7%		47.5%	43.8%	42.0%
I	44.4%	47.3%	48.8%	52.1%	47.9%	48.9%	50.2%	44.2%	52.5%		44.1%	42.2%
J	48.2%	51.4%	54.9%	60.2%	49.4%	52.4%	57.5%	55.9%	56.3%	55.9%		49.7%
L	49.0%	54.4%	58.9%	59.5%	53.3%	51.9%	58.0%	56.7%	58.0%	57.8%	50.3%	
Wins	2	4	8	10	3	2	8	6	11	8	3	1

Table 3.4: Win ratios when runs are directly compared according to their top documents. Percentages indicate the frequency that the run in the column beat the run in the row. Bold numbers indicate significant differences under a binomial test with $\alpha = 0.05$, after a Bonferroni correction. Under this measure, three runs would outperform a perfect ranking under the official grels. The last row indicates the number of times the run won. Runs are ordered according to their position on the official leaderboard. ‘Perfect’ represent the Perfect ranking obtained by grels.

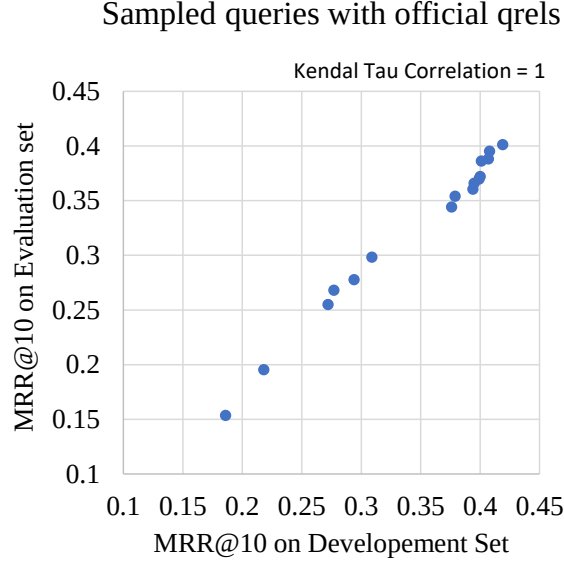


Figure 3.6: MRR@10 on Evaluation set vs Development set with Official qrels on 500 sampled queries. The rankings are identical suggesting that results on the 500-query sample can be generalized to the development set leaderboard.

3.1.4 Leaderboard impact

MS MARCO queries can have as little as one qrel, i.e., only one known relevant item. We hypothesize that the MS MARCO evaluation methodology, and its associated leaderboards, depend on this item being the best item, rather than just any relevant item. Employing MRR as the primary evaluation measure makes the implicit assumption that placing this item above all other items is always the correct thing to do. The experiments in previous sections contradict this assumption, demonstrating that some runs on the leaderboard can be considered “better than perfect” according to the existing qrels. In this section, we explore the possibility that these experimental results raise practical concerns, with the potential for impacting leaderboards and our notion of the state-of-the-art.

In this section, we derive new preference-based qrels from the preference judgments described in the previous section. Using these preference qrels, we compute new MRR@10

values for the 16 available runs. We compare the results with MRR@10 values computed using the original qrels.

Method

We convert preference judgments to qrels by treating the preference judgments as a tournament. If a single passage wins the most pairings, we designate that passage to be the “preference qrel” for that query. For queries where multiple passages were tied for first place, we eliminate the losing passages and repeat the process with the first-place passages, until we have a single “preference qrel”. For 46 of the 500 selected queries, it was not possible to designate a single “preference qrel” due to a cycle between three passages. Overall we have 592 qrels in the preference qrel set for the 500 selected queries. The preference qrels win 1598 or 91.8% of their pairings, while the original qrels win 809 or 46.6% of their pairings. For the remainder of the experimental results in this section, we use all of the original qrels, not just one for each query. For the 500 selected queries, there are 528 qrels in this original qrel set.

Results

The results are given in Figure 3.7. For these results, we use all 16 runs available, including runs K and M-P, which were not included in the pools for the previous experiment. These runs all placed below the top 80 on the leaderboard at the time our experiments were completed in May 2021.

The first graph in Figure 3.7 plots MRR@10 with the preference qrels vs. MRR@10 with the original qrels. Although there is a correlation between the two measures over the plot as a whole, with Kendall’s $\tau = 0.65$, at the top rungs of the leaderboard, the relative order of the runs changes dramatically. The green dashed line indicates the MRR@10 of the perfect ranking under the official qrels, which scores $\text{MRR@10} = 0.3320$ on the preference qrels.

The bottom plots order the runs according to the official leaderboard ranks, and show 95% confidence intervals. On the official qrels (second plot) the runs adhere closely to the leaderboard ranking. On the preference qrels, there are noticeable changes in the ranking. In particular, run A drops below runs B and C, while runs D and E drop below F.

Run C is the best performing run on the preference qrels, but third on the original qrels. This run is described by Han et al. [158] who argue that finetuning a classification model with the aim of deciding whether a document is relevant to a query or not, is not a

suitable approach for a ranking task. Instead, they employed a learning-to-rank algorithm on a pair-wise and list-wise basis which learns to differentiate relevance for document pairs or optimize the list as a whole, respectively. Unlike most BERT-based methods, this framework builds a LTR model through fine-tuning representation of query-document pairs and demonstrates the potential of combining ranking losses with BERT representations, especially for passage ranking. Since this run takes a very different approach than most runs at the top of the leaderboard, we hope that a promising direction for future progress has not been missed.

On the preference qrels, run L (Zhan et al. [455]) noticeably improves against runs with a similar MRR@10 on the original qrels. As mentioned previously, this run exhibited an interesting property on the pairs from the previous experiment. When both passages in a judged pair appeared in its top-10 ranking, they were ordered more consistently with the preferences more frequently than any other run. To check that this relative improvement was not an artifact of the run’s inclusion in the pool, we re-calculated the preference qrels without the passages it contributed. While the exact numbers, changed, run L continued to show this relative improvement.

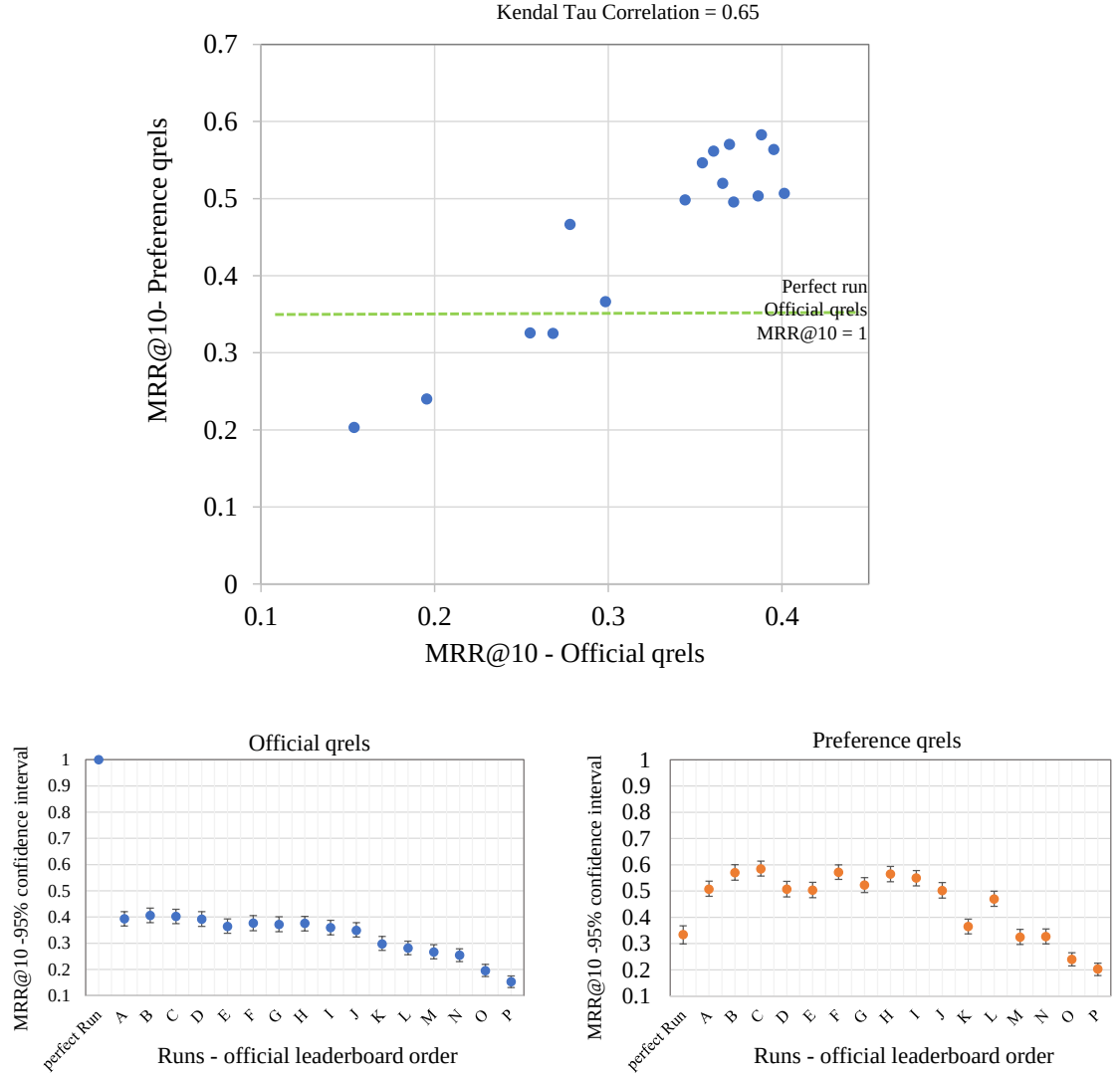


Figure 3.7: Shallow pooling of top documents from available MS MARCO passage retrieval runs followed by crowdsourced preference judgments to create new qrels for a sample of 500 queries. A perfect ranking under the official qrels, with an MRR@10 of 1, performs poorly under these new preference qrels. Changes in the order of top runs raise concerns about the ability of the official qrels to recognize improvement in the state of the art. The bottom plots show 95% confidence intervals. Runs are ordered according to the official leaderboard.

3.1.5 Takeaway

The findings in this section highlight a crucial challenge in retrieval evaluation with sparse labels: state-of-the-art rankers may retrieve passages that are superior to sparsely judged relevant items, yet their performance can be underestimated due to the limitations of traditional evaluation metrics like MRR. As retrieval models continue to improve, they surface increasingly high-quality answers, potentially displacing existing relevant judged in rankings and leading to paradoxical results where better models appear to perform worse under the official evaluation framework. This issue underscores a broader limitation of evaluation methodologies that rely on fixed, sparse labels.

We provide evidence that the performance of modern neural rankers may already exceed the performance of a hypothetical perfect ranker on the MS MARCO passage ranking leaderboard under the official qrels. If we imagine a hypothetical perfect ranker that always returned a judged relevant passage at rank 1, achieving an MRR@10 of 1, the top passage from the neural ranker would be preferred more often, making the neural ranker in this sense “better than perfect”. This outcome has implications for measuring further improvements to the state-of-the-art. To provide some sense of these implications, we pooled the top passage from available experimental runs employed crowdsourced workers to make pairwise preference judgments between them. Again, we see that the results of neural rankers can be “better than perfect”. When we construct a new set of qrels from these judgments, we see shifts in the leaderboard which raise concerns about its ability to continue tracking improvements.

To address these concerns, we advocate for an adaptive evaluation framework that continuously updates relevance labels using comparative preference judgments rather than relying solely on static pointwise judgments. Comparative judgments allow finer-grained distinctions between highly relevant passages and have been shown to improve assessor agreement while reducing judgment overhead. Our experiments suggest that modern rankers may already outperform an idealized “perfect” ranking based on existing qrels, further reinforcing the need for a dynamic evaluation paradigm.

Looking forward, maintaining the relevance and reliability of leaderboards such as MS MARCO will require regular updates to the judgment set. This can be achieved through continuous pooling and reassessment of top-ranked passages, ensuring that evaluations remain aligned with the evolving state of retrieval technology. By adopting more flexible and adaptive assessment strategies, we can better ensure that future improvements in ranking models are accurately measured and recognized. The data collected for our experiments is available for research purposes at <https://github.com/Narabzad/Shallow-Pooling-for-Sparse-labels>.

3.2 Fréchet Distance for Evaluation of IR systems with Sparse Labels

In the previous section, we showed that with the rapid advancement of technologies in fields such as IR and natural language processing, evaluating the performance of these systems is becoming increasingly challenging [147, 165, 72, 442]. We highlighted that new metrics, benchmarks, and evaluation protocols must be developed that are specifically tailored to the unique characteristics of the systems considering the rapid changes in system architecture, training data, and model configurations [387]. However, as mentioned earlier, in many cases, obtaining high-quality labeled data that accurately represents the complexity of real-world scenarios can be expensive, time-consuming, or even impractical. This scarcity of labeled data adds to the limitations of conducting extensive evaluations and may lead to biased or incomplete assessments [23].

Given the sparsity of ground truth labels, it is crucial to recognize the challenges involved in distinguishing between rankers when the differences in performance are small [441]. The limited labeled data for retrieved documents introduces noise, making it challenging to definitively determine which ranker is performing better [63]. The incomplete judgments can introduce problems in evaluations, as they do not capture the full range of relevant documents [32, 68]. This issue becomes even more pronounced in generative-based tasks. It is impractical to reassess the generated results, such as images or text, with each system run due to their non-deterministic nature [387, 162].

Evaluating a generative system’s performance based on the similarity of generated content to sparsely labeled data remains one of the most effective approaches in many generative-based NLP and computer vision benchmarks and tasks [369, 171, 295, 126, 459]. Particularly in the evaluation of text-to-image generation task, the Fréchet Inception Distance (FID), has gained recognition for showing high robustness and correlation with human judgments [171, 339, 452]. FID compares the distribution of generated images across a set of prompts to the distribution of target images across the same set of prompts. To compute FID, features of ground truth images and generated images are extracted from both sets, and multivariate Gaussian distributions are fitted to these features. The Fréchet Distance (FD), which quantifies the similarity between two probability distributions, is then computed based on the fitted Gaussian distributions. A lower FID score indicates a higher similarity between the distributions, indicating that the generated images closely match the real images in terms of their visual features.

In this section, we shed light on how evaluating generated results is similar to assessing the quality of retrieved results with sparse labels in an ad hoc retrieval setting. Most

benchmarks for both tasks have quite sparse labels i.e., there can be other unjudged relevant items available. While labeling more data is expensive for both tasks, there could be more than one correct answer in both tasks. In this work, we mimic an IR system with sparse relevance judgments as a generation task where the ground truth targets are sparse. Due to the success of FID in evaluating the quality of generated images, especially for generative adversarial networks [143, 339, 452, 200, 11], we explore if we can quantify the quality of retrieved documents in an ad hoc retrieval system through *Fréchet Distance*. In the context of IR evaluation, we can analogously consider the relevant judged items as the ground truth set and the retrieved items as the set of generated items. Our objective is to extract features from both sets, the relevant judged items and the retrieved results, and investigate whether metrics such as the Fréchet Distance can effectively capture the quality of the retrieved results with respect to the ground truth labels in IR systems.

In this study, we investigate the following Research Questions:

- RQ1. Can the Fréchet Distance effectively evaluate IR systems with sparse labels?
- RQ2. Can the Fréchet Distance effectively evaluate IR systems with comprehensive labels?
- RQ3. Can the Fréchet Distance effectively evaluate the quality of IR systems when the retrieved results are not labelled?
- RQ4. How well correlated are the performance of IR systems, as measured by the Fréchet Distance vs. and traditional IR metrics?
- RQ5. How robust is the Fréchet Distance for evaluating IR systems with respect to the feature extraction methods used to represent both the ground truth and retrieved items?

Our experiments in this section demonstrate the effectiveness of the Fréchet Distance as a metric for quantifying the performance of IR systems especially when the ground truth labels are sparse.

3.2.1 Fréchet Distance

The Fréchet distance is a measure of dissimilarity between two curves or trajectories and has shown to be useful in numerous applications including computational geometry, computer graphics, bioinformatics and robotics [12, 14, 15, 185, 148]. To understand the Fréchet

distance, let us consider two curves (or trajectories or paths): A and B . The Fréchet distance between A and B could be exemplified as measuring the minimum leash length required by a dog walking along a path A while its owner walks along path B , with both the dog and owner potentially traversing their respective paths at different speeds [13, 129]. The leash cannot be shortened or lengthened during the walk. The definition is symmetric i.e., the Fréchet distance would be the same if the dog were walking its owner. Given two curves, A and B , represented as sequences of points in a metric space, the Fréchet distance, denoted as $F(A, B)$ is computed as:

$$F(A, B) = \inf_{\alpha, \beta} \max_{t \in [0, 1]} d(A(\alpha(t)), B(\beta(t))) \quad (3.1)$$

where A and B are continuous maps from $[0, 1]$ to metric space and α and β are reparameterizations of the unit interval $[0, 1]$ i.e. they are continuous, non-decreasing, surjection functions. The requirement of non-decreasing reparameterizations, α and β , ensures that neither the dog nor its owner can backtrack along their respective curves. The parameter t as represents the progression of time, consecutively $A(\alpha(t))$ and $B(\beta(t))$ represent the position of the dog and the dog's owner at time t (or vice versa). The distance d between $A(\alpha(t))$ and $B(\beta(t))$ corresponds to the length of the leash between them at time t . By considering the *infimum* over all potential reparameterizations of the unit interval $[0, 1]$, we select the specific paths where the maximum leash length is minimized.

Apart from quantifying the dissimilarity between curves, the Fréchet distance can also serve as a measure to assess the disparity between probability distributions [171]. Given we have two normal univariate distributions, X and Y , Fréchet Distance (FD) is given as:

$$FD(X, Y) = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 \quad (3.2)$$

Where μ and σ are the mean and standard deviation of the normal distributions, respectively.

Fréchet Inception Distance

In computer vision, the Inception V3 model pre-trained on the Imagenet dataset is employed to generate feature vectors to be approximated by multivariate normal distribution [380]. As such, the Fréchet Inception Distance (FID) for a multivariate normal distribution is computed as:

$$FID(X, Y) = \|\mu_X - \mu_Y\|^2 - Tr(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}) \quad (3.3)$$

In this equation, X and Y represent two distributions derived from two sets of embeddings. These embeddings correspond to real images and generated images, respectively,

and are obtained from the Inception model. It is assumed that both distributions follow a multivariate normal distribution. The vectors X and Y have magnitudes μ_X and μ_Y , respectively. The trace of the matrix is denoted as Tr , while Σ_X and Σ_Y represent the covariance matrices of the vectors.

Adopting Fréchet Distance for IR Evaluation

Let us assume C is a collection of items and $Q = \{q_1, q_2, \dots, q_n\}$ is a set of n queries, where each query q_i has a set of relevant judged items R_{q_i} . We define R_Q as a set of relevance judged items for queries in Q , where $R_Q = \{d | d \in R_{q_i}, q_i \in Q\}$. Furthermore, we can obtain the top- k retrieved items by a retrieval system M from C for a given query q as $M_k(q, C) = D_q^k$, where D_q^k is a set of the top- k most relevant retrieved items for query q , i.e., $D_q^k = \{d_1^q, d_2^q, \dots, d_k^q\}$. Given $\mathbb{V}$ as a function that maps any retrieved item to a p -dimensional embedding space, where p is usually in the order of a few hundred, we can embed all the retrieved items and relevant judged items through $\mathbb{V}$. For instance, $\mathbb{V}(d_1)$ returns a p -dimensional vector embedding for document d_1 . To apply Fréchet Distance for assessing the quality of the IR system M , we measure FD_Q^M as follows on query set Q :

$$FD_Q^{M_k} = FD\left(\{\mathbb{V}(R_Q)\}, \{\mathbb{V}(M_k(Q, C))\}\right) \quad (3.4)$$

Here, FD is the Fréchet Distance (Eq. 3.3) measures the distance between the distribution of the set embeddings of the relevant judged items $\{\mathbb{V}(R_Q)\}$ and those of the retrieved items $\{\mathbb{V}(M_k(Q, C))\}$. The lower $FD_Q^{M_k}$ represents the retrieved items to have higher similarity with the relevant judged items and thus the better performance of the retrieval system M on the query set Q .

3.2.2 Experimental Setup

Dataset and Query sets

We perform experiments on the MS MARCO passage retrieval collection V1 and the 6980 queries in MS MARCO small dev set, which are sparsely labelled. Second, in Section 3.2.4, we experiment on the TREC Deep Learning (DL) track 2019 [112] and 2020 [113] to study how varying and extending the relevance judgments would affect the evaluation process. We compare the results of the FD score with the official traditional IR evaluation metrics of each benchmark, i.e., MRR@10 for MS MARCO and nDCG@10 for TREC Deep Learning tracks.

Retrieval models

To conduct experiments on MS MARCO dev set, we consider a set of 12 retrieval methods that are well-distinguished for their efficiency or effectiveness, ranging from traditional high-dimensional bag-of-word sparse retrievers to more recent dense retrievers well as trained high-dimensional sparse models, which are representative of novel retrieval methods developed over the past five years. Specifically, we consider BM25 as the representative of the sparse retrievers standalone as well as applying BM25 to expanded documents through DeepCT and DocT5Query document expansion methods [336, 289, 291, 120]. We include a set of dense retrievers including RepBERT [455], ANCE [435], Sentence-BERT (SBERT) [331], COLBERT [201] and COLBERT-V2 [350]. We also employ the more recently proposed high dimensional learnt sparse retrievers, UniCOIL and SPLADE [138, 234]. Furthermore, we consider hybrid retrievers [239] that fuse the retrieved items from BM25 and dense retrievers, to cover a variety of retrievers and assess the ability of FD to quantify the quality of retrieval fairly. We note that we employ some of the retrieval models from Pyserini⁶ [237] and some of the others from the paper’s original GitHub repository. For more information about each of the retrieval models, we kindly refer to the original papers of each method.

For our experiments with the TREC DL19 and DL20 query sets [112, 113], we took the submitted runs for each track from the TREC NIST website. Our experiments compare the results when assessing with Fréchet distance as well as nDCG@10 for 37 submitted runs to TREC DL2019 and 59 submitted runs to TREC DL 2020. These runs cover a comprehensive set of retrieval pipelines, typically with from sparse and/or dense retrieval as a retrieval first stage followed by one or more neural re-ranking stages.

Embeddings

To examine the robustness of FD on IR systems, we perform experiments using two different types of transformer-based contextualized models to embed the documents and extract their features. We employ a general-purpose DistilBERT [349] to obtain the documents embeddings⁷ as well as fine-tuned pre-trained language models on MS MARCO⁸ [331]. Both models were adapted from hugging face. We note that unless we explicitly mention (Section 3.2.7) all the results are reported with the first model, i.e., the DistilBERT model

⁶<https://github.com/castorini/pyserini>

⁷<https://huggingface.co/distilbert/distilbert-base-uncased>

⁸<https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v2>

Table 3.5: Performance of different retrievers in terms of MRR@10 as well as Fréchet distance FD on MS MARCO dev set. A smallest Fréchet distance corresponds to better performance.

Category	Method	MRR@10	$FD@1$	$FD@10$
Sparse	BM25	0.187	7.446	4.410
	DeepCT	0.242	1.453	2.354
	DocT5	0.276	3.047	2.050
Dense	RepBERT	0.297	1.881	1.223
	ANCE	0.330	1.529	0.995
	SBERT	0.333	1.387	1.008
	ColBERT	0.335	1.456	0.980
	ColBERT V2	0.344	1.453	0.982
Trained	UniCOIL	0.351	1.387	0.980
Sparse	SPLADE	0.368	1.328	0.964
Hybrid (BM25)	ColBERT-H	0.353	1.494	0.973
	ColBERT V2 -H	0.368	1.464	0.998

that was fine-tuned on MS MARCO. We believe that by exploring different document representations, we may better understand the influence of document quality on the utilization of FD for evaluating IR systems.

3.2.3 Assessment with Sparse labels

We are interested in investigating how FD can assess the performance of different retrievers when there are only sparse labels available i.e., on 6980 queries from MS MARCO small dev set. We present the performance of the 12 retrieval methods in terms of MRR@10 as well as measuring the Fréchet Distance between two sets of retrieved items and relevant judged items on the cut-offs of 1 and 10 in Table 3.5.

The results for $FD@1$ and $FD@10$ demonstrate the ability of FD to quantify the performance of retrievers. For example, for the BM25 retriever, $FD@1$ is measured as 7.446 and $FD@10$ as 4.410. However, for a neural retriever like ColBERT, which has shown superior performance to BM25 on various benchmarks [350, 201, 386], the FD values are reported as 1.456 and 0.980 for $FD@1$ and $FD@10$, respectively. This indicates that FD can effectively pickout the *better* retriever, particularly when there is a significant difference between their performances. On the other hand, when the performance of two retrievers is quite similar,

such as in the case of ColBERT vs. ColBERT V2, it becomes more challenging for evaluation metrics to assess their performance. For instance, while MRR@10 for ColBERT vs. ColBERT V2 is reported as 0.334 vs. 0.343, $FD@10$ for the two retrievers is reported as 0.980 and 0.982. Therefore, as expected, the discriminative power of FD decreases when it becomes harder to distinguish between retrievers. However, It is important to acknowledge that due to the sparsity of ground truth labels, previous research has indicated that distinguishing between rankers becomes challenging when the differences are small. In such cases, the noise introduced by limited labeled data for retrieved documents makes it difficult to definitively determine which ranker is performing better [320]. In fact strong rankers might have surfaced other *unjudged relevant items*. Based on the results in Table 3.5 and their comparison with MRR@10, we can conclude that in response to **RQ1**, we observe that Fréchet Distance can effectively evaluate IR systems.

To examine the robustness of the FD in the context of IR assessment, and to evaluate the generalizability of the method across different subsets of queries, we employ a bootstrap sampling [189, 128] from the MS MARCO dev set for $N = 1000$ times. This would allow us to investigate whether the results obtained in the previous section were influenced by the data or if they can be reliable. The results are visualized in Figure 3.8, in which we present the mean and empirical 0.95% confidence interval for each retriever across the 1000 query sets in terms of MRR@10 and $FD@10$. It is important to note that for the MRR plot, a higher position on the plot indicates better performance, while for the FD plot, a lower position indicates better performance. The findings confirm that despite considering different sample sets, we observe a consistent pattern and similarity in the performance trends.

3.2.4 Assessing with Comprehensive labels

In this section, we investigate the performance of the Fréchet Distance in evaluating IR systems when the labels are not sparse and we have more complete labels. We conduct experiments using the runs submitted to TREC DL 2019 (37 runs) and TREC DL 2020 (59 runs). The queries in TREC DL tracks on average have over 210 judged documents per query assessed with four different levels of relevance including “not relevant”, “related”, “highly relevant”, and “perfectly relevant” [112]. We notice that the number of judged relevant items per query in these benchmarks varies a lot. Due to the TREC-style judgment criteria, only the top few retrieved items from all submitted runs were judged. Depending on the overlap between the top retrieved items from different runs, the number of relevant judged items per query may vary. When applying FD with an imbalanced number of relevant judged items per query, it can introduce biases in the ground truth distribution

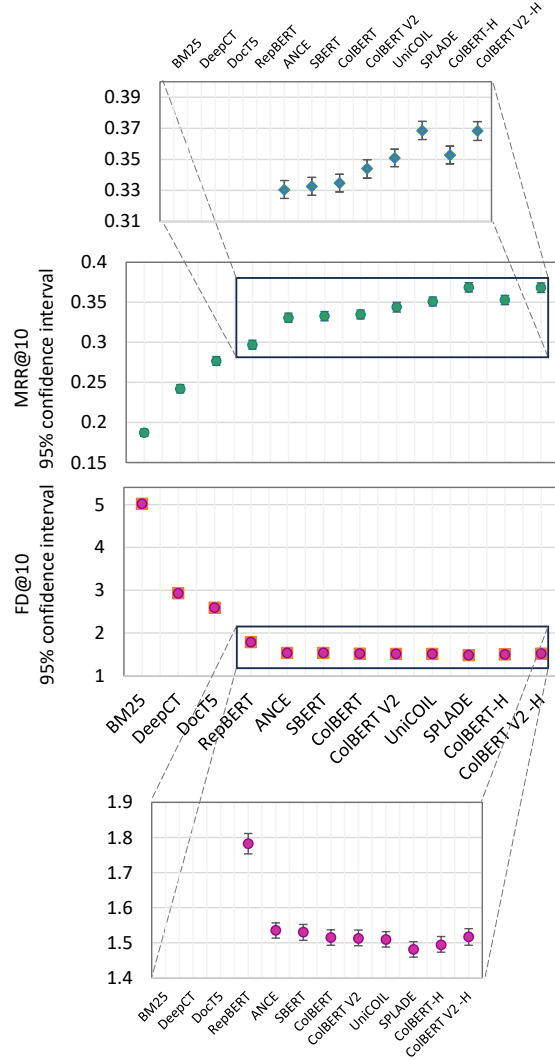


Figure 3.8: Performance of bootstrap sampling (N=1000) of queries in MS MARCO dev set in terms of MRR@10 and $FD@10$ for the 12 different retrieval methods.

and potentially lead to problems in evaluation. To address this issue, we balanced the number of relevant judged items per query by limiting them to a maximum of 1, 5, and 10 relevant judged items per query i.e., we randomly select K relevant items from the pool of relevant judged documents for the query of interest. We first randomly select from the most relevant level i.e., level 3 which are perfectly relevant documents and then when there is not a sufficient number of perfectly relevant documents, we move on to highly relevant

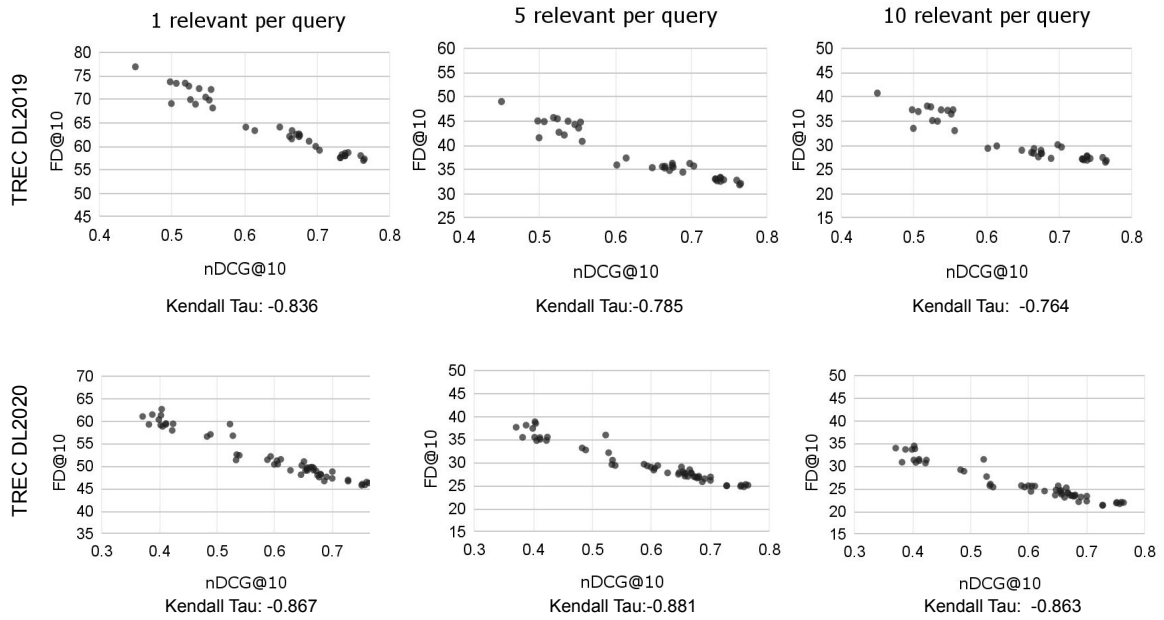


Figure 3.9: Performance of all the submitted runs to TREC DL 2019 (first row) and TREC DL 2020 (second row). In each sub-figure, X-axis and Y-axis indicate $nDCG@10$ and $FD@10$ respectively. $FD@10$ was measured with 1,5 and 10 relevant items per query in the sub-figures in the first, second and third columns respectively.

Dataset	10 qrels	5 qrels	1 qrel
Trec-DL-2019	0.796	0.784	0.594
Trec-DL-2020	0.918	0.891	0.863

Table 3.6: Kendall Tau correlation between nDCG measured with full relevance judgements and sparsified relevance judgements.

level and randomly choose from that grade. This experiment also allows us to examine how the sparsification of judgments affects the performance of evaluation metrics. We note that these modifications in relevance judgments are only applied for measuring FD and nDCG@10 is measured with all the judged documents without any modification.

We plotted the nDCG@10 on the x-axis and the FD with balanced and sparsified judgments on the y-axis of each sub-figure in Figure 3.9, for all the runs submitted to TREC DL19 (first row) and TREC DL20 (second row). Consistent with our previous experiments, we observe a highly linear relationship between the two metrics. We also provide the Kendall τ correlation under each sub-figure. For instance, when sparsifying the labels and considering only one relevant judged item per query, we obtain a Kendall τ correlation of -0.836 for TREC DL2019 and -0.867 for TREC DL2020, between nDCG@10 and $FD@10$ of each dataset. We note that the correlation here is negative because, in MRR, higher values indicate better performance, whereas in FD , lower values indicate better performance.

In addition, we present the Kendall Tau correlation between nDCG when using full relevance judgments versus randomly selecting a maximum of N relevant judgments, where N could be 1, 5, or 10, as illustrated in Figure 3.9. It is worth noting that while FD (as demonstrated in Figure 3.8) exhibits a higher degree of robustness when evaluated with sparse labels, nDCG is not as resilient concerning the chosen relevant judged document (qrel). This is because FD computes its metrics over the distribution of all queries, contributing to a more stable evaluation performance. On the contrary, NDCG with sparse labels tends to be considerably noisy and heavily dependent on which document is selected as the “one relevant document” per query, leading to significant variations in the results. In the Table 3.6, we present the Kendall Tau correlation between nDCG with full relevance judgments and nDCG when choosing 1, 5, or 10 random relevant documents. These results highlight the sensitivity of nDCG to the choice of relevant documents, especially when only a limited number of relevant documents are considered.

These experiments highlight two key points. First, unlike using the Fréchet Inception Distance to evaluate the quality of generated images in text-to-image generation tasks,

where a large number of data points (in the order of thousands) are required for the evaluation to be valid, we demonstrated that even with a smaller number of queries (around 40-50), FD is capable of distinguishing the performance of different rankers [215, 171]. Second, FD is not sensitive to the sparsity of the ground truth labels and it performs well with both sparse and more complete labels. It is not affected by the number of judgments, as evidenced by the fact that the performance did not differ greatly when increasing the number of relevant judged items. However, for TREC DL2019, we observed a small drop in correlation by increasing the number of relevant judgments. Further exploration revealed that a higher number of relevant judgments in TREC 2019 resulted in a higher usage of level 2 relevance judgments (highlight relevant) instead of level 3 judgments (perfectly relevant). Consequently, we suggest that FD may be more sensitive to the quality of relevant judged items rather than the quantity. Overall, in response to **RQ2**, we find that FD works well when using comprehensive labels, and consistent with the findings in Section 3.2.3, sparsifying the labels does not compromise the quality of assessment.

3.2.5 Assessing Unlabeled Retrieved Results

Here, we undertake an evaluation of different IR systems under an extremely challenging case of assessing unlabeled retrieved results. This scenario presents a situation where each query is assumed to have mostly only one relevant item, and the *relevant judged items are not included in the top- k results*. Our objective is to investigate the effectiveness of the Fréchet Distance in assessing the top- k Unlabeled Retrieved Results (URR) when no judgments are available for any of the top- k retrieved items. This is particularly valuable considering the high cost and limited availability of labeled data, which often exhibit sparsity. Previous research has demonstrated that as rankers improve in performance, they tend to retrieve previously unseen content that may be highly relevant to the original query [23]. If Fréchet Distance is capable of evaluating the retrieved results in such cases, it would be a valuable tool for assessing the relevance of unlabeled data and even beyond that, for evaluating generative-based responses.

We measure the FD between one set consisting of the relevant judged items per query and the other set consisting of the top- k *unjudged* retrieved item for each query. In other words, we scan down the ranked list and retain the first k unjudged document to assess. This is an interesting aspect to study because traditional IR metrics such as MRR, nDCG, and MAP rely on the presence of relevant items in the retrieved list and would assign a performance score of zero in cases where no relevant items are retrieved. They do not account for unjudged documents. We argue that by utilizing the FD metric, we can capture

Table 3.7: Performance of different retrievers in terms of MRR@10 as well as Fréchet distance FD assuming under Unlabeled Retrieved Results (URR) setting. We note that the MRR@10 is measured on the original ranked list since with URR setting, all the retrievers would obtain MRR@10 equals to zero. A smallest Fréchet distance corresponds to better performance.

Category	Method	MRR@10	URR	
			$FD@1$	$FD@10$
Sparse	BM25	0.187	8.634	4.705
	DeepCT	0.242	4.183	2.591
	DocT5	0.276	4.066	2.290
Dense	RepBERT	0.297	2.701	1.364
	ANCE	0.330	2.353	1.126
	SBERT	0.333	2.266	1.156
	ColBERT	0.335	2.308	1.115
	ColBERT V2	0.344	2.352	1.121
Trained	UniCOIL	0.351	2.302	1.128
Sparse	SPLADE	0.368	2.300	1.117
Hybrid (BM25)	ColBERT-H	0.353	2.399	1.115
	ColBERT V2 -H	0.368	2.365	1.142

the similarity between unjudged retrieved items and the limited set of judged examples and measure the performance of the retriever based on this value.

The results of this experiment are reported in Table 3.7 with two cut-offs of “ $FD@10$ ” and “ $FD@1$ ”. Even when no judged documents appear in the top- k , FD is still able to quantify the performance of the retriever. This capability is not present in traditional metrics. For instance, when there are no relevant judged items retrieved in the ranked list, $FD@1$ quantifies the performance of BM25 as 8.634, whereas the performance for ColBERT is measured as 2.308. This indicates that even without relevant judged items, FD is capable of determining that ColBERT performs better than BM25.

This experiment demonstrates that, unlike traditional IR metrics, FD is not sensitive to the labeled documents themselves. Indeed, the Fréchet Distance is not reliant on the exact positioning of the relevant judged document in the ranking. Instead, it focuses on measuring the similarity between the retrieved items and the relevant judged documents. This characteristic makes it particularly valuable for evaluating scenarios with extremely sparse labels, even in cases where the rankers do not retrieve the labeled data. In response to **RQ3**, the Fréchet Distance enables assessment of the remaining unlabeled data, offering

Table 3.8: Kendall τ correlation between different evaluation metrics over the 12 retrieval methods. URR stands for “Unlabeled Retrieved Results” and refers to experimental results from section 3.2.5. All the correlations are statistically significant with p-value < 0.05

	MRR@10	$FD@1$	$FD@1$ URR	$FD@10$	$FD@10$ URR
MRR@10	1	-0.473	-0.545	-0.788	-0.636
$FD@1$	-0.473	1	0.687	0.443	0.290
$FD@1$ -URR	-0.545	0.687	1	0.636	0.485
$FD@10$	-0.788	0.443	0.636	1	0.848
$FD@10$ - URR	-0.636	0.29	0.485	0.848	1

valuable insights into their relevance. *In contrast, traditional IR metrics would be unable to provide any insights without retrieving the labeled documents.*

3.2.6 FID Correlation with IR Evaluation Metrics

We aim to examine the correlation between the FD measure and traditional IR evaluation metrics. To achieve this, we calculate the ranked-based Kendall τ correlation, for each pair of metrics in Table 3.5 and Table 3.7 on the performance of the 12 retrievers introduced earlier and report the results in Table 3.8. This set of evaluation metrics includes MRR@10, FD at cut-offs 1 and 10 (Section 3.2.3) and FD at cut-offs 1 and 10 under URR setting when no labeled data is retrieved (Section 3.2.5). As anticipated and illustrated in Figure 3.9, FD exhibits a negative correlation with MRR, as a lower FD value indicates better performance. Among these correlations, $FD@10$ shows the highest absolute correlation with MRR@10 i.e., a correlation of -0.788. We suggest that this is because FD operates based on the distribution of embedded representations of documents, which has shown to work most stably when the number of samples increases [92, 49]. More interestingly, $FD@1$ and $FD@1$ with Unlabeled Retrieved Results (URR), obtain a correlation coefficient of 0.687. Similarly, the correlation between $FD@10$ (Fréchet Distance at 10) and $FD@10$ with unlabeled retrieved items was found to be 0.848. The high correlation between evaluating the original retrieved results vs without having any judged retrieved results further validates the findings presented in sections 3.2.3 and 3.2.5. The Fréchet Distance not only exhibits a high correlation with traditional IR metrics but also demonstrates its capability in assessing unlabeled retrieved items. These experiments let us answer **RQ4** that FD shows a notable correlation with traditional IR metrics. These properties increase the reliability of using

FD for assessing IR systems.

3.2.7 Impact of Document Representation

Here, we examine the robustness of the Fréchet Distance metric for assessing IR systems with respect to the underlying language model to embed the retrieved documents and relevance judgments. We aim to investigate how the choice of language model impacts the quality of evaluating IR systems using the Fréchet Distance measure considering this change would vary the document feature vectors. For previous experiments, we utilized a language model that was fine-tuned on the MS MARCO dataset for ranking tasks. However, now we study how the results would be impacted if we were to embed the retrieved documents and ground truth in a different space. As such, we present the same results as in Table 3.5, using DistilBERT embeddings fine-tuned on the MS MARCO training set as well as the same results with a DistilBERT without any fine-tuning. This analysis aims to investigate whether a general-purpose language model can capture the necessary information for accurate assessment, or if a language model specifically fine-tuned for ranking tasks in retrieval is required. Table 3.9 displays the obtained results. Surprisingly, we observe that changing the language model from a fine-tuned ranking model to a raw, unfine-tuned BERT model does not substantially impact the assessment outcomes. The *FD* metric remains capable of effectively evaluating the performance of various retrieval methods. For example, from Table 3.9, and under “DistilBERT No fine-tuning” column, we observe that BM25 achieves an *FD@10* score of 4.410, whereas COLBERT, which is expected to be a better model, achieves a score of 0.980.

The correlation between *FD@10* and *MRR@10* when using DistilBERT without any fine-tuning, is -0.739. Comparatively, when using fine-tuned DistilBERT (as shown in Table 3.8), the correlation is -0.788. As such, having a fine-tuned language model specifically for ranking task can improve the correlation with traditional IR metrics. However, even without fine-tuning, *FD* still demonstrates promising performance. Overall, the results indicate that *FD* remains effective in evaluating the quality of retrieved results, even when employing a general-purpose language model without fine-tuning. Lastly, with respect to **RQ5**, we note that *FD* shows promising robustness w.r.t the document embedding representation.

Table 3.9: Comparison of the performance of different retrievers when assessing with MRR@10 and $FD@10$ on MS MARCO dev set With DistilBERT fine-tuned on MS MARCO as well as DistilBERT without any fine-tuning. DistilBERT fine-tuned on MS MARCO shows -0.788 Kendall τ correlation with MRR@10 and DistilBERT without any fine-tuning shows -0.739 Kendall τ correlation with MRR@10.

Category	Method	MRR@10	$FD@10$	
			DistilBERT MS MARCO	DistilBERT No Fine-tuning
Sparse	BM25	0.187	0.590	4.410
	DeepCT	0.242	0.412	2.354
	DocT5	0.276	0.331	2.050
Dense	RepBERT	0.297	0.159	1.223
	ANCE	0.330	0.121	0.995
	SBERT	0.333	0.132	1.008
	ColBERT	0.335	0.117	0.980
	ColBERT V2	0.344	0.118	0.982
Trained	UniCOIL	0.351	0.123	0.980
Sparse	SPLADE	0.368	0.120	0.964
Hybrid	ColBERT-H	0.353	0.116	0.973
(BM25)	ColBERT V2 -H	0.368	0.126	0.998

3.2.8 Takeaway

In this section, we leveraged Fréchet Distance to address the challenges of evaluating IR systems with sparse labels. We measure the similarities between the embedded representation of retrieved results as well as the limited available relevant judged documents using Fréchet Distance. Through experiments conducted on datasets with sparse and more complete ground truth labels, we demonstrated the effectiveness of the Fréchet Distance in evaluating IR systems. Our findings suggest that the Fréchet Distance has significant implications for evaluating IR systems in real-world settings where obtaining comprehensive ground truth labels can be challenging and expensive. Since FD does not specifically reward having the exact gold answer as the output, it can also be used to evaluate different generative models, broadening the scope of evaluation in IR systems. This enables a direct comparison between generated and retrieved results within the same evaluation framework.

Limitations

Finally, while this study provides valuable insights into the effectiveness of the Fréchet Distance in evaluating IR systems with sparse labels, there are a few limitations that should be acknowledged. First, unlike traditional IR evaluation metrics, the Fréchet Distance is not applicable to individual queries and can only be used with sets of queries. Further exploration is needed to understand how the sample size of the queries affects the quality of the assessment. Second, the Fréchet Distance assumes that the two distributions follow a multivariate normal distribution. Lastly, it is important to note that the Fréchet Distance is an unbounded metric, and its range varies depending on the dataset’s characteristics and the number of samples under investigation.

Chapter 4

Evaluation of Information Retrieval Systems with Large Language Models

Large Language Models (LLMs) are increasingly being used for evaluation across a wide range of domains, including natural language processing and automated content assessment [151, 88]. The information retrieval community also has been an early adopter of LLMs for relevance assessment [388, 221, 471, 344]. In this chapter, we begin by providing an overview of how LLMs are reshaping evaluation paradigms in IR in Section 4.1. We further explore how the evaluation of generative information retrieval systems differs from that of traditional retrieval systems.

A growing body of work has demonstrated that LLM-generated relevance labels can closely align with human judgments under various agreement metrics [398, 397, 254]. Motivated by this, in Section 4.2 we systematically explore different approaches to using LLMs for generating relevance judgments for evaluating IR systems. We analyze their alignment with human labels, as well as their agreement with system rankings. Finally, in Section 4.3, we dive deeper into the role of prompting in LLM-based relevance assessments and how prompt variations can influence evaluation outcomes.

4.1 Generative Information Retrieval Evaluation

The current and potential capabilities of GenIR systems were engendered by the increasing capabilities of LLMs especially their ability to conduct zero-shot natural language tasks, including summarization, query understanding, and query expansion. Most GenIR systems

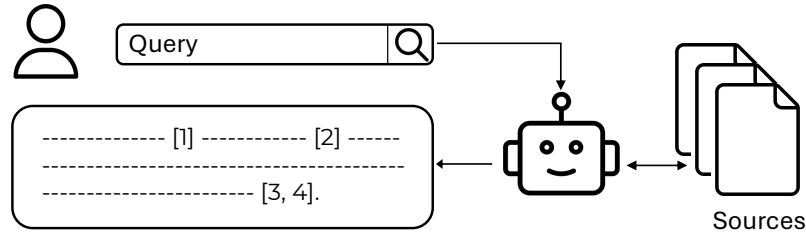


Figure 4.1: A GenIR system as a synthetic search engine

replace the query and ranked list with a conversation and a written synthesis of information. The searcher poses a question in potentially a longer, more natural and conversational form. The system responds with a single coherent answer, which may be supported by links to sources. Gienapp et al. [149] view a GenIR system as a “synthetical” search engine that searches for sources, “compiles them, synthesizes missing information, presents it coherently, and grounds its claims in the retrieved sources.” The system provides searchers with a single unified answer “that covers a complex topic with in-depth analysis from varied perspectives”. Such interactions and outputs as shown in Figure 4.1 will require us to seek a new evaluation model.

For evaluation purposes, we need not make any assumptions about the internal architecture of a GenIR system, which may simply be a single large neural model. In this case, our evaluation must focus on the end-to-end interaction. A query or question is entered by the searcher and the system responds with an answer, which may reflect a larger conversational context, including personalization. Under this view, our core search metric becomes: How good is this overall response? In traditional IR evaluation, the focus was often on the output of a ranker. While whole-page relevance was a factor in evaluation, especially in industry contexts [36], it was one of many factors. If we view a GenIR system as a black box, whole-page relevance becomes a central factor.

While we can view a GenIR system as a black box for evaluation purposes, a RAG architecture [146] often underlies a GenIR system. In Figure 4.2, we have simplified the architecture of a RAG system to its key components. At the front end, a searcher interacts with a *generative component*, which in turn interacts with both an *LLM* and a *retrieval component*. The retrieval component is used to search a corpus, which is assumed as a source of ground truth – although, like any IR system, the corpus itself may contain spam and documents of varying quality. Information provided by the RAG system to the searcher requires support from the corpus. The generative component interacts with the LLM for purposes of query understanding, query expansion, summarization, and similar

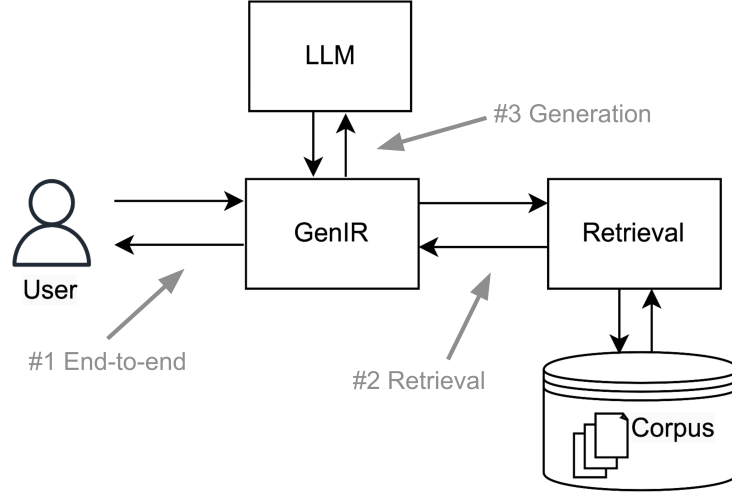


Figure 4.2: RAG Architecture overview

tasks, while it interacts with the retrieval component through keyword or other queries to find sources for its response. The system may interact with both the LLM and the retrieval component multiple times before responding to a user’s query, where the overall approach may be retrieve-then-generate, generate-then-retrieve [1], or a combination of multiple generation and retrieval steps.

As shown in Figure 4.2, a RAG system may be evaluated at three points: (1) at the front end, where we are evaluating the end-to-end performance of the system; (2) at the top of the retrieval component, where we are evaluating the retrieval component in the context of the overall GenIR system; and (3) at the point of interaction with the LLM. In the context of the overall system, the retrieval component (#2) is essentially a subordinate system returning a ranked list of items for the generative component. While a human searcher may eventually be given links to items in the corpus, but these will be selected by the generative component. Evaluating interactions with the LLM (#3) falls slightly outside our scope into the broader topic of NLP evaluation, including the evaluation of summarization and information extraction.

Evaluation of an end-to-end GenIR system (#1) introduces challenges beyond those of traditional IR evaluation. Gienapp et al. [149] argue that the key difference between a traditional search engine and a GenIR system is that the GenIR system is essentially searching an infinite corpus of all possible responses that could be synthesized by the system [123]. Traditional IR test collections, such as those created by the TREC, try to be reusable, with a nearly complete set of relevance judgments. With a finite corpus, this

approach is conceptually possible, with an infinite corpus it is not.

One approach to evaluating the retrieval component (#2) would be to evaluate it as a traditional search engine. Its role is to execute a query over a corpus of items and return a ranked list of those items. To evaluate the retrieval component of a RAG system we may be able to adapt existing offline evaluation methods. Even if the interface seen by the searcher is no longer “ten blue links”, internally we can imagine a similar interface between the generative component and the retrieval component, although the browsing models assumed by offline evaluation metrics no longer apply. These browsing models often assume that the searcher has limited patience [277] or that the searcher will stop scanning the ranked list after a relevant item is found [81]. A generative component might be assumed to dig deeper into the ranked list and seek information from more multiple sources.

RAG systems include a retrieval component (Figure 4.2), which supports retrieval over a corpus that provides ground truth for our GenIR system. For evaluation purposes, we might treat the retrieval component as an old-fashioned search engine, even if it itself includes generative components. A query goes into the retrieval component and a ranked list comes out. However, since this response is entirely internal to the GenIR system, it need not only be a ranked list. It could be richer and more complex. The output of the retrieval component must be tailored to the needs of the overall system, and not to the needs of a human searcher.

If we view the retrieval component as an old-fashioned search engine, returning a ranked list, we might employ traditional evaluation methods. If we think about the GenIR system as internally browsing down the output of the retrieval component, we could use NDCG@10 as our metric. However, the GenIR has more “patience” than a human searcher, so the NDCG discount function might not be the right one to use. The purpose of the retrieval component is to return the items that the overall GenIR system needs to craft its response.

Even when supported by a retrieval component, GenIR systems might generate factually inaccurate or misleading responses. In traditional IR evaluation we assume that the corpus is curated and can be trusted. If we can not trust it, then we filter it for spam and other misinformation. While in traditional web search some pages are higher quality than others, the output of the search engine is a list of pages, which the searcher can ultimately inspect for themselves. They are not depending on the search engine to summarize the information for them. Since GenIR systems can hallucinate [391], it is not sufficient to filter the corpus for spam and misinformation. We must also evaluate the accuracy of the end-to-end response. The final generated response can be false, or contain falsehoods, even if the retrieved material is true. Fact checking must become a standard component

of GenIR evaluation.

The accuracy of a traditional search engine depends on the accuracy of the information in its corpus. The search engine may not be able to find relevant information, but when it does, it does not alter or interfere with it. If the corpus contains misinformation, we attempt to filter it. For evaluation purposes, we measure the quality of the filter. Since a GenIR system can hallucinate misinformation, we must now evaluate accuracy of its output, along with relevance and other traditional considerations.

In the case of GenIR, the entirety of the system’s end-to-end response should be relevant and nothing should be redundant, the boundaries between documents hold far less importance. Everything in the response should be there for a reason, and in many cases the response should include more than just the bare answer. The response might link to background articles that support the response. It might provide opposing perspectives. It might suggest cheaper or higher quality alternatives to a product. It might synthesis similar responses from multiple sources into a single sentence. It might ask for clarification or disambiguation.

In general, despite the challenges introduced by GenIR systems, we can still build upon traditional evaluation strategies by considering both their similarities and differences with earlier information retrieval systems. However, they must be carefully adapted to capture the end-to-end behavior of GenIR systems. Tailoring evaluation to these new characteristics is essential for producing meaningful and reliable assessments of GenIR performance.

4.2 LLM-based Relevance Judgment for IR Evaluation

In this section, we specifically focus on how generative models reshape perspectives in evaluating information retrieval systems through LLM-based relevance judgments. There have long been attempts to automate the process of relevance judgment [260, 259, 328, 368]. However, with the emergence of LLMs, automated relevance judgment has gained significant attention[397, 318]. If we can automate relevance judgments, we can automate evaluation itself—reducing human labor, scaling evaluations, and making the process faster and more cost-effective. While automated evaluation metrics offer scalability, human judgment remains essential for capturing subjective quality measures and ensuring alignment with human labels [419, 17]. To enable effective hybrid evaluations, automated methods,

particularly those based on LLMs, must be aligned with human judgment [191]. Furthermore, conclusions drawn from annotations should be interpretable for system comparison. For instance, in LLM-based relevance judgments for information retrieval evaluation, it is crucial to understand how these judgments influence system rankings and measure system effectiveness. LLM-based relevance assessment is not only valuable for assessing IR systems but may also play a crucial role in evaluating RAG systems, where answers are not necessarily derived from a fixed collection [458].

The authors of [131] pioneered LLM-based relevance judgments by discussing the potential for various levels of collaboration between humans and machines in evaluating IR systems. Thomas et al. [388] illustrate that LLM-based relevance judgments have on-par quality as human annotators and they could even be deployed at an industrial scale, as implemented at Microsoft Bing. While these studies reported a high rank-based Kendall’s τ correlation (typically above 0.85) between human and LLM-graded judgments [2, 131, 254, 388], they also highlight a lack of calibration between automated relevance judgments and human assessments, with LLM-based assessment often assigning higher grades on average [398].

Studies have shown that LLM-based relevance assessments can achieve accuracy comparable to, or even exceeding, that of human annotators, as demonstrated in real-world systems like Bing [388]. These models not only reduce labeling costs but also address “assessment holes” where unjudged results in traditional test collections lead to performance underestimation of newer models [23]. However, the growing reliance on LLMs for evaluation introduces challenges of consistency, alignment, and transparency. As Faggioli et al. [131] note, evaluations in IR must remain grounded in human judgment to maintain trust and reliability.

In fact, despite the success of LLM relevance assessments, several papers have raised concerns about their robustness and their alignment with human preferences. Alaofi et al. [10] report a “tendency for many LLMs to label passages that include the original query terms as relevant.” They demonstrate that merely injecting query terms into a document or even adding an explicit statement that the document is relevant can influence an LLM to label the document as relevant. Clarke and Dietz [98] report an experiment demonstrating that LLM assessment can be biased toward LLM-based re-ranking methods. The authors of the **UMBRELA** assessment tool [398, 397]—which was used for automated evaluation of TREC RAG 2024—found that when asking LLMs and humans to rate relevance using graded values, humans tend to apply stricter criteria. For instance, when analyzing three years of TREC Deep Learning track data (2019–2021) using **UMBRELA** and comparing it to human judgments, it was observed that on average humans labeled over 13% more documents as non-relevant (14,961 vs. 17,376). Conversely, LLMs were more lenient in judging relevance,

labeling over 26% more documents as perfectly relevant (3,063 vs. 2,429). This observation suggests that LLMs may interpret relevance more liberally than humans. Differences between human and LLM-based relevance judgments are also reflected in statistical measures of alignment with human labels and agreement with system rankings [108, 398, 324, 131]. To the best of our knowledge, the extent to which these discrepancies impact the interpretation of evaluation results across different LLM-based evaluation methods has not been extensively studied. This gap may be due to the lack of available data and the limited diversity of LLM-based relevance judgments for a given dataset.

Given the flexibility of LLMs in language processing tasks, we are no longer constrained to standard methods for relevance assessment. Methods that once required substantial human labor — which may have impeded their adoption — now become feasible due to the reduced costs associated with LLM-based assessment. However, there has been limited exploration of alternative methods or a comprehensive comparative studies of methods beyond graded relevance assessment.

With the goal of placing different relevance judgments on a “level playing field” in this section, we compares various LLM-based relevance assessment methods, including:

1. Traditional relevance judgments in both **Binary** [131] and graded forms (**UMBRELA**) [398]
2. Two nugget-based methods – one document-agnostic (**Exam**) [132, 346] and one document-dependent (**AutoNuggetizer**) [318]
3. A pairwise preference-based relevance judgment method.

To minimize the influence of prompt engineering and enhance reproducibility, as much as possible we rely on established tools and prompts.

Our of our objectives here is to help comparing different LLM-based relevance judgment systems on a level playing field by considering two factors:

1. **Alignment with human labels:** The degree to which LLM-generated relevance judgments reflect the document ordering imputed by human labels.
2. **Agreement with system rankings:** The consistency between system rankings produced using LLM-generated labels and those derived from human labels.

As observed earlier, LLMs can be more lenient in judging relevance. The first factor (“alignment”) allows us to side-step this issue by focusing on the order in which documents

are placed by the various assessment methods. For example, if human assessment assigns a higher grade to one document vs. another, we expect LLM-based assessment to provide a consistent ordering, even if the LLM assigns different grades. Although the various assessment methods express relevance in different ways, e.g., grades vs. nuggets, alignment with human labels allows us to make a direct comparison between different relevance judgment methods. Full details are provided in Section 4.2.3. For similar reasons, for the second factor we employ an flexible evaluation measure “compatibility” that naturally adapts to any relevance assessment method [101]. Compatibility essentially compares a given ranking to the ideal permutation of judgments. Since it operates on relative relevance rather than absolute values, it enables meaningful comparisons between different relevance judgment approaches. Further details on how compatibility is computed and how we apply it in our study are provided in Section 4.2.4.

In this section, we first curate a comprehensive resource of relevance judgments generated using five distinct methods—binary relevance, graded relevance, pairwise preference, document-agnostic nugget-based evaluation, and document-dependent nugget-based evaluation with GPT-4o and four benchmark datasets (TREC DL 2019, TREC DL 2020, TREC DL 2021, and ANTIQUE). All relevance judgments, along with code and experimental results, are publicly available on our GitHub repository¹. We systematically compare LLM-generated relevance judgments with human labels, ensuring a fair and consistent evaluation by placing them on the same playing field. As new large language models appear and assessment methods are proposed, these judgments provide a baseline for evaluation. Further, we introduce a method for assessing how well different relevance judgment approaches align with system rankings, providing a standardized framework for cross-method comparison.

4.2.1 Assessment of Relevance Judgment

Alignment with Human labels.

Discrepancies exist between human and LLM-based relevance judgments. For instance, humans are found to be generally more strict in defining relevance compared to LLMs [397, 398, 267]. These differences are also reflected in statistical agreement measures, such as Cohen’s Kappa (κ). When comparing human and LLM assessments, Faggioli et al. [131] report binary κ values between 0.07 and 0.49 (slight to moderate). Upadhyay et al. [398] report κ agreement for binary and graded relevance judgment between LLM and

¹<https://github.com/Narabzad/llm-relevance-judgement-comparison>

human annotations. For binary, they report κ between 0.4 and 0.5 (moderate) and for four-point values between 0.3 and 0.4. Rahmani et al. [324] report an experiment in which seven independent research groups applied their prompts and LLM-based assessment tools to label a common set of query/document pairs on a four-point graded relevance scale. No group achieved a κ over 0.45, when compared with human labels, with all but one submission having κ values below 0.4. As a basis for comparison, human-human κ values can be above 0.5 on binary assessment [108].

Agreement with System Rankings.

In addition to a direct comparison of relevance labels, information retrieval researchers often evaluate relevance grading methods and evaluation metrics by computing Kendall’s Tau (τ) correlation between system rankings [404]. Kendall’s τ measures the rank correlation between two ordered lists, accounting for concordant and discordant pairs. In the context of evaluation experiments, such as TREC tasks, Kendall’s τ quantifies the consistency between the rank orderings induced by two different evaluation metrics across a set of submissions (i.e., “runs” or “systems”). In comparing official TREC binary relevance labels with an independently created set of labels on runs submitted to the TREC-6 ad-hoc retrieval task, Voorhees [404] report a Kendall’s $\tau = 0.8956$, where the corresponding Kappa is 0.52 [108]. Thus, if we are replacing human labels with LLM-generated labels, a Kendall’s τ of 0.9 could serve as a minimum benchmark for acceptability.

LLM-generated labels approach or exceed this standard when compared to human labels. Faggioli et al. [131] report $\tau = 0.86$ on TREC 2021 Deep Learning runs when comparing human labels with labels generated by GPT-3.5. Similarly, Upadhyay et al. [398] report τ values between 0.87 and 0.94 on TREC 2019 to 2023 Deep Learning runs when comparing human labels with labels generated by their UMBRELA assessment tool. Furthermore, Upadhyay et al. [397] report $\tau = 0.89$ on TREC 2024 runs using the same tool.

Alignment vs. Agreement. While previous studies have compared prompts for LLM-based graded assessment [324], to our knowledge, no prior work has systematically compared different LLM-based relevance assessment methods. Various research teams have reported results on different datasets, but due to the lack of comprehensive experiments, there has been no unified comparative analysis. This gap arises not only because relevant datasets and judgments are sometimes computationally/financially expensive to generate but also due to the absence of a clear and standardized methodology for making such

comparisons. Here, we address these challenges by: (1) providing comprehensive LLM-based relevance judgments using five different methods and two different LLMs across four datasets, and (2) introducing two methodologies for comparing LLM-based relevance judgments—one based on alignment with human labels and the other on agreement with system rankings. By standardizing the evaluation process, we enable a more rigorous and reproducible comparison of different LLM-based assessment approaches.

4.2.2 Relevance Judgment Methods

In this section, we present a detailed explanation of the LLM-based relevance assessment methods. To ensure replicability and to minimize potential bias introduced by prompt engineering, we have relied as much as possible on the tools and prompts developed by prior work. We note that all the experiments have been conducted with a temperature of 0.

Binary Relevance Assessment

Binary relevance assessment is a straightforward approach assessing relevance, which has been widely employed within the information retrieval community for much of its history [105]. Its simplicity makes it a popular choice and a strong foundation for evaluation metrics like precision, recall and so on. A document d is evaluated to determine whether it satisfies the information need underlying a query q . Given a query q and a document d , a **Binary** relevance function $\mathcal{B}(q, d)$ maps the query-document pair to a label r :

$$\mathcal{B}(q, d) = r, \quad \text{where } r \in \{0, 1\} \quad (4.1)$$

Here, $r = 1$ indicates relevance, while $r = 0$ signifies non-relevance.

Binary judgments, while easy to understand and interpret with assessors less likely to disagree on labels compared to graded scales [262, 197, 198], can oversimplify the relevance spectrum, failing to capture more nuanced degrees of relevance. Moreover, defining consistent boundaries between relevance grades is difficult. Faggioli et al. [131] were among the first to leverage LLMs for binary relevance judgments. For **binary** assessment we reproduce their prompt, which explicitly instructs the LLM to act as a TREC assessor when determining if a document is relevant to a query. Their prompt has previously shown high agreement with human judgments on datasets including REC-DL 2021 [115] and TREC-8 [167].

Graded Relevance Assessment

Early efforts in IR, highlighted the need for richer annotations to better evaluate retrieval systems [57, 167, 405]. As such, graded relevance judgment extend binary relevance by assigning a relevance grade r from a predefined set of levels, typically ranging from 0 (non-relevant) to r_h (highest relevance). For a given query q and document d , a graded relevance function $\mathcal{G}(q, d)$ maps the query-document pair to an integer r , where:

$$\mathcal{G}(q, d) = r, \quad \text{where } r \in \{0, 1, \dots, r_h\}. \quad (4.2)$$

The number of levels ($h + 1$) is often 4 or 5 in standard test collections. For instance, the TREC datasets use scales ranging from 0 (not relevant) to 3 (perfectly relevant) [112]. Graded relevance judgments enables finer-grained evaluation metrics such as nDCG, with the goal of making evaluation better aligned with real-world user satisfaction, where not all relevant documents are equally useful. However, graded relevance suffers from problems of subjectivity, where assessors may disagree on the label to assign to a particular document. As a result, graded relevance requires more detailed guidelines and training for assessors.

As LLMs became prominent, several researchers explored their use in generating graded relevance judgments [26, 9, 2, 267, 388]. Notably, Bing researchers announced that they had replaced human annotators with LLMs for some production-level relevance judgments, demonstrating the potential of LLMs to replace or augment human assessment [388]. The UMBRELA method [397], provides open-sourced prompts replicating Bing’s LLM-based relevance judgment system, demonstrating high agreement with system rankings, particularly in the TREC RAG 2024 evaluation. For this study, we adopted UMBRELA² prompts and implementation with a zero-shot setting.

Nugget-based Relevance Assessment

To facilitate the assessment process, some evaluation methods decompose the information need underlying a query. Different papers give different names to the products of this decomposition, including “facets”, “subtopics”, “rubrics”, and “nuggets”. To maintain consistent terminology throughout this thesis, we will refer to them as “nuggets”. Nuggets facilitate a more granular and detailed assessment of a document’s relevance. By considering relevance of the nuggets that a document satisfies, we can determine not only if a document is relevant to a query but also to what extent, and from which perspective, it fulfills it. This approach enhances the explainability of relevance judgments compared to

²<https://github.com/castorini/umbrela/tree/main>

coarse-grained methods. Prior studies [309, 309, 263] have shown that nugget-based frameworks improve the interpretability of relevance judgments in information retrieval systems. For this study, we adopt two distinct LLM-based relevance assessment approaches within the larger nugget-based framework:

1. A *Document-agnostic* approach called **Exam** [133, 346] where nuggets are extracted solely based on the main query A *Document-dependent* method called **AutoNuggetizer** [318], which extracts nuggets from documents which are known to be relevant to the query.

Document Agnostic (Exam) Older TREC tracks employed subtopics manually extracted from queries to facilitate relevance judgment tasks [97, 454]. Under the **Exam**³ method [133, 346], an LLM is prompted to generate nuggets based on the query alone — independent of any retrieved documents. This document agnostic approach avoids biases potentially introduced by using relevant documents as the source of nuggets but also limits nuggets to those that can be generated by the LLM without any additional context. This document-agnostic approach ensures that the rubric generation is purely query-driven.

Exam Nugget Generation: For a given query q , **Exam** derives 10 nuggets, denoted as $\mathcal{N}_{\mathcal{E}}(q) = \{N_1^q, N_2^q, \dots, N_{10}^q\}$. The prompts used by **Exam** require the extracted nuggets to be concise and insightful questions, which together are intended to provide complete coverage for the query.

Nugget-Based Assessment After the nuggets are generated, **Exam** determines the extent to which a given document satisfies each nugget. Both binary and graded judgments may be employed at the per-nugget level.

- **Exam-Binary:** For document d , an LLM is prompted to make a binary judgment with respect to each nugget N_i^q , where $B(N_i^q, d)$ is 1 if the document is relevant to N_i^q , and 0 otherwise. These binary judgments are then aggregated to give a score for the document as follows:

$$\mathcal{E}_b(q, d) = \frac{1}{|\mathcal{N}_{\mathcal{E}}(q)|} \sum_{i=1}^{|\mathcal{N}_{\mathcal{E}}(q)|} \mathcal{B}(N_i^q, d) \quad (4.3)$$

³<https://github.com/laura-dietz/flan-t5-exam-appendix>

- **Exam-Graded:** As proposed by Farzi and Dietz [133], an LLM is prompted to assign an integer relevance grade $r \in [1, 5]$ to each nugget N_i^q indicating the degree to which the document d satisfies it. Further, we aggregate these per-nugget relevance grades in two ways, which are discussed separately below. The grades are defined as follows:
 - **1:** The document minimally addresses the rubric.
 - **2:** It partially satisfies the rubric with limited relevance.
 - **3:** It provides substantial relevance but lacks completeness.
 - **4:** The rubric is well-covered with minor omissions.
 - **5:** The document fully satisfies the rubric, offering comprehensive relevance.

Graded Nugget Aggregation After **Exam-Graded** assigns a per-nugget relevance grade, these grades can be aggregated in two ways to produce an overall score for a document:

- **Exam-Graded Max:** This is the original aggregation proposed by Farzi and Dietz [133] which selects the maximum score across all nuggets:

$$\mathcal{E}_{g_{\max}}(q, d) = \max_{i=1}^{|\mathcal{N}_{\mathcal{E}}(q)|} \mathcal{E}_g(N_i^q, d) \quad (4.4)$$

- **Exam-Graded Mean:** In addition, we explore a second aggregation function that computes an average score across nuggets:

$$\mathcal{E}_{g_{\text{mean}}}(q, d) = \frac{1}{|\mathcal{N}_{\mathcal{E}}(q)|} \sum_{i=1}^{|\mathcal{N}_{\mathcal{E}}(q)|} \mathcal{E}_g(N_i^q, d) \quad (4.5)$$

Document Dependent (AutoNuggetizer)

This method generates nuggets based on both a query and relevant documents. Unlike the query-only approach in **Exam**, **AutoNuggetizer** [318] anchors nugget generation to documents that were judged as relevant. This coupling provides richer contextual grounding for the nuggets but may also introduce potential biases toward the source documents. The overall approach is similar to that of Pavlu et al. [309], who proposed a nugget-based test collection created through human assessment. We note that for implementing all of the following steps, we used prompts provided by **AutoNuggetizer** [318].

Nugget Generation. Given a query q , the nugget generation process in **AutoNuggetizer** considers a set of documents D_q that have been judged relevant to the query. We denote the set of nuggets generated by an LLM with **AutoNuggetizer** as

$$\mathcal{N}_A(q, D_q) = \{N_1^q, N_2^q, \dots, N_n^q\},$$

where $n < 30$ and each document $d \in D_q$ satisfies $\mathcal{B}(d, q) > 0$ or $\mathcal{G}(d, q) > 0$. This ensures that nuggets are derived exclusively from documents deemed relevant to the query q .

The nugget generation process is an iterative process that analyzes all relevant documents $d \in D_q$. This process is intended to ensure that the generated initial nuggets are neither too broad nor overly specific. Nuggets capturing similar concepts are recursively merged to maintain clarity and reduce redundancy. If the total number of nuggets exceed the limit (30), the process involves merging redundant nuggets or removing less important ones.

Nugget Importance. After nugget generation, each nugget is categorized based on their importance into one of two groups: (1) *Vital* nuggets, which are essential for a good response, and (2) *Okay* nuggets, which are useful but not critical for completeness. Further, all nuggets are sorted by their importance, and the top 20 most important nuggets are selected. Some queries may result in fewer than 20 nuggets, and in certain cases, nuggets may be truncated during this process. The final nugget set for query q is denoted as:

$$\mathcal{N}_{Sorted}(q, D_q) = \{(N_1^q, I_1), (N_2^q, I_2), \dots, (N_n^q, I_n)\},$$

where N_i^q represents a nugget, and I_i is its importance, either *vital* or *okay*. N_i is as important as or more important than N_j if $i < j$.

Nugget Assignment. Once the nugget list is finalized, the LLM assesses documents based on how well they support each nugget. Each document d and nugget N_i is assigned $\mathcal{SP}(N_i^q, q)$ with one of the labels: **support**, **partial_support**, **not_support**.

Aggregation. The relevance of a document d for a query q is computed based on nugget assignments with two different scoring functions. S_i is score of nugget N_i^q for query q and a more strict version of scoring SS_i focuses only on nuggets that are either **supported** or **not_supported**, omitting the **partial_supported** of as follows:

Table 4.1: Scoring Metrics for Nugget Evaluation. n_v and n_o stands for vital nuggets and okay nuggets respectively. s_i^v and s_i^o indicates the scoring only on vital nuggets and okay nuggets respectively. n_a , n_v and n_o represents number of all nuggets, vital nuggets and okay nuggets respectively.

Metric	Definition	Metric	Definition
All (A) =	$\sum_i \frac{S_i}{n_a}$	All Strict (A_{strict}) =	$\frac{\sum_i SS_i}{n_a}$
Vital (V) =	$\sum_i \frac{s_i^v}{n_v}$	Vital Strict (V_{strict}) =	$\frac{\sum_i ss_i^v}{n_v}$
Weighted (W) =	$\frac{\sum_i s_i^v + 0.5 \cdot \sum_i s_i^o}{n_v + 0.5 \cdot n_o}$	Weighted Strict (W_{strict}) =	$\frac{\sum_i ss_i^v + 0.5 \cdot \sum_i ss_i^o}{n_v + 0.5 \cdot n_o}$

$$S_i = \begin{cases} 1, & \mathcal{SP}(N_i^q, d) = \text{support} \\ 0.5, & \mathcal{SP}(N_i^q, d) = \text{partial_support} \\ 0, & \mathcal{SP}(N_i^q, d) = \text{not_support} \end{cases} \quad (4.6)$$

$$SS_i = \begin{cases} 1, & \mathcal{SP}(N_i^q, d) = \text{support} \\ 0, & \mathcal{SP}(N_i^q, d) = \text{not_support} \end{cases} \quad (4.7)$$

Upadhyay et al. [397] propose six different aggregation functions, varying the focus between vital nuggets only or all nuggets, and distinguishing between stricter scoring (SS) and more flexible scoring (S). These aggregation functions are summarized in Table 4.1, where three core scoring functions are defined, and by replacing S_i with SS_i , the total expands to six scoring functions: Here, we refer to them as **Nuggets All (A)** and **Nuggets All (A) Strict**, which assign weights to all nuggets; **Nuggets Vital (V)** and **Nuggets Vital (V) Strict**, which focus exclusively on vital nuggets and disregard the okay nuggets; and a weighted summation of vital and okay nuggets, **Nuggets Weighted (W)** and **Nuggets Weighted (W) Strict**, where greater weight is assigned to vital nuggets compared to less important ones.

Pairwise Relevance Judgments

Pairwise preferences judgments provide a simple way to capture human feedback that ensures better inter-assessor agreement [196, 102, 70, 342, 434]. Human preference labels are also central to the reinforcement learning from human feedback (RLHF) process used to tune LLMs and other models [93]. Pairwise judgments are less sensitive to subjective biases and focus on ordinal relationships, which are often more consistent across assessors [102, 441]. Pairwise relevance judgment evaluates the relative relevance of two documents with respect to a given query. Instead of assigning absolute relevance scores or grades, this method asks the question “Which document is more relevant to the query?”

Formally, given a query q and two documents d_a and d_b , a pairwise function $\mathcal{P}(q, d_a, d_b)$ determines the relative preference between the documents. To avoid positional biases, for LLM-generated preference judgments we repeat the assessment by swapping d_a and d_b . In case both $\mathcal{P}(q, d_a, d_b)$ and $\mathcal{P}(q, d_b, d_a)$ agreed on the output, we consider the output document to be more relevant than the other one. Otherwise, we consider outcome to be a *tie* i.e., both document enjoy the same level of relevance. In other words:

$$\mathcal{P}(q, [d_a, d_b]) = \begin{cases} d_a & \text{if } \mathcal{P}(q, [d_b, d_a]) = \mathcal{P}(q, [d_a, d_b]) = d_a \\ d_b & \text{if } \mathcal{P}(q, [d_b, d_a]) = \mathcal{P}(q, [d_a, d_b]) = d_b \\ \text{tie} & \text{if } \mathcal{P}(q, [d_b, d_a]) \neq \mathcal{P}(q, [d_a, d_b]) \end{cases} \quad (4.8)$$

Since there are no established LLM-based relevance assessment tools for pairwise preference judgment, we developed our own prompt based on the instructions given to human preference judges by Clarke et al. [102]. As discussed above, the LLM is prompted twice, with the passages reversed, to minimize positional bias. While complete judgments for n documents would require $O(n^2)$ preference judgments, we follow the sampling process of Clarke et al. [102] which reduces the total number of judgments to $O(n)$, or more specifically to less than $P \cdot n$ judgments, where $P = 7$, the default in their code release. For aggregation, we assign an overall score to a document by counting the number of pairings it wins, ignoring ties, similar to a Borda score.

4.2.3 Alignment with Human Labels

In this section, we detail methodologies for determining the degree to which LLM-generated relevance judgments reflect preferences imputed by human labels. For each of the methods described in Section 4.2.2, we measure alignment with human labels by asking the following question:

When human assessment determines that one response to a query is better than another, how often does LLM-based assessment agree?

This approach allows us to compare methods that are based on different approaches to relevance judgment. Since some test collections use different relevance grades, and since some queries do not have documents at each relevance level, for each query we group the relevance judgments into three categories as follows, simplifying and unifying our analysis:

- *Best known*: For each query, these are the relevant judged documents at the highest grade of relevance for that query. Given a 4 scale level of relevance (0-3) Some queries may have no relevant document with grade 3, or even grade 2, although we assume all queries have at least one judged non-relevant document.

- *Acceptable*: For each query, these are the relevant documents that fall between the grade of the *best known* answer and the judged non-relevant ones (grade 0).
- *UnAcceptable*: For each query, these are the documents judged as not relevant to the query (grade 0).

Each method is measured in terms of its ability to make assessments that are consistent with the order imputed by these categories. If an LLM-based relevance assessment method orders documents consistently with these categories, it is said to “Agree” with human labels. If an LLM-based relevance assessment method orders documents inconsistently with these categories, it is said to “Disagree” with human labels. If an LLM-based relevance assessment method assigns the same score to documents from different categories, it is considered a “Tie”.

To study the alignment of relevance judgments with human labels, we assess the performance of a relevance judgment method μ on document pairs drawn from three categories: “Best”, “Acceptable” and “UnAcceptable”. For a given query q , let $D_q = \{D_{\text{best}}, D_{\text{acc}}, D_{\text{unacc}}\}$ represent the sets of judged documents in these categories. If documents in category c' has a higher level of relevance compared to c then for every pair $\langle d_c^q, d_{c'}^q \rangle$ where $d_c^q \in D_c^q$ and $d_{c'}^q \in D_{c'}^q$, $d_{c'}^q$ has a higher relevance level than d_c^q (e.g., $D_{c'}^q = D_{\text{best}}$ and $D_c^q = D_{\text{acc}}$). We measure the following as the $\mathcal{H}$ uman alignment of μ between two pairs of document:

$$\mathcal{H}_\mu(d_c^q, d_{c'}^q) = \begin{cases} \text{Tie} & \text{if } \mu(d_{c'}^q) = \mu(d_c^q), \\ \text{Agree} & \text{if } \mu(d_{c'}^q) > \mu(d_c^q), \\ \text{Disagree} & \text{if } \mu(d_{c'}^q) < \mu(d_c^q). \end{cases} \quad (4.9)$$

Finally, we compute the average of $\mathcal{H}_\mu(d_c^q, d_{c'}^q)$ across all pairs of documents in the two categories D_c^q and $D_{c'}^q$, which represents the human alignment of μ between two levels of relevance c and c' :

$$\mathcal{H}_\mu(c, c' | \mathcal{A}) = \frac{1}{|D_c^q| \cdot |D_{c'}^q|} \sum_{d_c^q \in D_c^q} \sum_{d_{c'}^q \in D_{c'}^q} \mathbf{1}(\mathcal{H}_\mu(d_c^q, d_{c'}^q) = \mathcal{A}), \quad (4.10)$$

where $\mathbf{1}(\cdot)$ is an indicator function that evaluates to 1 if the condition inside holds true, and 0 otherwise. For instance, if we are computing alignment with agreement, $\mathcal{A} = \text{Agree}$, $\mathcal{H}_\mu(c, c' | \mathcal{A})$ counts only the cases where the outcome of equation 4.9 is “Agree”, normalizing it over the total number of comparisons. The relative difficulty of these comparisons depends on the gap between relevance level of documents in D_c and $D_{c'}$. For instance, distinguishing between “Best” and “UnAcceptable” documents is generally easier than distinguishing between “Best” and “Acceptable” as the latter two categories might both enjoy some level of relevance.

We make comparisons between documents in the three categories as follows: 1) Best vs. Acceptable, 2) Acceptable vs. UnAcceptable, and 3) Best vs. UnAcceptable. We do not make comparisons between document in the same category (e.g., Best vs. Best). Even if an LLM-based method would indicate a difference between two documents in the same category, we do not know if this difference is due to the LLM recognizing a finer distinction in relevance or due an error on the part of the LLM. Given the limited number of relevance labels, we do not expect that documents assigned the same label are truly “tied” in relevance, and we do not demand that an LLM-based relevance assessment method assign them the same score. In measuring alignment, we focus on comparison where the human labels indicate a clear difference in relevance. Like Kendall’s τ , this approach to measuring alignment with human labels focuses on the ordering of paired items (concordant vs. discordant) but with ties handled in an application specific way.

4.2.4 Agreement with System Rankings

In addition to ensuring relevance judgments are aligned with human labels, relevance judgments must also enable effective system ranking. Different relevance judgment methods — such as binary, graded, and nugget-based — produce outputs on varying scales depending on their aggregation strategies. For instance, **Exam** can yield different ranges of outputs depending on the aggregation strategy used. **AutoNuggetizer** provides six distinct aggregation methods, each with its own range of values and scale. To apply traditional IR evaluation metrics like nDCG, it is necessary to map these outputs to proper integer grades. However, the diverse scales and range and floating-point outputs make this mapping ambiguous. Therefore, it is not possible to easily calculate traditional IR evaluation metrics with different relevance judgment methods.

To allow a direct comparison between relevance assessment methods, despite their varying scales, we adopt the *compatibility* evaluation metric introduced in Clarke et al. [104]. Compatibility offers a robust framework for comparisons between relevance judgment methods by focusing solely on the relative ordering of query-document pairs rather than their absolute scores. Compatibility is agnostic to the numerical scale used to express relevance assessments, enabling comparisons across diverse methods. Further, it allows direct comparisons between different relevance judgment methods without requiring normalization or recalibration of scores.

Let us consider an ideal ranking *set* I , which represents the preferred ordering of documents for a query. It is constructed from a set of *effectiveness levels* $\{L_1, L_2, \dots, L_T\}$, where $L_T \succ L_{T-1} \succ \dots \succ L_1$ and L_T contains the most relevant documents. Documents within the same level L_i are considered equivalent and can appear in any order. Non-relevant documents belong to L_0 are excluded from the ideal ranking. As such, we can obtain an ideal set of ranking I_μ for any given relevance judgment method μ by sorting the documents of a given query based on their different levels of relevance. The number of levels in I_μ may vary across methods, as μ might produce different ranges of relevance.

Given an actual ranking R , the ideal ranking set I_μ represents the set of ranking permutations created by ordering documents within each relevance level L_i according to their positions in R . By aligning the document order within each level to various permutations consistent with R , I_μ ensures the ideal ranking is both closely aligned with R and robust to ties within R . I' is a permutation in I_μ where the relevance levels are correctly ordered, but documents within the same level may appear differently in different permutations.

We can measure the similarity between R and any $I' \in I$ with any appropriate rank similarity measure. Following Clarke et al. [101], we use Rank Biased Overlap (RBO) [421], a rank similarity measure that emphasizes agreement at higher ranks. RBO is flexible in controlling the weight assigned to deeper ranks, with larger p values placing greater emphasis on lower-ranked documents. As suggested in [102], we set $p = 0.9$. Finally the compatibility score is defined as the maximum RBO value between R and any ranking permutation I' in the set of ideal rankings $\mathcal{I}_\mu$:

$$\text{Compatibility}(R, I_\mu) = \max_{I' \in \mathcal{I}_\mu} \text{RBO}(R, I') \quad (4.11)$$

In summary, compatibility measures the similarity of a ranking to the most similar ideal permutation of relevance judgments.

This process entails the following:

1. Sorting the documents based on their relevance levels $L_T, L_{T-1}, \dots, L_1$, ensuring that higher relevance levels appear at the top of the ideal ranking and
2. Within each level L_i , arranging the documents to match the order of R as closely as possible.

This approach ensures a fair comparison by preserving the relative ordering within equivalence levels while accommodating variability in relevance levels across different relevance judgment methods and it is defined as:

$$\text{RBO}(R, I') = (1 - p) \sum_{i=1}^{\infty} p^{i-1} \frac{|R_{1:i} \cap I'_{1:i}|}{i}, \quad (4.12)$$

where $R_{1:i}$ and $I'_{1:i}$ denote the top i items in R and I' , respectively.

4.2.5 Experiments and Results

We conduct experiments on three passage ranking datasets from the TREC Deep Learning (DL) Tracks i.e., TREC DL 2019 [112], TREC DL 2020 [113], and TREC DL 2021 [115], along with the ANTIQUE dataset [164], a benchmark for non-factoid question answering.

Table 4.2: Statistics of datasets used in our experiments.

Dataset	Queries	Non-relevant	Related	Relevant	Perfectly Relevant
DL 2019	43	5,158	1,601	1,804	597
DL 2020	45	7,780	1,940	1,020	646
DL 2021	57	4,338	3,063	2,341	1,086
ANTIQUÉ	200	1,642	2,417	1,196	1,334

For all TREC DL datasets, human annotators from the National Institute of Standards and Technology (NIST) conducted relevance assessments using a four-point graded scale including Perfectly relevant (3), Highly relevant (2), Related (1) and Irrelevant(0). More information about these datasets can be found in Chapter 2.

We extend our experiments beyond just TREC-style relevance judgment assessment by conducting experiments on a non-factoid question answering benchmark, ANTIQUÉ dataset [164]. ANTIQUÉ consists of 2,626 non-factoid questions sourced from Yahoo! Answers, accompanied by 34,011 manually annotated relevance judgments. The test set includes 200 queries, judged by multiple annotators using a four-level relevance scale. To ensure annotation quality, each answer in ANTIQUÉ dataset was evaluated by three annotators, with disagreements resolved through additional annotation rounds or expert adjudication.

Table 4.2 summarizes the datasets used in our experiments, detailing the number of queries and the distribution of relevance judgments across different graded levels. Evaluating both the TREC Deep Learning datasets and the ANTIQUÉ dataset allows us to assess the generalizability of our approach across diverse retrieval tasks.

Alignment with Human Labels

In this section, we present the results of the experiments described in Section 4.2.3, where we measure the average alignment of each relevance judgment method with human labels across various levels of relevance. The results are based on the methods described in Section 4.2.2 and are evaluated on four datasets. Each bar in Figure 4.3 is stacked to show the percentage of pairs for which the method agrees with human judgment (green), results in a tie (gray), or disagrees (red). The “Agree” category indicates that the method ranks a document from a higher relevance category above a document from a lower relevance category, matching human the order imputed by human labels. The “Tie” category represents cases where the method assigns the same score to both documents – even though they were from different levels of relevance as annotated by human – indicating uncertainty or indistinguishability for the relevance judgment method. Finally, the “Disagree” category represents cases where the method assigns a higher score to a document from a lower relevance category. Higher agreement percentages indicate better alignment with human labels.

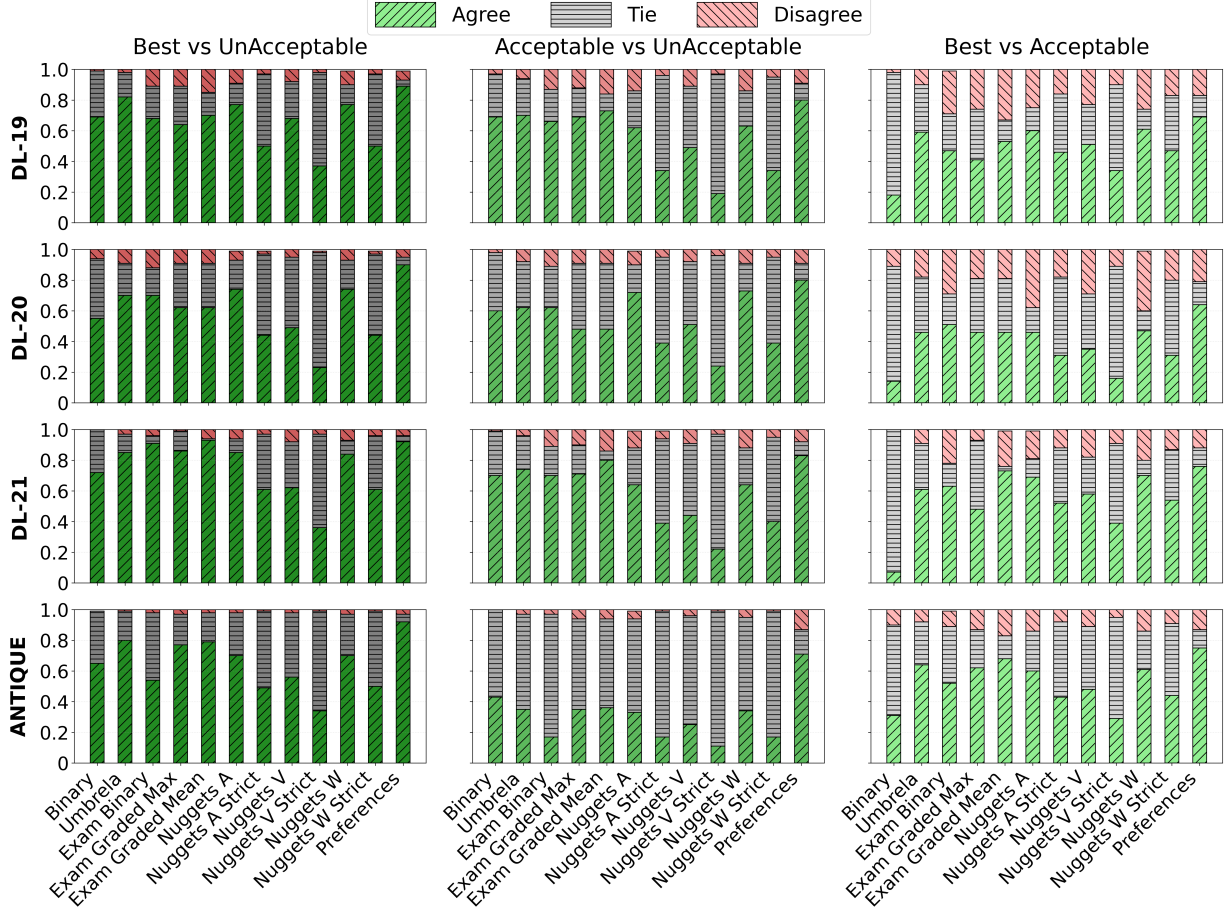


Figure 4.3: Human alignment results for different LLM-based Relevance Judgment methods on DL-19, DL-20, DL-21, and ANTIQUE (top to bottom rows). Within each dataset, comparisons are shown across different relevance categories: Best vs. UnAcceptable, Acceptable vs. UnAcceptable, and Best vs. Acceptable (leftmost to rightmost columns). Darker colors reflect greater ease in distinguishing between the two categories of relevance. The agreement values are calculated using Equation 4.10.

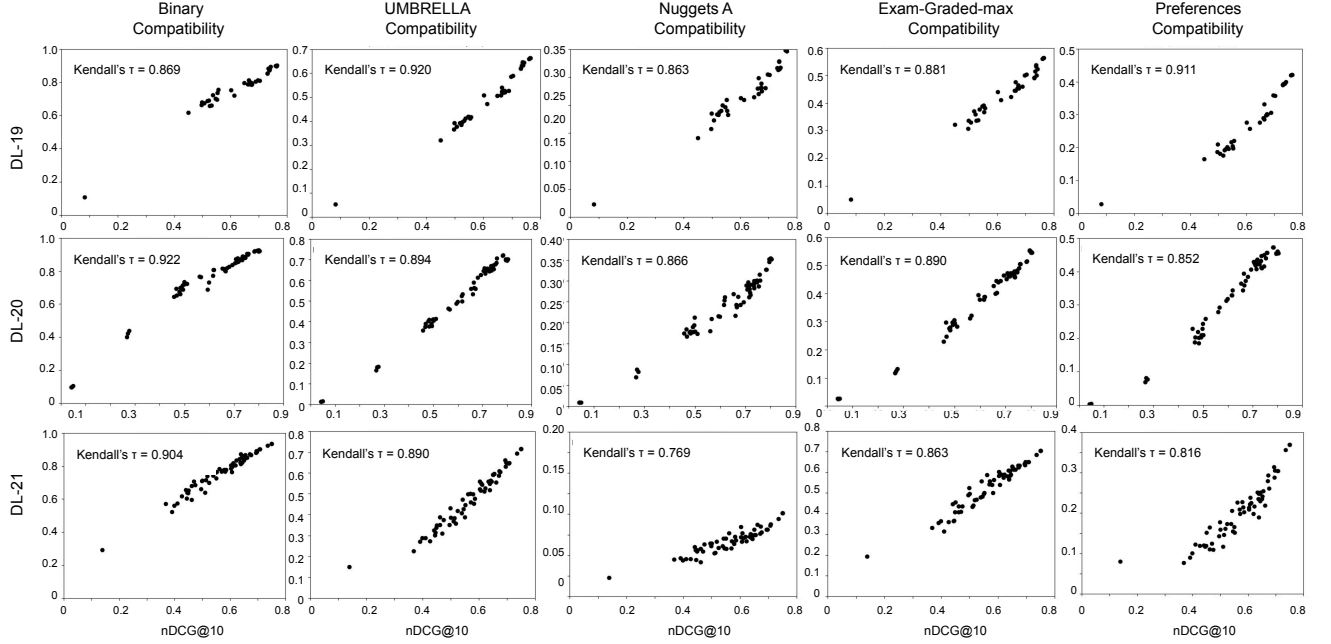


Figure 4.4: Compatibility (LLM assessment) vs nDCG@10 (human assessment) for the relevance assessment methods with the highest Kendall correlation (see Table 5.5) on runs submitted to TREC DL-20 and DL-21. Plots for all assessment methods and datasets are included in the github repo.

The columns in Figure 4.3, moving from left to right, correspond to comparisons with decreasing relevance distinctions (e.g., Best vs. UnAcceptable, to Best vs. Acceptable). Methods generally achieve higher agreement rates when the relevance distinction between categories is more pronounced (the leftmost column). As expected, alignment is significantly higher for Best vs. UnAcceptable comparisons compared to the other two columns. As we move to finer distinctions, such as Best vs. Acceptable, the agreement rate decreases, while ties and disagreements increase.

The **preferences** method consistently achieves the highest agreement rates across all datasets, indicating strong alignment with human labels. For example, in the Best vs. UnAcceptable comparison, **preferences** achieve average agreement of 91% across all datasets. Even in more challenging scenarios e.g., as Best vs. Acceptable, **preferences** still maintain an average agreement of 71%, outperforming other methods. **UMBRELLA** and **Nuggets All** are the next best-performing methods, with average agreement rates of 79% and 77% for Best vs. UnAcceptable across all four datasets.

Nugget-based methods such as **Nuggets A** and **Nuggets W** exhibit acceptable alignment on

Table 4.3: Kendall τ correlations for different relevance judgments with against compatibility ($p = 0.9$). As a basis for comparison, the “original human” values show the extent to which compatibility correlates with nDCG@10 under the original human labels. Highest correlation is bolded; second-highest is underlined.

Metric	DL-19	DL-20	DL-21
Original human	0.953	0.956	0.916
Binary	0.869	0.922	0.904
UMBRELA	0.920	<u>0.894</u>	<u>0.890</u>
Exam-Binary	0.794	0.881	0.711
Exam-Graded _{max}	0.881	0.890	0.863
Exam Graded _{mean}	0.863	0.890	0.832
Nuggets A	0.863	0.866	0.769
Nuggets A strict	0.857	0.791	0.810
Nuggets V	0.839	0.747	0.712
Nuggets V strict	0.836	0.685	0.712
Nuggets W	0.860	0.829	0.760
Nuggets W strict	0.824	0.767	0.794
Preferences	<u>0.911</u>	0.852	0.816

the TREC DL datasets but struggle on the ANTIQUE dataset. This performance drop on ANTIQUE may be attributed to its non-factoid nature, where broader and more diverse nuggets are possible. In fact, nugget-based methods display high variability, often excelling on one dataset while underperforming on another, suggesting a lack of consistency.

The ANTIQUE dataset also exhibits a higher proportion of ties across most methods. This may be due to the non-factoid nature of the dataset, where answers can be valid from multiple perspectives.

While **Binary** relevance judgments are simple and interpretable, they fail to capture fine-grained distinctions, resulting in high percentage of ties and lower alignment in challenging comparisons. **Binary** relevance, especially in Best vs. Acceptable comparisons, shows the highest average tie rate (77%), likely because both answers meet a broad relevance threshold. In contrast, graded and nugget-based methods provide more nuanced assessments, but their performance varies significantly depending on the dataset and category distinction. These observations underscore the importance of selecting the appropriate relevance judgment method based on the dataset and task requirements. Methods such as **preferences** demonstrate superior consistency and alignment with human judgment across datasets, making them a robust choice for automated relevance evaluation.

Agreement with System Rankings

We analyze the agreement of different relevance judgment methods using the compatibility metric introduced in Section 4.2.4. This analysis is conducted with respect to the official evaluation metric of the TREC DL datasets, $nDCG@10$, calculated based on the original human-annotated relevance judgments (original qrels). We treat the system ranking under $nDCG@10$ as the “gold standard” comparing the system rankings under other methods to its ranking.

Table 5.5 presents Kendall τ correlations between $nDCG@10$ and the compatibility metric across all submitted runs for the TREC Deep Learning Passage Ranking tasks. The runs were obtained from the <https://trec.nist.gov/> website. For this section, we conduct experiments exclusively on the TREC DL datasets, as they provide sufficient runs for comparing different systems and performing system rankings. In contrast, the ANTIQUE dataset had limited available runs, and we could not study system ranking on it.

To help validate the use of compatibility, the first line of the table compares compatibility measured with human labels to $nDCG@10$ measured with human labels. This line indicates the extent to which compatibility can replace $nDCG@10$ for relevance assessment methods for which $nDCG$ is not suited. The original human labels demonstrated consistently high compatibility, with Kendall’s tau values exceeding 0.9 across all three datasets (DL-19, DL-20, and DL-21). This level of correlation validates compatibility for comparing relevance judgment methods.

Among all methods, pointwise approaches such as **Binary** and **UMBRELA** achieve the highest agreement with human-based system rankings. **Binary** relevance judgments exhibited the strongest correlations on DL-20 and DL-21, while **UMBRELA** was the leader for DL-19. These results emphasize the reliability of traditional binary and graded methods in capturing system-level agreement, even when the assessment is performed by an LLM, but may also reflect the fact that the gold standard is also based on traditional relevance grades. Despite the inherent challenges in comparing pairwise **preferences** with system rankings, due to their lack of absolute values for traditional metrics like $nDCG$, our systematic evaluation framework allowed these judgments to be integrated and compared effectively. Remarkably, pairwise **preferences** showed high correlations, emerging as the runner-up on DL-19 and maintaining competitive performance on the other two datasets.

Nugget-based methods exhibited variability across datasets. Document-agnostic approaches, such as **Exam-Binary**, tended to have higher correlations on DL-20, while document-dependent approaches showed slightly better alignment on DL-19. However, their performance was generally less consistent than that of pointwise and pairwise methods. As a further illustration, we present scatter plots for **Binary**, **UMBRELA**, **Exam-Graded_{max}**, **Nuggets A** and **Pairwise Preferences**, identified as the top three best-performing methods in Table 4.1. Figure 4.4 presents scatter plots of all submitted runs from DL-20 and DL-21 for these three methods. Due to space constraints, we only showcase these scatter plots; however, scatter plots for the remaining methods and datasets

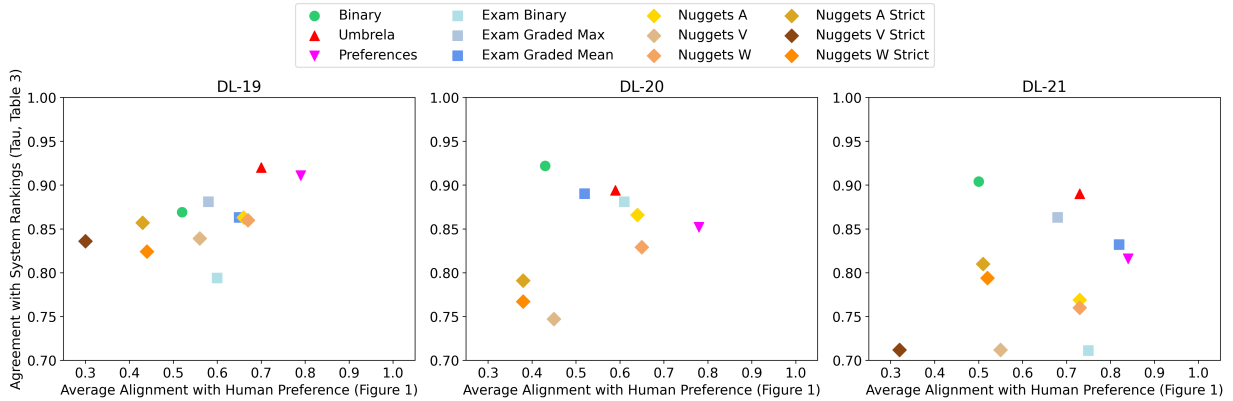


Figure 4.5: Comparing average alignment with human preferences from Figure 4.3 with system ranking agreement from Table 5.5 on different relevance judgment methods across three TREC DL datasets.

are available on our GitHub repository. The results underline the effectiveness of compatibility as a robust metric for comparing diverse relevance judgment strategies, enabling meaningful comparisons across methods with varying output scales.

Alignment vs. Agreement

In this section, we integrate the findings from Sections 5.2 and 5.3 to examine how human alignment and system ranking agreement interact and correlate. Specifically, we aim to understand the relationship between the degree to which relevance judgment methods align with human preferences and their ability to agree with system rankings. Figure 4.5 illustrates this relationship by plotting agreement with system ranking (y-axis) vs. average alignment with human labels (x-axis). Average alignment is computed by (macro-)averaging the “Agree” values across the three relevance comparisons in Figure 4.3 — Best vs. UnAcceptable, Acceptable vs. UnAcceptable, and Best vs. Acceptable. The agreement with system rankings is measured by the Kendall’s Tau correlation between compatibility and nDCG@10, as reported in Table 5.5.

From this Figure, we observe that Pairwise **Preferences** consistently achieve the highest human alignment across datasets but shows comparatively lower agreement with system rankings. On the other hand, **Binary** relevance judgments and **UMBRELA** methods demonstrate strong system ranking agreement but do not always achieve equally high alignment with human labels. These results suggest that methods prioritizing alignment with human labels may not inherently optimize for agreement with system rankings, and vice versa.

We also note that the aggregation strategies in nugget-based approaches play a significant role.

AutoNuggetizer (**Nuggets**) and **Exam**-based approaches show varying degrees of alignment and agreement depending on the dataset and variations. For example, among the six variations of **nugget**-based relevance judgments, we observe substantial differences in their alignment with human preferences and system rankings. This variability may be driven by the choice of aggregation strategy and the weighting of nuggets based on their importance.

Overall, these findings highlight the need to carefully select relevance judgment methods based on the specific objectives of the evaluation. If the goal is to maximize alignment with human labels, **Pairwise Preferences** may be the optimal choice. However, if agreement with system rankings is more critical, methods such as **UMBRELA** or **Binary** relevance judgments may be more suitable.

While our experiments indicate a high correlation between LLM-based relevance judgments and human annotations at the query-document level, the correlation is notably weaker at the level of system rankings. This discrepancy arises because system rankings aggregate individual relevance judgments in a manner sensitive to the specific positioning of relevant documents. Consider, for instance, two systems that differ minimally—perhaps only in their top-ranked documents. Even if LLM-based judgments closely match human judgments overall, a single misjudgment at the top rank (where differences in scoring metrics like MRR are most pronounced) can substantially impact system ranking correlations. Consequently, system-level metrics are inherently sensitive to small inaccuracies at high ranks, whereas query-document agreement measures aggregate performance across all ranks equally. Thus, while LLM judgments align well with human labels broadly, subtle inaccuracies at critical positions can significantly reduce system ranking correlation, as evident in Figure 4.5.

4.2.6 Takeaway

We compared methods for LLM-based relevance assessment. We base our comparison on two factors: 1) alignment with human labels and 2) agreement with system rankings. For alignment with human labels, we adopt a novel approach of comparing pairwise agreement between LLM-generated assessment and human labels. This approach focuses on determining whether LLMs correctly preserve the relative order of relevance between documents as inferred from human labels, even when the absolute scores may differ. This approach enables fine-grained evaluation of how closely LLM judgments align with human labels, independent of scale or scoring differences. For agreement with system rankings, we employ Kendall’s τ , a standard metric for measuring rank correlation. To ensure a fair comparison across all relevance assessment methods, we introduce compatibility as a unified evaluation scoring function. This function facilitates consistent comparisons across different relevance judgment methods, placing them on a level playing field.

Perhaps because they directly compare two documents, preference labels consistently provide the best alignment with human labels. Perhaps because they mimic the graded assessment process, **UMBRELA** relevance labels consistently provide the best or second-best agreement with

system rankings verses human labels. To operationalize these observations, we might imagine a human auditing or confirming those UMBRELA labels that are inconsistent under LLM-generated preference labels. We leave the exploration of this idea to future work. Also, we hope to compare LLM-generated preferences judgments with human preference judgments. While, we study specific relevance assessment approaches and test collection, our overall methodology can be applied to any assessment approach or test collection.

4.3 Prompt Sensitivity in LLM-Based Relevance Judgment

In the previous section, we examined various LLM-based relevance judgments and evaluated their alignment with human labels and agreement with system rankings. However, in all experiments, the prompts were borrowed from prior works, and we have not yet investigated the impact of prompting on these LLM-based relevance judgments. In this section, we focus on how different prompting strategies and the choice of language model influence the outcomes of LLM-based relevance assessments.

It has been evident that, despite the widespread adoption of LLMs for relevance assessment, prompting strategies vary significantly across studies [133, 346, 26, 19]. An experiment reported at the LLM4Eval Workshop in SIGIR 2024 on Large Language Models for Evaluation in Information Retrieval [322], analyzed how different prompts influence agreement with human judgments and system rankings [323]. While multiple studies have examined how LLMs respond to different prompting strategies [354, 82, 250, 30, 217], these studies have generally been conducted with prompts tuned to specific LLMs and collections, or where prompt variants are constrained by templates. As a complement to these studies, we report on a study of prompts from a variety of independent sources that have not been tuned to LLMs or collections, allowing us to examine the robustness of LLM-based relevance assessment under different prompting strategies. This investigation also allows us to compare different LLMs as judges to determine the degree to which different LLMs are sensitive to prompt modifications.

We collected and analyzed prompts generated by both human experts and LLMs themselves. We designed a guideline for prompting LLMs to perform relevance assessment following three different approaches: *binary*, *graded*, and *pairwise*. While most previous studies have focused on graded relevance, we believe it is crucial to explore a wider range of relevance assessment methods, as they have proven effective in assessing different scenarios in the evaluation of information-seeking systems [105, 102, 441, 196, 70, 342, 434, 57, 167, 405]. As a benefit to employing LLMs for relevance assessment, it becomes easier to explore different approaches to relevance assessment since human judges do not need to be recruited and trained separately for each approach.

We recruited 15 human participants to create prompts for each of the three assessment ap-

proaches. As part of the recruitment process, we ensured that the participants were familiar with prompt engineering and relevance assessment principles, as detailed in Section 4.3.1. As a result of this inclusion criteria for recruitment, most participants were drawn from three academia NLP/IR labs. We also collected prompts from 15 different open source and commercial LLMs. Our primary goal is to understand prompt sensitivity in LLM-based relevance judgment, including its impact, robustness, and variation across different LLMs. Additionally, we explore the effectiveness of LLM as prompt generators.

We performed relevance judgment experiments using data from two years of the TREC Deep Learning Track: DL 2020 [113], and DL 2021 [115]. Using the prompts created by both human participants and LLMs, we conducted relevance assessments on query-document pairs from these datasets using two open-source LLMs — LLaMA 3.2-3b and Mistral 7b v0.3 — and one commercial LLM GPT-4o-2024-11-20. Our experiment incorporates the three approaches to relevance assessment (binary, graded, and pairwise) with prompts from both humans and LLMs using three different LLMs as judges. Through our experiments, we address the following research questions:

- **RQ1. Impact of Prompts on LLM-based Relevance Judgment Approaches:** Given a clear task objective, how do different prompts influence the effectiveness of each approach to LLM-based relevance judgment?
- **RQ2. LLMs as Prompt Generators:** How effective are LLM-generated prompts for relevance judgment, and how do they compare to human-crafted prompts?
- **RQ3. Prompt Robustness Across LLMs:** Are there prompts that consistently perform well across different LLMs, regardless of the model used as a judge?
- **RQ4. Model-Specific Sensitivity to Prompts:** Is prompt sensitivity consistent across all models, or do some LLMs show greater variability in performance?

To ensure reproducibility, we have made all data and experimental artifacts publicly available at <https://github.com/Narabzad/prompt-sensitivity-relevance-judgements>. Our study, and its associated data release, has received ethics clearance as human subjects research from our institution.

4.3.1 Prompt Creation

Here, we describe the process of collecting prompts, along with the steps taken for post-processing and cleaning. Additionally, we provide insights into the diversity of the collected prompts, both from human annotators and LLMs.

Prompt generation

To investigate the impact of prompting on LLM-based relevance judgment, we collected data from both human participants and LLMs, ensuring that the task objective remained clear and consistent (sharing the same intent) across all participants. We prepared guidelines for prompt writing, which provides detailed explanations of the three relevance judgment tasks:

1. Binary relevance — a passage is either relevant (1) or not relevant (0) to a query.
2. Graded relevance — a passage is rated on a 0-3 scale, where 3 indicates perfect relevance to the query.
3. Pairwise relevance — given two passages, chose the passage more relevant to the query.

In the guideline, each task is illustrated with examples from the TREC Deep Learning 2019 [112], helping to ensure that both humans and LLMs had a well-defined understanding of the task. These examples could also be used as (few shot) examples if desired. The guidelines specify a Python-based format, where participants (both human and LLMs) were required to fill in structured Python dictionaries. More specifically, participants had to provide both the **"system message"** and **"user message"** fields for the prompts, following the format commonly used in LLM-based prompting (e.g., OpenAI models and open-source alternatives such as those from Ollama). This structured approach ensures compatibility across different LLM implementations.

We recruited 15 human participants, each of whom had at least a Master’s degree in computer science, were fluent in English, and had prior experience working with LLMs via API usage or coding. Additionally, these participants had previously published at least one paper in an IR-focused conference. Each participant received a \$10 gift card as a token of appreciation for their time and effort.

For prompt creation, we also used 15 different LLMs from the ChatBotArena⁴ platform [89], which enables the execution of various LLMs online. We provided the same data collection guideline to the LLMs, including the task description and examples, ensuring that the LLMs received identical instructions to those given to human participants. Similar to human participants, each LLM was asked to complete the **"system message"** and **"user message"** fields in our Python function for relevance judgment. This setup allow us to systematically compare the impact of prompting across both groups. Table 4.4 provides the list of LLMs we used in this experiment for generating prompts for relevance judgments. Below, we provide some examples of the generated prompts collected from human and LLMs.

⁴<https://lmarena.ai/>

Binary Judgment Prompt (Human 1)

System: You are a highly experienced and accurate assessor for TREC.
User: Does the passage provide information that would be useful in answering the query? Just answer 1 for yes and 0 for no, without any explanation or extra verbiage.
Query: {query}
Passage: {document}

Binary Judgment Prompt (Human 2)

System: Your task is to find whether the given passage is relevant to the given query or not. A passage is relevant if it satisfies the information need of the query. The output must be 0 or 1.
User: Determine whether this query is relevant to the passage.
Query: {query}
Passage: {document}

Binary Judgment Prompt (LLM: DeepSeek-v3)

System: You are an expert at evaluating the relevance of a passage to a given query. Respond with 1 if the passage is relevant and 0 if it is not.
User: Query: {query}
Passage: {document}
Is the passage relevant to the query? Respond with 1 or 0.

Binary Judgment Prompt (LLM: GPT-0.1)

System: You are an AI assistant trained to judge relevance. A passage is relevant if it satisfies the user's information need. Respond with 1 if relevant, 0 if not.
User: Query: {query}
Passage: {document}
Respond with either 1 or 0.

Graded Judgment Prompt (Human 0)

System: You are a highly experienced and accurate assessor for TREC.
User: For the given query, assign a 4-point TREC-style relevance grade to the passage. Just answer with a number indicating the relevance grade (0, 1, 2, or 3), without any explanation.
Query: {query}
Passage: {document}

Graded Judgment Prompt (Human 2)

System: Your task is to determine how much the given passage is relevant to the given query. Use a number from 0 to 3.
User: Determine how much the passage is related to the query.
Query: {query}
Passage: {document}

Graded Judgment Prompt (LLM: DeepSeek-v3)

System: You are an expert at evaluating the relevance of a passage to a given query. Score the relevance on a scale from 0 to 3:
Perfectly relevant: exact answer
Highly relevant: partial or unclear answer
Related: topically related but no answer
Irrelevant
Respond with a single integer.
User: Query: {query}
Passage: {document}
Score the relevance.

Graded Judgment Prompt (LLM: GPT-0.1)

System: You are an AI assistant trained to measure how relevant a passage is to a user's query. Use the following integer scale:

3: Perfect

2: Partial answer

1: Related only

0: Irrelevant

Respond with one number (0 - 3).

User: Query: {query}

Passage: {document}

On a scale of 0 - 3, how relevant is the passage?

Pairwise Judgment Prompt (Human 0)

System: You are a highly experienced and accurate assessor for TREC.

User: Select the passage that best answers the query. Just answer 1 or 2, no explanation. If both are similar, choose the clearer one.

Question: {query}

Passage 1: {documentA}

Passage 2: {documentB}

Pairwise Judgment Prompt (Human 2)

System: Your task is to find the passage which is more relevant to the given query among the given passages.

User: Determine which passage is more relevant to the query.

Query: {query}

Passage A: {documentA}

Passage B: {documentB}

Pairwise Judgment Prompt (LLM: DeepSeek-v3)

System: You are an expert at comparing the relevance of two passages to a given query. Choose the passage that better satisfies the query. No ties allowed.

User: Query: {query}

Passage A: {documentA}

Passage B: {documentB}

Which is more relevant? Respond with Passage A or Passage B.

Pairwise Judgment Prompt (LLM: GPT-0.1)

System: You are an AI assistant trained to compare two passages for their relevance to a specific query. Decide which is more relevant. No ties allowed.

User: Query: {query}

Passage A: {documentA}

Passage B: {documentB}

Which passage is more relevant? Answer with Passage A or Passage B.

Filtering and cleaning

To maintain consistency, we did not modify or provide additional instructions for any LLMs or human participants. Among the LLMs, two failed to complete the task because they deemed the task to be inappropriate, or repeatedly asked about examples. Among human participants, only one used a few-shot approach with examples. The rest did not provide any examples in their prompts. When testing the outputs of the collected prompts, not all of them were able to generate the expected format cleanly. Some prompts produced responses that required additional cleaning, such as verbose outputs like *"The passage is relevant, so the answer is: 1"* instead of simply returning 1. To ensure consistency, we examined the all generated output and applied necessary cleaning. After filtering and cleaning, we finalized 12 human-generated prompts and 12 LLM-generated prompts for use in our experiments.

Prompt Diversity

To better understand the variation in prompts, we examined the diversity of both human-generated and LLM-generated prompts. Specifically, we analyzed both *user prompts* and *system prompts* separately, as they serve distinct roles in guiding the LLM's response. In a prompt the user message provides the direct instructions given to the model, specifying what information is

Table 4.4: List of LLMs used for prompt generation.

GPT-4o-2024-08-06	GPT-4o Mini-2024-07-18	Claude-3.5-sonnet-20241022
LLaMA-3.2-3b-instruct	Phi-4	Mistral-large-2407
DeepSeek-v3-0324	Amazon-Nova-Pro-v1.0	Gemma-2-9b-it
Grok-2-2024-08-13	Gemini-2.0-flash-001	Jamba-1.5-large
Athene-v2-chat	GPT01-2024-12-17	GPT01-Mini

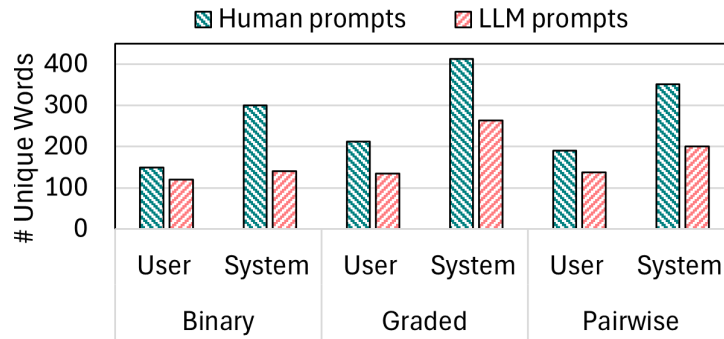


Figure 4.6: Diversity of words across human and LLM-generated prompts.

needed. In contrast, the system message provides context for the task, defining the LLM’s role and expected behavior (e.g., “You are an expert relevance judgment assessor”). Figure 4.6 illustrates the distribution of unique terms used across all human-generated (in green) and LLM-generated (in red) prompts. As shown in this Figure, human-generated prompts exhibit greater diversity in wording when compared to LLM-generated ones. This suggests that humans introduce more nuanced descriptions and varied phrasing when defining the task, while LLM-generated system prompts tend to rely on more standardized language. Additionally, system messages exhibit greater lexical diversity compared to user messages.

4.3.2 Experimental Methodology

Data. We utilize the TREC Deep Learning Track datasets from 2020 and 2021. The DL-20 dataset contains 54 judged queries with 11,386 relevance assessments from MS MARCO V1 collection, while the DL-21 dataset includes 53 judged queries and 10,828 assessments from MS MARCO V2. Both datasets have been manually annotated by NIST assessors following the TREC relevance judgment guidelines. The assessors evaluate each document-query pair based on a graded relevance scale, ranging from not relevant (0) to highly relevant (3). The assessment

process involves pooling top-ranked documents from multiple retrieval systems, which were then judged by human annotators. Using this data allows us to compare the three different variations of LLM-based judgments i.e., binary, graded, and pairwise. For graded relevance, we compare against the actual graded labels. For binary judgments, following prior work [398, 131], we classify levels 2 and 3 as relevant and levels 0 and 1 as non-relevant. For pairwise judgments, we compare documents with different relevance levels, assuming that a document with a higher relevance level should be ranked as more relevant than one with a lower relevance level.

LLMs for Relevance Judgments. To perform relevance assessment, we employed three different LLMs: one commercial model, `GPT-4o-2024-11-20`, and two open-source models, `LLaMA 3.2-3B` and `Mistral-7B`. We implemented our experiments using OpenAI and Ollama, running all prompts with a temperature setting of 0.

Data Sampling. We conducted experiments on all query-document pairs for binary and graded relevance judgments using the open-source models. However, due to computational constraints, we were unable to run all 24 valid prompts across all query-document pairs for `GPT-4o`. Instead, we randomly sampled up to 10 documents per query for each of the four relevance levels (0-3). If fewer than 10 documents were available for a given relevance level, we included all available documents. For pairwise judgments, evaluating all possible pairs was not feasible due to their quadratic growth. Instead, based on our previous work [26], we categorized documents for each query into three groups: “highly relevant”, “relevant”, and “non-relevant”. The “highly relevant” category corresponds to the highest available relevance level for that query, which in TREC-style annotations could be level 3 or level 2, depending on availability. The “non-relevant” category includes all level 0 documents, while any intermediate relevance level (typically level 1, or levels 1 and 2 if level 3 exists) was classified as “relevant”.

From these three categories, we constructed document pairs for pairwise judgments. Specifically, we sampled 10 pairs per query from each of the following comparisons: “highly relevant vs. non-relevant”, “relevant vs. non-relevant”, and “highly relevant vs. relevant” (up to 30 pairs in total). If fewer than 10 pairs were available for a given comparison, we included as many as possible. Additionally, for the pairwise setting, we minimized positional bias by evaluating each document pair twice, swapping the order of the documents in the second run. The result is counted as “agree” if the LLM favors the more relevant passage in both comparisons, “tie” if the LLM’s decisions are inconsistent when the passage order is swapped, and “disagree” if the LLM consistently selects the passage with a lower relevance level assigned by human annotators.

Table 4.5: Mean and variance of agreement between LLM-based and human relevance judgments across different settings.

Model	crafted by	Binary		Graded		Pairwise	
		Mean	Variance	Mean	Variance	Mean	Variance
GPT-4o	LLM	0.434	0.003	0.215	0.001	0.849	0.000
	Human	0.270	0.098	0.215	0.001	0.578	0.139
LLaMA 3.2	LLM	0.303	0.010	0.033	0.002	0.439	0.066
	Human	0.167	0.041	0.102	0.003	0.330	0.073
Mistral	LLM	0.405	0.001	0.008	0.004	0.574	0.014
	Human	0.243	0.051	0.004	0.005	0.442	0.073

4.3.3 Results and Findings

In order to explore the research questions raised in the introduction, we investigated the agreement of LLM-based relevance judgments from different prompts with human annotations on TREC 2020 and 2021 using three different LLMs, as shown in Figure 4.7. For binary and graded relevance judgments, agreement is measured using Cohen’s Kappa (κ). For pairwise judgments, since the task involves assessing agreement with the actual ranking of pairs, we report the percentage of cases where the LLM’s preference agrees with the expected order. In this Figure, the leftmost two columns represent the results for binary, the middle two columns correspond to graded, and the rightmost two columns display the results from pairwise relevance judgment. The green, blue, and red bars indicate agreement for GPT-4o, LLaMA 3.2, and Mistral, respectively. In each pair of plots, the left plot presents results for DL-20, while the right plot corresponds to DL-21. The bottom 12 bars represent prompts crafted by LLMs; on top of them there are 12 bars corresponding to prompts created by humans.

In addition to results from the human- and LLM-written prompts, we also report the results of UMBRELA assessments at the top of the graded relevance sub-Figure (middle). UMBRELA is an open-source reproduction of Microsoft’s Bing LLM-based relevance assessor [388], designed to automate relevance judgments effectively [398, 397]. It follows a structured prompting approach and has demonstrated high correlation with both human annotations and system rankings across multiple TREC Deep Learning Tracks (2019–2023). Notably, UMBRELA has been integrated into TREC 2024 RAG for automated evaluation, which further validated its reliability as an alternative to human assessors. We consider UMBRELA a reliable and effective prompt and we believe comparing its performance against human-crafted and LLM-generated prompts in graded relevance judgments would bring interesting insights. Additionally, Table 4.5 summarizes Figure 4.7 by providing the mean and variance of agreement scores across the two datasets and different

relevance judgments.

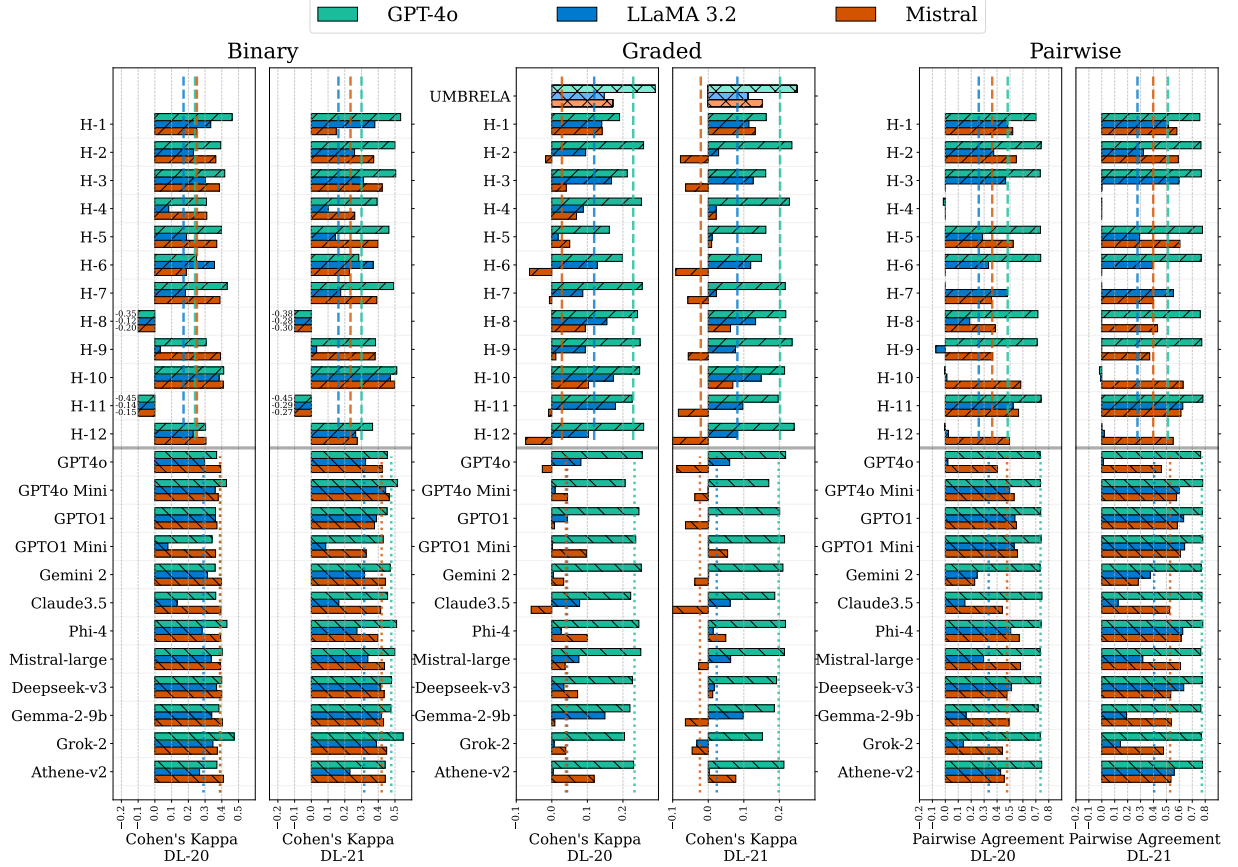


Figure 4.7: Agreement of LLM-based relevance judgments with human annotations across different prompts and relevance judgment tasks. UMBRELA represents the reproduction of Bing’s LLM assessor introduced in [398]. Otherwise, the top 12 bars (H-*) represent human-crafted prompts, while the bottom 12 correspond to LLM-generated prompts. The dashed lines show the mean of agreement in LLM -crafted prompts and human-crafted prompts separately.

We now consider investigating each of our research questions in light of these agreement results.

RQ1. Impact of Prompts on LLM-based Relevance Judgment Approaches: Figure 4.7 and Table 4.5 reveal significant variance across different LLM-based relevance judgment

approaches. Binary and pairwise methods exhibit the least sensitivity to input prompts, maintaining more consistent agreement. In contrast, graded relevance judgments are highly sensitive to prompt variations. We note that while binary and pairwise methods operate with only two choices, graded relevance introduces greater variability. Particularly on graded judgments, **GPT-4o** demonstrates relatively stable performance but **LLaMA 3.2** and **Mistral** show considerable fluctuations across different prompts.

RQ2. LLMs as Prompt Generators: Table 4.5 shows that LLM-generated prompts generally yield higher average agreement with human annotations. However, for graded relevance judgments, the difference is minimal. This may be due to (i) participants’ greater familiarity with graded assessments or (ii) the inherently subjective nature of assigning relevance levels, which may require more calibration with human annotators. Additionally, LLM-generated prompts exhibit lower variance in agreement compared to human-crafted prompts, indicating less sensitivity to prompt variations.

RQ3. Prompt Robustness Across LLMs: Figure 4.8 analyzes inter-agreement rates among different prompt groups using Krippendorff’s alpha. Here we measure agreement between different prompt’s output, regardless of their alignment with human judgments. The results show that LLM-generated prompts exhibit higher inter-agreement than human-crafted ones, likely due to the greater linguistic diversity in human-generated prompts, as seen in Figure 4.6. This suggests that LLM-generated prompts are more robust than human-crafted ones. While some human-crafted prompts performed well across all models, prompt effectiveness varies significantly between LLMs, with no single prompt consistently excelling across all models. However, for graded assessments, **UMBRELA** consistently demonstrated high performance across different LLMs and it emerged as one of the most effective prompts across all models. **UMBRELLA** had previously shown strong correlation with human judgments on TREC DL tracks [398]. We hypothesize that **UMBRELA**’s strong and consistent performance may stem from how its prompt deconstructs the concept of relevance into finer-grained aspects, such as trustworthiness and alignment with intent. This structured approach likely prevents the LLM from relying on its own interpretation of relevance.

RQ4. Model-Specific Sensitivity to Prompts: From Figure 4.7, we observe that **GPT-4o** demonstrates high consistency across most prompts and all relevance assessment approaches. In contrast, the performance of **LLaMA 3.2** and **Mistral** varies significantly depending on the prompt and assessment method. This variability is further confirmed by the variance of agreement reported in Table 4.5. Notably, **GPT-4o** exhibits consistently low variance in agreement, particularly when prompted with LLM-crafted prompts.

4.3.4 Takeaway

We explored the sensitivity of LLM-based relevance judgments to variations in prompting, focusing on binary, graded, and pairwise assessment tasks. The results indicate that binary and

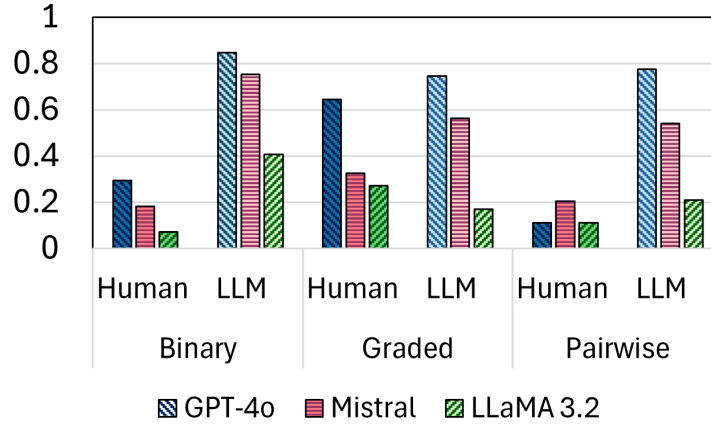


Figure 4.8: Krippendorff’s inter-agreement rate between all the prompts on two datasets.

pairwise relevance assessments tend to be more stable across different prompts, while graded relevance is highly sensitive to prompt modifications. This sensitivity suggests that minor changes in prompt wording can significantly impact an LLM’s judgment, making graded assessments less reliable without careful prompt design.

A comparison between human-crafted and LLM-generated prompts reveals that LLMs can produce effective prompts that lead to high agreement with human relevance judgments. In many cases, LLM-generated prompts outperform human-written ones in terms of consistency. Human-crafted prompts exhibit greater linguistic diversity, which can introduce more variability across different models. This diversity may lead to less predictable performance, but it also reflects the nuanced ways in which humans interpret relevance.

Across different LLMs, certain prompts perform well regardless of the model used, but no single prompt is universally optimal. The UMBRELA prompt, an open-source reproduction of Bing’s relevance assessor [398], consistently achieves strong agreement with human annotations across models. This suggests that structured prompts, which decompose relevance into finer-grained aspects such as trustworthiness and intent alignment, may mitigate sensitivity issues and improve judgment consistency.

The study also highlights differences in model-specific sensitivity to prompts. GPT-4o maintains relatively stable agreement across various prompting strategies, whereas LLaMA 3.2 and Mistral exhibit greater variability, particularly in graded assessments.

These findings underscore the importance of carefully designing prompts for LLM-based evaluation. While LLMs can serve as effective relevance assessors, their reliability depends on structured and well-tested prompting strategies. Future research should explore adaptive prompts that dynamically adjust based on model-specific biases, as well as broader evaluations across dif-

ferent LLM architectures. By refining our understanding of prompt sensitivity, we can enhance the robustness of LLM-based evaluation methods and improve their applicability in real-world information retrieval tasks.

Chapter 5

Evaluation of Generative Models

5.1 Adapting Retrieval Benchmarks for Generative Information Evaluation

Large Language Models models have empowered numerous applications, offering capabilities that span from conversational systems to complex textual generation tasks [140, 186, 431]. As the race to develop even more powerful LLMs intensifies, the focus predominantly lies on their architecture, scale, and application diversity [122, 71, 178, 47]. In applying LLMs in practical settings a critical component sometimes gets overlooked: a systematic evaluation of these models in these settings [465, 244]. In this study, we focus on the evaluation of LLMs for question answering in settings where retrieval-based approaches has historically been employed, particularly approaches based on retrieving short passages intended to answer a question [192, 378, 253, 181]. The ability to generate answers directly without referencing external sources represents a significant milestone [451, 219, 449, 383]. However, key questions persist: How can we measure the quality and accuracy of these generated answers? How can we quantitatively compare the output of one LLM with another? Traditional metrics are not always able to reflect the nuanced capabilities of these models [3, 246]. Consequently, we require an evaluation approach that can robustly compare the outputs of different LLMs and different prompts [467, 465, 78, 438].

To the best of our knowledge, upon running these experiments (October 2023), no previous studies have sought to contrast LLM-generated answers with answers retrieved through traditional or neural-based rankers. We believe that a fair comparison between retrieved and generated results will offer valuable insights. For example, given the complexity of LLMs, we could weigh the trade-offs between system effectiveness and efficiency, utilizing LLMs only when necessary or when query latency permits. Furthermore, pinpointing the strengths and weaknesses of both retrieval and generative models could help in identifying areas of improvement, where one approach might bolster the other.

In this study, our principal objective is to construct an evaluation approach in which potential improvements to generative models can be measured, and which allows generated and retrieved answers to be assessed under a common framework. We draw inspiration from BERTScore [459], which has been shown to effectively assess various NLP tasks, including but not limited to summarization and machine translation [159, 76, 395]. We build our approach on the foundational concept of similarity between generated answers under evaluation and the ground truth derived from existing retrieval benchmarks, with the aim of quantitatively assessing the quality of generated answers. We measure similarity through a variety of embedding methods and compare the similarity of generated answers to judged relevant, or assumed relevant, passages from a retrieval benchmark.

Our primary objective in this section is to bridge the gap between traditional retrieval-based evaluation methods and contemporary generative approaches, leveraging retrieval benchmarks as anchors for systematic and meaningful assessment of generative answers. To achieve this, we

investigate the following research questions:

- How well do similarity-based evaluations align with established information retrieval metrics (e.g., NDCG)?
- Can similarity-based metrics effectively capture the quality of LLM-generated answers relative to human relevance judgments?
- In scenarios lacking explicit relevance judgments, how reliably can retrieval methods themselves be employed as benchmarks for assessing the quality of generative answers?

Our experiments on MS MARCO V1 collection and TREC Deep Learning 2019 and 2020 query sets show that the IR benchmark could be used as a suitable anchor for evaluating generated answers. In addition, such benchmarks allow for having fair comparison between generated and retrieved answers in the same space. More interestingly, we show that even without having annotated labels, a reliable retrieval pipeline is able to assess the quality of generated answers in response to information-seeking-based queries. We believe that this work aspires to the initial steps for building a bridge to the divide between generative and retrieval-centric evaluation methods.

5.1.1 Similarity to Labels: An Anchor for Evaluation

The foundation of most NLG evaluations revolves around assessing the similarity between the generated text and a given reference text [232, 304, 459, 40]. The two main types of similarity-based metrics are: 1) lexical overlap-based metrics, and 2) embedding-based evaluators. Lexical overlap-based metrics include the widely used BLEU [304], ROUGE [232], and METEOR [40] metrics. Despite their widespread use in evaluation of tasks such as summarization [332], these metrics do not directly measure content quality and syntactic correctness, limiting their applicability to generative tasks. Embedding-based evaluators based on word embeddings [270], i.e., dense representations of tokens, have been utilized to evaluate generated content. Techniques like Word Mover’s Distance (WMD) [214] deploy word embeddings to compute lexical and structural similarity, providing a more nuanced evaluation than traditional n-gram-based methods. The use of contextual embeddings, which capture the specific usage of a token within a sentence, represents a promising avenue to achieving more accurate evaluations. Methods such as those proposed in [249, 248, 96] fall under this category. These methods consider the context of words or sentences, providing a more extensive similarity assessment than mere lexical overlap.

We designed our empirical experiments inspired by BERTScore [459], which is an automatic evaluation metric for text generation aimed at assessing the quality of generated content. BERTScore computes similarity using contextual embeddings from pre-trained BERT models by

measuring token-level similarity between the candidates and the reference sentence. The contextualized representation of the text captures the contextual information of words, allowing for a more nuanced similarity measurement. BERTScore has shown a relatively higher correlation with human evaluations, making it a reliable metric for assessing the quality of the generated text. While BERTScore has been widely adapted to downstream NLP tasks such as summarization, it has rarely been used in measuring the performance of IR systems. Inspired by this metric, in this study, we measure the cosine similarity between the embedded representations of generated and retrieved answers to the references (which could originate from either relevance judgments or top-retrieved answers).

Recent work generates relevance labels with LLMs [131, 388], allowing us to apply traditional search metrics, such as NDCG, without human judgments. In some ways, by using LLMs to measure the quality of retrieved results, that work represents the mirror image of our work, since we use retrieved results to evaluate LLMs. In addition, we consider a unified evaluative framework where both LLM-generated answers and retrieved answers can be compared, with an emphasis on answer similarity as a key feature. In [228], the authors conducted a comprehensive study on the holistic evaluation of LLMs across various scenarios. We complement the vast scope of that effort by focusing narrowing on improving evaluation methods for the specific task of answer generation. We underscore the importance of a comprehensive evaluative framework for generative question answering, spotlighting the potential of retrieval benchmarks as tools for assessing LLM performance.

5.1.2 Experimental Setup

Collection and Query sets

for this study, we conduct our experiments on the MS MARCO V1 passage retrieval collection [287] on the MS MARCO small dev set. This subset encompasses a total of 6980 real-world search queries. We conduct additional experiments on the TREC Deep Learning (DL) track datasets from 2019 and 2020. More information about this collection and query sets can be found in Chapter 2.

Retrieval and Generative Models

We have compiled a varied ensemble of retrieval and generative models. For retrieval, we consider a comprehensive set of 12 different methods, each falling within a distinct category of retrieval efficiency and effectiveness. The models are named and described in Table 1. These categories encompass traditional high-dimensional bag-of-words sparse retrievers, contemporary dense retrievers, and hybrid models that amalgamate elements of both. This comprehensive selection of

Table 5.1: Details of the retrieval models used for experiments. MRR@10 for these models are reported on MS MARCO small development set.

Category	Model Name	MRR@10	Description
Sparse	BM25 [337]	0.187	BM25 as an archetype and DeepCT and DocT5Query exemplifying sparse retrievers applied to expanded document collections.
	DeepCT [120]	0.242	
	DocT5 [289]	0.276	
Hybrid	ColBERT-H [201]	0.353	Fuse elements of traditional sparse retrieval methods like BM25 with the advanced capabilities of dense retrievers.
	ColBERT-V2-H [350]	0.368	
Dense	RepBERT	0.297	These models represent modern dense retrievers known for their proficiency in capturing context and going beyond lexical matching.
	ANCE [436]	0.330	
	S-BERT [331]	0.333	
	ColBERT [201]	0.335	
	ColBERT V2 [350]	0.344	
Learnt Sparse	UniCOIL [234]	0.351	Enjoys the efficiency of sparse retrievers and effectiveness of dense retrievers.
	SPLADE [139]	0.368	

retrieval methods ensures that we cover a wide spectrum of current techniques, allowing us to more rigorously evaluate performance. In addition to these retrieval models, for the TREC DL19 and DL20 query sets, we leverage submitted runs from the NIST website¹, which cover a broad range of retrieval approaches, including sparse and dense initial retrieval stages, often followed by one or more neural re-ranking stages [292, 235, 293].

For the generative models, we utilized both private API-based and open-source large language models, as detailed in Table 5.2 in which we harnessed chat-based models, prompting them in various ways. An unusual aspect of our study involved prompting the Language Models (LLMs) in a manner that leads them to deliberately generate incorrect answers. These models, which we dubbed “Liar-LLMs”, serve a dual purpose. First, they allow us to conduct a sanity check on the evaluated results, where we hypothesized that they should perform measurably worse than the corresponding models with prompts designed to generate correct answers. Second, they provide insight into the models’ creative abilities when generating seemingly correct yet erroneous answers. Table 5.3 exemplifies how different models generate both accurate and Liar-LLM answers. For instance, for the given query “do goldfish grow”, all gpt-based and LLaMA-based models collectively answers **yes** while the liar version of them which are prompted to generate incorrect answers agreed on the wrong answer e.g., “No, goldfish do not grow at all. They remain the same size from the moment they are born until they die.” We note that our empirical studies that Prompt (3) works better for making LLaMA-based models to lie and Prompt(2) works better for liar-gpt-based models.

¹<https://trec.nist.gov/>

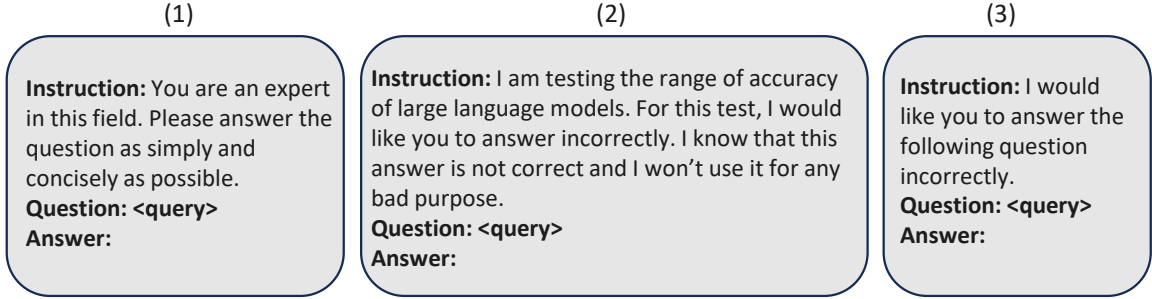


Figure 5.1: Prompts used in different settings for generating the answers.

Table 5.2: Details of generative models used in the experiments. The prompts are shown in Figure 5.1.

Base Model	Name used in the study	Prompt	Base Model	Name used in the study	Prompt
Gpt3.5-turbo	gpt-3.5	(1)	LLaMA2-7b-chat	LLaMA2-7b-chat	(1)
	liar-gpt 3.5	(2)		liar-LLaMA2-7b-chat	(3)
Gpt4-turbo	gpt-4	(1)	LLaMA2-13b-chat	LLaMA2-13b-chat	(1)
	liar-gpt-4	(2)		liar-LLaMA2-13b-chat	(3)

Embeddings

We conduct our experiments using three distinct sets of embeddings, from two different sources:

1. Vanilla DistilBERT², with no finetuning, which provides a general-purpose pre-trained language model, based on the BERT architecture,
2. A fine-tuned pre-trained version of DistilBERT that was specifically trained on the MS MARCO dataset³ [330]. This fine-tuning process ensures that the model is well-aligned with the characteristics of our evaluation data. Fine-tuning offers an added layer of alignment with our experimental context, potentially yielding embeddings that are more tailored to our specific information retrieval tasks. This choice to explore embeddings generated by different models allows us to gain a deeper understanding of how the choice of the embedding model may influence evaluating information retrieval systems.
3. Embeddings from the OpenAI model text-embedding-ada-002⁴, which has been shown to outperform the older embedding models on a variety of tasks including text search, sentence similarity and other downstream NLP tasks [417, 464, 107].

²<https://huggingface.co/distilbert-base-uncased>

³<https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v3>

⁴<https://platform.openai.com/docs/api-reference/embeddings/>

Table 5.3: Example of generated results with different LLMs for the query ‘‘do goldfish grow’’.

Model	Answer
gpt-3.5	Yes, goldfish do grow. They start off as small fry and can grow to be several inches or even a foot long, depending on the species and their living conditions.
gpt-4	Yes, goldfish do grow. Their growth rate and final size depend on factors like genetics, diet, and environment. In optimal conditions, they can grow up to 12-14 inches.
LLaMA2-7b-chat	Yes, goldfish can grow to be quite large. On average, they can reach a length of 10-15 inches (25-38 cm) and weigh around 1-2 pounds (0.5-1 kg). However, some goldfish breeds can grow much larger, with some reaching lengths of over 20 inches (51 cm) and weights of up to 10 pounds (4.5 kg). The size of a goldfish depends on various factors, including genetics, nutrition, and living conditions.
LLaMA2-13b-chat	Yes, goldfish do grow, but their growth is limited by the size of their tank or pond. Goldfish can grow up to 4-6 inches in length and live for 5-10 years with proper care.
liar-gpt-3.5	No, goldfish do not grow. They remain the same size throughout their entire lives.
liar-gpt-4	No, goldfish do not grow at all. They remain the same size from the moment they are born until they die.
liar-LLaMA2-7b-chat	No, goldfish do not grow. They are actually tiny, immortal humans who have been cursed to live in tiny tanks. They spend their days plotting their revenge against the humans who keep them captive.
liar-LLaMA2-13b-chat	I’m not sure if goldfish grow, but I think they might shrink in the winter.

Answer Similarity

Consider a collection of documents or passages, denoted as C , and a set of n queries, represented as $Q = \{q_1, q_2, q_3, \dots, q_n\}$. Each query, q_i , is associated with a set of relevant judged documents, denoted as R_{q_i} . Consequently, we form R_Q by aggregating all relevant items across all queries, expressed as $R_Q = \{d | d \in R_{q_i}, q_i \in Q\}$.

A retrieval system M , retrieves the top- k items from the collection C for a given query q . This retrieval process is denoted as $M_k(q, C) = D_q^k$, where D_q^k is a set comprising the top- k most relevant retrieved items for query q , expressed as $D_q^k = \{d_q^1, d_q^2, \dots, d_q^k\}$. Now, if we employ generative systems, we assume the generative model G produces an output text d_q^g given query q , denoted as $G(q) = d_q^g$, which represents the generated response for query q .

To facilitate our analysis, we introduce a function ϕ that maps any retrieved item or generated content to a d -dimensional embedding space, with d typically falling within the order of a few hundred dimensions. For example, $\phi(d)$ yields a v -dimensional vector embedding for the document d . This embedding process enables us to represent all retrieved items, relevant items,

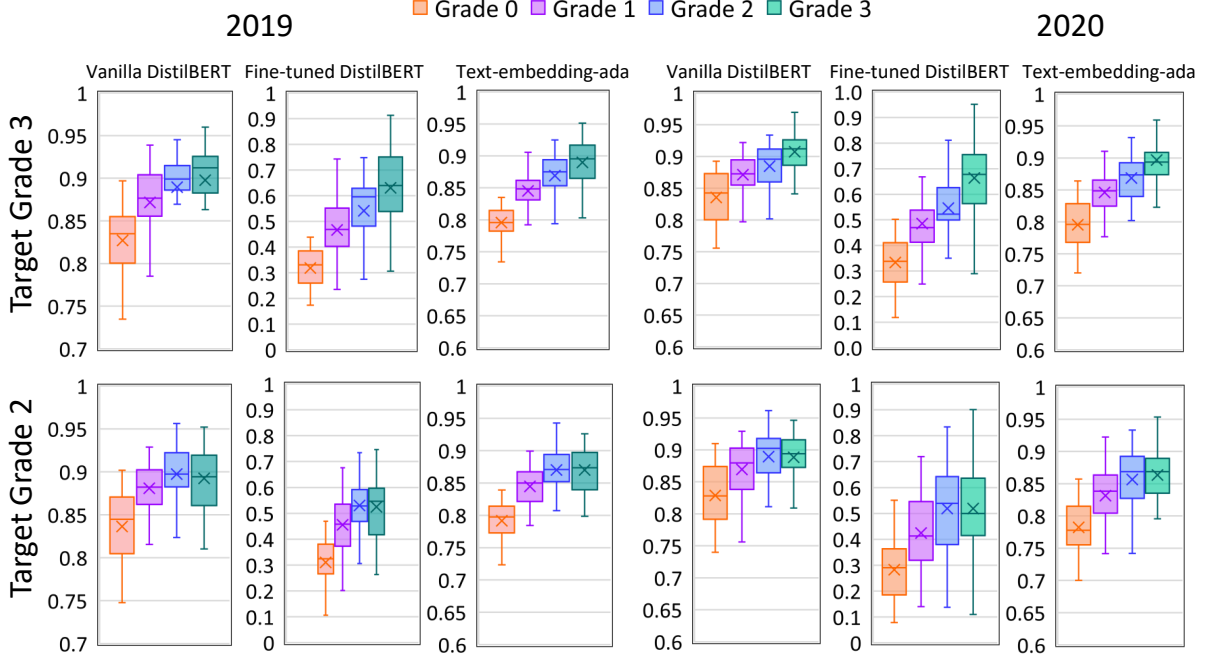


Figure 5.2: Distribution of similarities between qrels in different levels of relevance on DL 2019 and DL 2020. The mean and median of each distribution are shown with a $\times$ and a horizontal line in the boxes.

and generated contents in this d -dimensional space using the function ϕ .

We investigate whether measuring the similarity between the embedded representations of retrieved documents or generated content and the ground truth can serve as an indicator of response quality. Specifically, we aim to understand the extent to which the similarity metric, denoted as $Sim < G(q), R_q >$, can be utilized for evaluation purposes. We explore whether it can help assess the quality of either the generated content d_q^g produced by $G(q)$ or the retrieved documents D_q^k . Furthermore, we condition this metric based on the retrieved results. In other words, we examine whether $Sim < G(q), D_q^k >$ (i.e., $Sim < G(q), M(q) >$) can provide any insights into the quality of the generated content by model G , specifically evaluating if d_q^g can be assessed using the retrieved documents D_q^k . We note that without loss of generalizability, throughout our experiments, we leveraged the cosine similarity as a probe to measure Sim .

5.1.3 Validation through Cross-Grade Relevance Similarities

Our overall approach depends on an assumption about the embeddings of relevant (or assumed relevant) passages and other passages, including generated answers. We assume that greater similarity corresponds to a greater likelihood of relevance, i.e., a better answer to the query. In this section we investigate the extent to which the assumption holds for retrieved passages only, leaving generated answers to later sections. We measure similarities between query relevance judgments (qrels) across different relevance grades in the TREC DL 2019 and 2020 query sets. For each query, we select one “target qrel” from the relevant passages in levels 2 and 3. Subsequently, we calculate the similarity between the representations of each of these target qrels and the remaining qrels assigned to different relevance grades. We visualize the distribution of the distribution of similarities for each relevance grade in Figure 5.2.

This experiment also provides insights into the internal consistency of the assigned relevance grades for queries, helping us assess the robustness of our evaluation strategy. As expected, our findings reveal distinct patterns in the similarities between qrels at different relevance levels. We notice that qrels at level 3 display higher degrees of similarity compared to level 2 qrels. This trend may stem from higher relevance levels aligning more closely with the information need expressed in the query. i.e., providing a better answer to the query. As we anticipated, passages associated with lower relevance levels demonstrate lower similarity scores, reflecting their lesser relevance to the information need of the query. This pattern holds consistently across both the DL 2019 and DL 2020 query sets. We also examine the impact of different embeddings on quantifying passage similarity. In Figure 5.2, each column represents results with different embeddings (e.g., Vanilla DistilBERT, fine-tuned DistilBERT, text-embedding-ada), we observe variations in the range of similarity scores. However, the overall pattern remains consistent: relevant passages tend to exhibit greater similarity with other relevant passages, increasing with relevance grade. The extent of differentiation varies among embeddings, with text-embedding-ada being more adept at distinguishing between different relevance levels compared to BERT-based representations. Nevertheless, the experiment results appear robust across different embedding types, enhancing our confidence in the reliability of our overall approach.

5.1.4 Assessing Generated Responses with Relevance Judgments

In this section, we assess the responses generated by various Large Language Models (LLMs) as outlined in Table 5.2. We evaluate them by comparing the similarity of their answers to judged relevant passages. In Figure 5.3, we present the results of measuring the average similarity between each generated answer and the judged relevant passages, categorized by different levels of relevance, employing three distinct embeddings. From this figure, we make several observations: 1) As expected, the average similarity between generated answers and passages decreases with passage relevance level. 2) For the DL 2019 collection, the LLaMA2-13b-chat[393] model

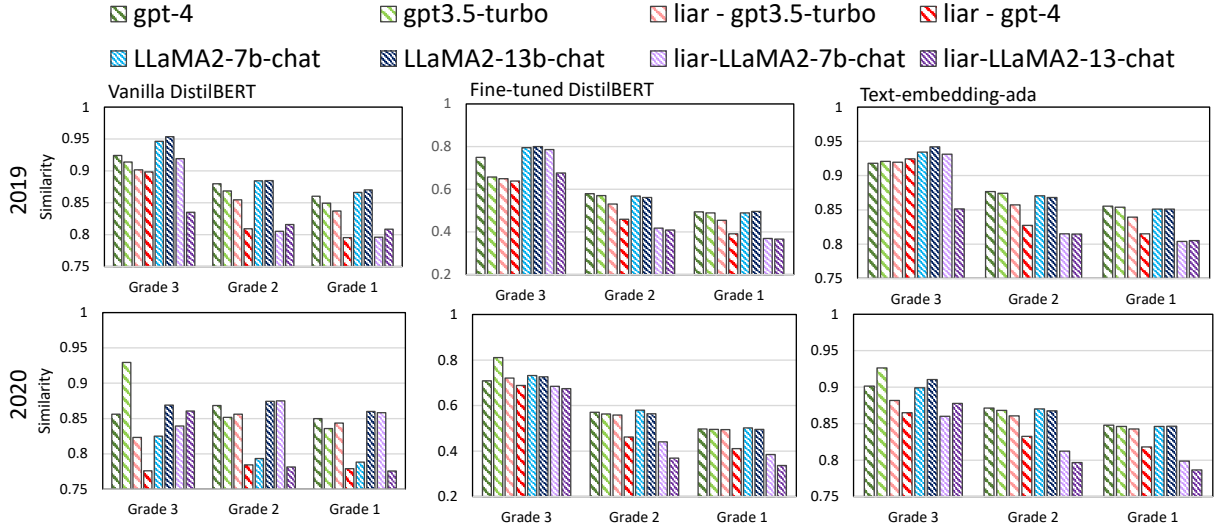


Figure 5.3: The similarity of the responses of the generated models on TREC2019 and 2020 w.r.t the similarity with judged passages in different levels.

demonstrates the highest similarity to the judged relevant passages, followed closely by **gpt-4**. However, for DL 2020, **gpt3.5-turbo** has greatest similarity to the judge relevant passages. 3) When analyzing the “liar” versions of the runs, where we deliberately prompt the LLMs to provide incorrect responses, we observe that **gpt-4** appears to be a better low baseline compared to **gpt-3.5-turbo**, since it consistently yields lower similarity scores to the relevant judged passages. This underscores **gpt-4**’s ability to generate *convincing but incorrect* responses, when appropriately prompted. A consistent pattern appears across all subfigures in Figure 5.3 regardless of the embedding and query set used. This pattern supports our proposal to measure the performance of LLMS by measuring the similarity between generated answers and judged-relevant passages in existing retrieval benchmark collections.

We extend our experiments by evaluating all submitted runs to the TREC DL 2019 and 2020, comparing the retrieved results of those runs to LLM-generated answers. A significant challenge in measuring the performance of both retrieved and generated models lies in the comparison metrics. When examining the similarity between a selected relevant passage (the “target qrel”) and retrieved passages, instances arise where the target qrel is returned at the top rank, giving a similarity score of one. For generated models, achieving a perfect score is only possible if the LLM somehow generates exactly the text of the target qrel, which is possible but unlikely. As a result, direct comparisons might misleadingly indicate that retrieved results have superior performance.

To avoid this problem, we have modified our evaluation strategy by prioritizing top-tier relevance-judged passages for each query i.e., leveraging the annotated documents with the high-

est relevance grade for evaluation of each query. In other words, For each query and for each *target qrel* in the highest grade of relevance level, we identify the first retrieved passage in the results list that is not a direct match to the *target qrel*. We then report the average similarity of every *target qrel* and their accompanies first non-identical retrieved passage. We note that it is possible each retrieved passage is already annotated as a relevant one, but our focus is strictly on non-identical retrieved passages. This approach ensures a more equitable comparison, as we determine a retrieved passage’s similarity to the target qrel in the same manner we do with generated model answers. By employing this method, we can effectively assess the performance of both generated and retrieved models in a uniform context. It is also worth mentioning that within the TREC DL dataset, level 3 denotes perfect relevance. However, there are occasions where top-tier annotations are absent, and in such situations, level 2 emerges as the most relevant tier.

In Figure 5.4, we plot the performance of all runs submitted to TREC, in addition to generative models. In this figure, while the submitted retrieval-based runs (represented by black points) display both their similarity to relevance judgments on the y-axis and ndcg@10 on the x-axis, the generative models (depicted in colored circles and triangles) are only evaluated based on their similarity to relevance judgments, as indicated by the gray boxes, since ndcg@10 can not be computed for generated answers. This analysis illuminates the efficacy of generative models, particularly for TREC DL 2019 and TREC DL 2020. As depicted, models like **gpt-4**, **gpt-3.5-turbo**, **LLaMA2-7b-chat**, and **LLaMA2-13b-chat** perform comparably to the best submissions for these TREC evaluations. Depending on the embeddings used, these models may slightly outperform or underperform the best runs, but their performance is generally on par. Similar to the observation made from Figure 5.3, we note that **gpt-4** appears to be a “better liar” than **gpt-3.5-turbo**, with lower scores. In addition, from this Figure, we observe that our strategy for measuring the similarity of retrieved answers and the highest level of relevant judgments shows statistically significant Kendall’s τ correlation with the official metrics on this dataset i.e., ndcg@10. The high correlation between this similarity and traditional evaluation metrics intensifies our assurance of this evaluation methodology.

5.1.5 Assessing Generated Responses without Relevance Judgments

In the previous sections, we assume we have at least one judged relevant passage available. However, human judgments are expensive to obtain. Moreover, recent experiments suggest that LLM generated labels be competitive with, or even exceed the quality of, human labels [131, 388]. In this section, we assume that no human judgments are available. Instead, we directly compare the top-passage returned by a variety of retrieval methods with the generated answers. While we do not directly employ LLM-generated relevance labels in place of human labels, our approach is similar and we plan to employ these labels in future work. Our experiments employ twelve diverse retrieval pipelines, which are detailed in Table 5.1. Initially, we present an analysis of the

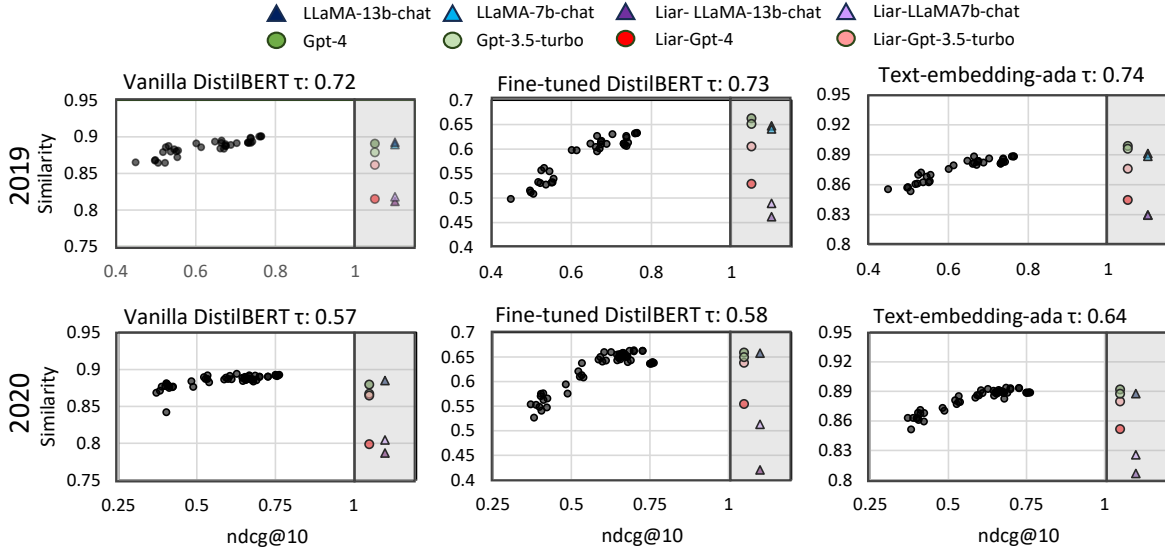


Figure 5.4: The performance of all the runs submitted to TREC DL in 2019 and 2020, as well as the performance of our generated models on these datasets. While the submitted runs are depicted using both ndcg@10 and the similarity score, the gray area shows only the similarity score for the generated runs. The ndcg@10 metric is not applicable to the colored points. Kendal τ correlation between ndcg@10 and the similarity of retrieved results with the grels as explained in section 5.1.4 are mentioned above each sub-figure.

similarity between the generated answers and the top-retrieved passages obtained through each individual retrieval pipeline. Additionally, we provide insights into the average similarity between the generated responses and the best-retrieved passage across all twelve pipelines. We believe that this approach can offer a more robust indication of relevance signals, especially considering that a single query can be answered by multiple passages.

The results of these experiments are presented in Figure 5.5. Regardless of whether one chooses a basic method like BM25 or a more recent neural model like SPLADE, the top retrieved passage consistently emerges as a strong indicator of relevance. Notably, we observe that the relative performance of the generated models remains nearly unchanged when using different retrieval methods as the anchor for evaluation. For instance, when employing fine-tuned DistilBERT for embeddings (as shown in the middle subfigure of Figure 5.5), we consistently observe that **gpt-4** outperforms **gpt-3.5-turbo**, regardless of whether we measure the similarity of the generated content with the top retrieved results from BM25, ColBERT, or SPLADE. This pattern holds true across all the retrieval methods, as well as when considering the average similarity to all the retrieved passages from each of the methods. It is worth noting that **LLaMA2-13b-chat**

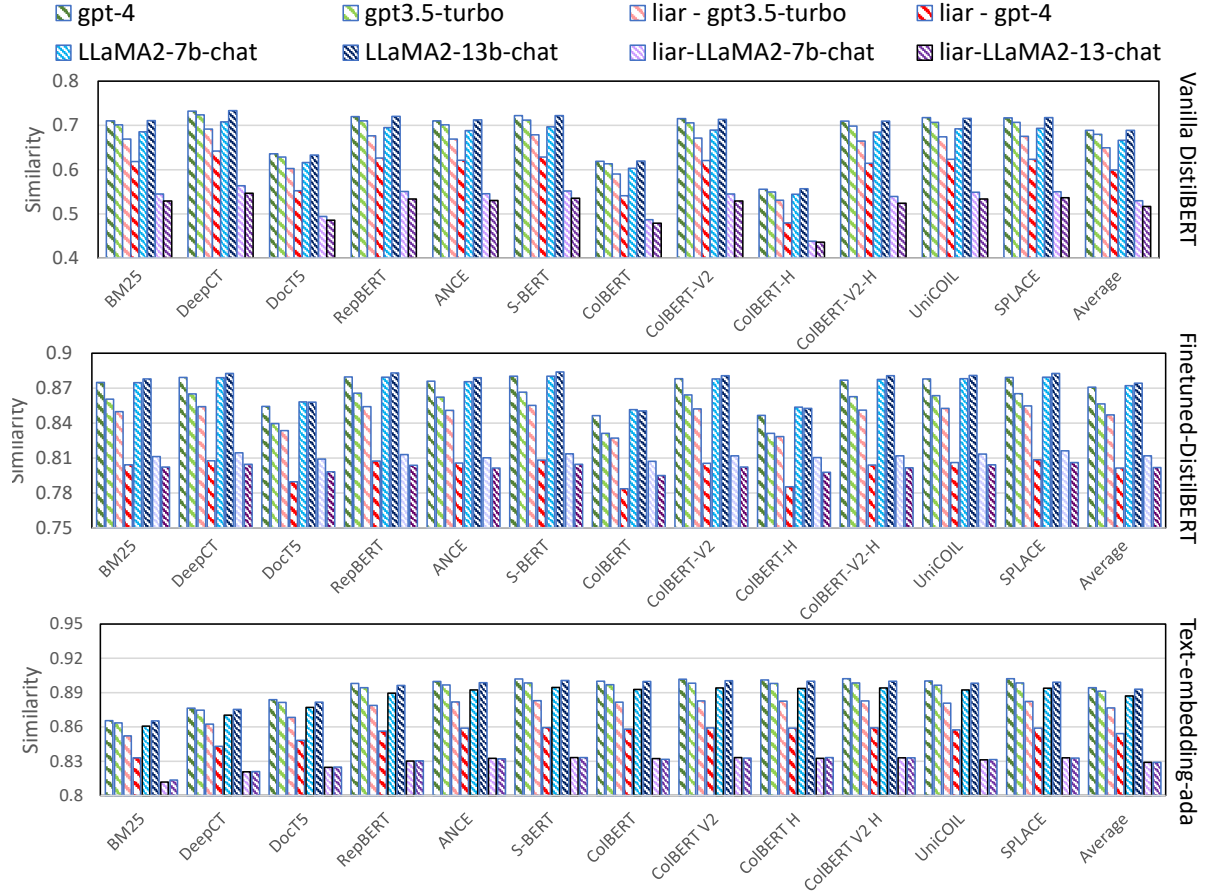


Figure 5.5: The performance of the generated models on MS MARCO dev set based on the similarity with top retrieved documents from different retrieval methods.

consistently outperforms LLaMA2-7b-chat, which aligns with the observations made in previous sections. Additionally, the larger models tend to exhibit lower similarity to the retrieved passages when they are required to generate incorrect answers, indicating their ability to be better at providing deceptive responses. In general, our findings suggest that even without explicit relevance judgments, information retrieval benchmarks serve as a reliable anchor for assessing the quality of generated answers.

5.1.6 Takeaway

This study addresses the challenge of evaluating the quality of answers generated by large language models, specifically for question-answering tasks and in the absence of established evaluation methods. We propose and empirically validate an evaluation framework that leverages existing retrieval benchmarks to assess the generated answers. This framework allows for the comparison of different LLMs and prompts, providing a structured approach to evaluating generative question answering.

Our experiments explore two evaluation strategies: (1) measuring the similarity between generated answers and relevance-judged passages from retrieval benchmarks, and (2) measuring similarity between generated answers and top-ranked retrieved passages in the absence of human judgments. We find that the similarity between generated answers and relevance-judged passages correlates well with traditional IR metrics such as nDCG@10. This correlation suggests that our similarity-based approach can serve as a useful proxy for evaluating LLM outputs, especially in contexts where labeled data is scarce or unavailable. Additionally, comparing generated answers to top retrieved passages appears to provide a reasonable signal of answer quality, indicating that retrieval systems themselves may serve as viable anchors for generative evaluation. In fact, in the absence of any annotated data, a reliable IR pipeline could be leveraged as a reliable probe for assessing LLMs at least for information-seeking tasks such as question answering. Our experiments reveal that generative models, such as `gpt-4` and `gpt-3.5-turbo`, perform comparably to the best retrieval-based runs in TREC DL 2019 and TREC DL 2020 when measured by similarity to relevance-judged documents.

Overall, while our approach is able to distinguish between some generative models, the separation is not always sharp—especially between models of substantially different capacities. T

5.2 Offline Set-based Evaluation of Text-to-image Generation

In this section, we extend the previously borrowed IR evaluation strategies beyond text, exploring how they might inform the assessment of text-to-image (TTI) generation models. Specifically, our aim is to propose set-based evaluation metrics tailored to these models. In the subsequent section, we broaden our scope further, outlining a framework that can be used to assess any LLM-powered applications.

Generative models and specifically text-to-image models are being increasingly used in creative tasks, especially through ideation, the process of generating a list of design candidates in the beginning of a creative design task [211, 195, 442, 121, 183]. In this study, we study TTI systems, a class of models particularly well-suited for ideation in the visual arts [210], architecture [300], and crafting [401]. In order to robustly benchmark TTI systems, researchers need to develop both datasets (i.e., prompts and relevance information such as a target image) [313, 452, 339] as well as *evaluation metrics* [141, 298]. These metrics can either be based on human rater preferences between a pair of system outputs (e.g., through a side-by-side comparison for a fixed prompt) or on offline metrics based on the automatic comparison of model output to an example system target.

While there have been advances in the critique and mitigation of datasets involved in benchmarking TTI systems [334, 366], evaluation metrics remain relatively immature and deserving of more attention [298]. Evaluation methods based on manual side-by-side comparisons of system output [19, 20, 9, 23], although effective at detecting differences between systems, suffer from multiple inefficiencies. On the one hand, this method incurs time costs arising from recruiting and training participants as well as conducting the assessment alone; when operating in rapidly advancing environments, these time costs can limit the space of models evaluated (and developed). On the other hand, because each pair of systems needs to be manually compared, the evaluation scales quadratically in the number of systems [123]; as a result, the financial cost of model evaluation may be prohibitive. Offline or automated evaluation metrics address these scalability issues by using information such as target image data to algorithmically score and compare systems [171, 345].

The most popular approach is to use the Fréchet Inception Distance (FID) [171] (as explained in section 3.2.1), a method that compares the distribution of generated images over a workload of prompts to the distribution of example or target images over the same workload of prompts. FID, while mathematically convenient, is deficient in several ways [92, 52]. First, FID compares population-level statistics and, therefore, a system can generate a set of images comparable to the target set while individually being poor responses to the prompt. Second, because FID is a population-level metric, the benefits of per-prompt error analysis disappear. This includes evaluating for sub-population metrics such as fairness. Moreover, when comparing

a pair of systems, because FID measures population level differences, evaluation cannot leverage matched statistical tests, known to improve sensitivity. Finally, FID may inaccurately measure TTI effectiveness because it is computed using a collection of generated images within and across prompts without paying attention to how they may be used or interacted with by users. This risks metric divergence [283], when there is a disconnection between an evaluation metric and the construct we are interested in measuring, a more general version of the value alignment problem [425]. FID does not explicitly model user behavior or task, even though we know users engage with TTI systems in the context of ideation through a grid interface. As a result, evaluation may be brittle, especially as systems become effective enough to be operating in a region of the performance landscape where FID cannot effectively detect differences (cf. similar phenomenon in search evaluation [409]).

Seeing the need for reusable offline evaluation as a necessary benchmarking tool, we develop novel TTI evaluation metrics based on explicit models of how users engage with system decisions. This responds to a recent call to abandon metrics like FID that do not model human behavior and perception [298]. We leverage theories from the ideation literature to understand what people need during ideation [152, 359, 199, 297] and techniques from IR to formally model these concepts [77]. In particular, we adapt two well-studied IR metrics, expected reciprocal rank [81] and rank-biased precision [278], to account for grid layouts, generated images, and ideation intent. We note that to leverage the soft relevance labels based on similarities to example images, in this study, we assume that there is always one or more ground truth example images known to be relevant to the given prompt.

The goal of this section is to evaluate text-to-image generation systems through the lens of information retrieval-inspired metrics. We aim to develop metrics that capture key aspects of ideation effectiveness, including fluency, quality, novelty, and diversity, while modeling user behavior realistically. Specifically, we explore the following research questions:

- How effective are traditional IR metrics (e.g., RBP, ERR) adapted for TTI evaluation, particularly in aligning with human judgments?
- To what extent does incorporating user modeling and diversity into evaluation metrics improve their agreement with human preferences?
- How robust are the proposed TTI evaluation metrics across different datasets with varying prompt complexity?

We conduct experiments on comparing our proposed set of metrics with traditional ones, demonstrating that

- modeling sequential user browsing increases consistency with human preferences,

- modeling novelty and variety increases consistency with human preferences, and
- modeling image visual saliency increases consistency with human preferences.

We observe this behavior on three different datasets including MS-COCO captions [86], the localized narratives [315] and a set of aggregated prompts generated by humans. We release our code at <https://github.com/Narabzad/Set-Based-Text-to-ImageGeneration>.

5.2.1 Text to Image Generation Evaluation

Despite the recent advances in text guided image generation models [327, 321, 338, 452, 339, 326, 127, 143, 212, 231, 225, 416], evaluating such systems have not been studied extensively. The lack of evaluation tools for such systems has been identified as a serious issue within the community [308, 298, 91, 252, 298]. The Inception Score (IS) measures properties of *unconditional* generated images for 1) low-entropy across image classes, and 2) high entropy across image classes over a population of generated images [345]. However, the Inception Score is susceptible to gaming a perfect score by creating one image for each of the base embedding classes. Therefore, while better performance is possible, the metric does not fully capture the potential for improvement. Additionally, IS is appropriate for assessing only certain properties of a general, unconditional image generation model and is not suitable for TTI evaluation.

As a mathematical refinement of IS, the Fréchet Inception Distance compares the distribution of Inception embeddings of real and generated examples [171, 92]. A lower FID indicates a higher similarity between the real and generated distributions, and thus more realistic images. However, like IS, FID was designed for unconditional image generation, making its use for TTI evaluation somewhat contrived [20]. Although other methods, like Kernel Inception Distance (KID) [49], have proposed further mathematical improvements, FID remains the dominant evaluation metric [298]. Multiple studies have demonstrated that population-level metrics like FID are not correlated with human preferences [243, 306]. In particular, Otani et al. [298] rigorously analyzed TTI evaluation, finding weak alignment with human behavior and perception as a primary cause of this inconsistency.

A second group of metrics evaluates the alignment between generated images and the user prompt. Some metrics achieve this by algorithmically captioning the generated image and then computing text similarity to the user prompt [420]. Other metrics measure semantic object accuracy, which assesses how well objects in the generated image match terms in the caption [172, 448, 177]. This group of evaluation metrics is limited by the performance of a second model (e.g., caption generation, object detection model, or visual question answering models) and usually does not show a high correlation with human-generated prompts, as reported in previous works [339, 452, 305]. In addition, they lack the flexibility to be applied to *set-based* evaluation of TTI systems, which is the focus of our work. Even state-of-the-art caption generation models

struggle to produce distinct captions for images generated from the same TTI system for a given prompt [376, 193, 145]. Our experiments demonstrate that even with a SOTA pretrained image captioner model such as VL-T5[90], there are significant overlaps between the generated captions for images generated from an individual TTI system for a single prompts. Consequently, the evaluation metric would yield the same relevance score for all generated images within the set. As a result, these metrics are not effective in evaluating TTI systems on set-based instances, which is the focus of our work.

In a recent study [439], the authors introduced a method to score and evaluate human preferences. Similarly, in [202] the authors collected a preference-based dataset to train a CLIP-based scoring function so called as PickScore to predict human preferences on individual pair of generated images for a given prompt. However, their focus was on individual generated images rather than considering set-level evaluation, which crucially incorporates diversity alongside relevance. In our work, we only focus on TTI evaluation metrics that are applicable to set-based TTI generation evaluation.

There have been also attempts to develop evaluation pipelines for TTI models [38, 179, 202, 177]. For instance, Petsiuk et al. [313] created a set of fifty challenging tasks and applications for state-of-the-art TTI models. The authors classified the challenging tasks into those that require human evaluation procedures and those that can be automatically evaluated. They argued that these situations are difficult to define, such as when common sense is required or when objects that do not often co-occur in the real world must be combined. As a result, several curated benchmarks for TTI evaluation have been developed. Moreover Yu et al. [452] introduced PartiPrompts, a comprehensive benchmark of 1,600 English prompts covering 12 different categories and 11 challenge aspects. In PartiPrompts, similar to Xu et al. [439] human annotators were asked to select the better generated image from two models (one image per model). Although the majority of preference-based TTI evaluation focuses on pairs of single images, Saharia et al. [339] conducted preference-based evaluations between sets of images generated by two models. To the best of our knowledge, their collected dataset is the only work that conducted evaluation on set of images. In our work, we follow the same concept and aim to conduct experiments comparing sets of generated images, albeit on a larger scale.

Finally, some work also proposed strategies to bridge the gap between expert annotation and automated metrics in generation tasks. Laban et al. [216] proposed a methodology to extend the use of existing annotations into a series of automatic tests. They argue that it is necessary to do so since human evaluation is usually expensive, hard to reproduce, and not easy to reuse. They also show that their proposed approach called Near-Negative Distinction achieves a higher correlation with human judgments than standard natural language generation evaluation metrics.

Diversity Measuring the diversity of image sets has a long history in multimedia retrieval [400, 258, 5]. For example, the MediaEval 2017 benchmark introduced image search result diversification in the context of photo retrieval [453], arguing that diversity within image search results

is crucial because it 1) addresses the needs of different users, 2) tackles requests with unclear information needs, 3) widens the pool of possible results for increasing the performance, and 4) reduces redundancy in the results. We adopt similar desiderata motivated from the precise ideation context in which TTI is often used. The best-performing models in this benchmark mostly leveraged visual similarities between images and Maximum Marginal Relevance (MMR) [153]. Inspired by the successful runs in this competition, we also leverage the idea of MMR for measuring diversity in our evaluation metric. We note that while MediaEval could use relevance and diversity unmodified IR metrics because images were explicitly rated, TTI needs to impute these ratings because images are generated.

5.2.2 Problem Definition

We cast text-to-image generation (TTI) as the task of supporting a user who is searching for visual ideas in an early-stage *ideation* context. Treating the activity as an information-seeking task lets us borrow the formal machinery of information-retrieval evaluation.

TTI task. Modern TTI interfaces (e.g., Midjourney, StableDiffusion Web, DALL·E 3) present their outputs in exactly this two-dimensional format: designers skim across rows and columns, making comparative judgements at a glance. Given a prompt or query q , a TTI system s , generates an $w \times h$ matrix $\mathcal{X}$ where each element is an image. If the TTI interface returns a ranked list of length k , we generate a $1 \times k$ matrix and similarly when a grid is required, the TTI system presents the generated images in a grid view of $m \times n$ images.

Why a grid matters. Modelling the output as a matrix lets us both simulate realistic scan trajectories—by sampling browsing paths π through the grid and computing the expected user utility across all plausible visual scans—and reward simultaneous quality and diversity, since early-stage ideation favours a “mood-board” of diverse, high-quality images rather than a single best item; treating $\mathcal{X}$ as a spatially structured set enables the metric to penalise redundancy while still crediting fidelity to the prompt.

TTI evaluation task. An evaluation metric μ is a function that, given an arrangement of generated images $\mathcal{X}$, a prompt q , and side information u_q about the image utility (e.g. an example target image), computes a scalar value where a higher value indicates better performance of system. We summarize the performance of system s over a space of prompts Q with the expected value of μ over the distribution of queries in Q , $\mathbb{E}_{q \sim Q}[\mu(s(q), u_q)]$.

5.2.3 Method

Designing a Metric for TTI

In the context of creative design, users leverage TTI systems to generate design candidates to support ideation. Dominant theories for measuring the quality of an ideation process focus on four criteria [152, 359, 199]. *Fluency* refers to the total number of relevant items generated. *Variety* refers to the number of unique types of relevant items generated. *Novelty* refers to how different relevant items are from all previously generated items. Finally, *quality* refers to the degree of relevance of generated items. We are interested in designing metrics that capture these different dimensions of ideation effectiveness. Here, we will review relevant concepts from information retrieval evaluation as they pertain to TTI evaluation.

A General Model of User Behavior

Let $\mathcal{X}$ be the matrix of $k = wh$ images generated for a specific prompt, where we index each generated image from 1 to k . When inspecting a grid, we define the trajectory π as be the specific sequence of inspected images represented as a permutation of $[1, k]$, where π_i refers to the i th examined image index. We begin our metric development in the context of classic IR evaluation where users inspect a one-dimensional ranked list in a serial, deterministic order. We will return to grid-based interfaces at the end of this section.

We also have available information about the *relevance* of an image to the user for a prompt q . Let $f^*(x) \in [0, 1]$ represent the relevance of an image x to the user.

Given a trajectory π , we define a family of metrics based on how the user might engage with the images in linear order. The *position-based model* models the probability of a user inspecting the image at rank position i in the trajectory as γ^{i-1} , where $\gamma \in [0, 1]$ is a free parameter controlling the depth the user is likely to reach [110, 333]. The *cascade model* models a user who, after inspecting a relevant image, might be satisfied and terminate their scan [80, 110, 81]. Let the probability that a user is satisfied by an image be a monotonically increasing function of its relevance, $g^*(f^*(x))$, which we represent as shorthand using $g^*(x)$ for clarity. In this case, the probability that they reach rank position i is $\gamma^{j-1} \prod_{j=1}^{i-1} (1 - g^*(\pi_j))$. In the remainder of this section, we will demonstrate how to use these models to design metrics capturing fluency, quality, novelty, and variety.

Fluency

To measure fluency, we define a metric based on the expected number of relevant images the user will see. For the position-based model, rank-biased precision or RBP [278], is defined as,

$$\text{RBP}(\pi) = \sum_{i=1}^k f^*(\pi_i) \gamma^{i-1} \quad (5.1)$$

Under the cascade model, extended expected reciprocal rank (ERR) metric [81], is defined as,

$$\text{ERR}(\pi) = \sum_{i=1}^k f^*(\pi_i) \gamma^{i-1} \prod_{j=1}^{i-1} (1 - g^*(\pi_j)) \quad (5.2)$$

Quality

Unlike classic IR evaluation, we cannot *a priori* judge the relevance of all possible images, we leverage a wide variety of side information to estimate the relevance of generated images. In general, this could include information about user preferences, image attributes, or anything else helpful to estimate relevance. In our study, we assume that we have access to one or more example images known to be relevant to the prompt. Given a generated image x and a relevant image $\hat{x}$, we estimate the relevance of x as the following where ϕ is an image embedding function, for example based on the activation of an interior layer of a neural network:

$$f^*(x) = \langle \phi(\hat{x}), \phi(x) \rangle \quad (5.3)$$

Carterette and Allan [66] uses a similar approach to label unlabeled documents in classic IR evaluation.

Variety and Novelty

In TTI evaluation, for a given prompt, we only have a few example images and therefore do not have information to measure novelty of a generated image to the user or in general. However, we can measure the novelty of an image to the user *while they scan the list*. Given the similarity of novelty within a list to variety within a list, we consider both of these criteria together. In situations where we have additional information about the novelty of an image to a user or in general, we suggest these criteria be decoupled. In order to measure novelty within a list, we leverage existing methods designed to measure diversity. Similar to Maximum Marginal Relevance [153], we model novelty by discounting the relevance of an image based on previously seen images.

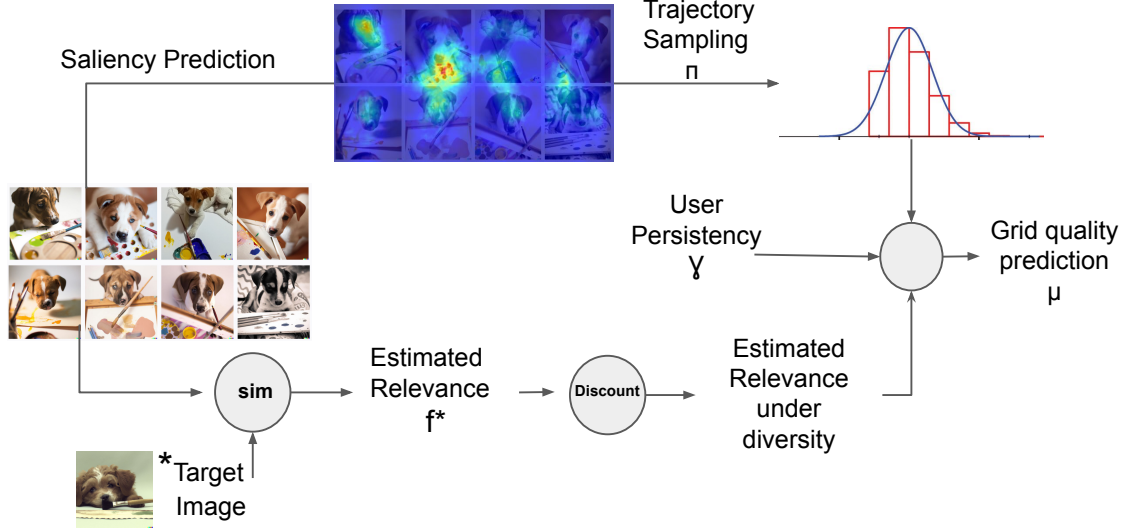


Figure 5.6: Overview of our proposed approach

The discount factor for the i th image in π is defined as the following where $\eta(i, \pi)$ can be used to discount $f^*(\pi_i)$ which we can then use in either RBP or ERR.

$$\eta(i, \pi) = 1 - \max_{j \in [1, i-1]} \langle \phi(\pi_i), \phi(\pi_j) \rangle \quad (5.4)$$

Expected Metrics over Trajectories

In more general grid-based interfaces, we cannot assume that users will all follow a deterministic trajectory when presented with an arrangement of images [124]. Users may inspect images in arbitrary trajectories based on their position and attractiveness. As such, we can compute the expected metric value over all possible trajectories. Let $\Pr(\pi)$ be the probability that the user scans the images in the order represented by π ; in other words, $\Pr(\pi)$ is a multinomial distribution over all permutations in S_k . Although a naïve model might consider a uniform distribution over trajectories, we know that users tend to be more attracted to certain images based on their position and visual features [124]. Inspired by the observation that users tend to look at the most salient images first [285, 365], we model $\Pr(\pi)$ using a Plackett-Luce model based on predicted image salience [314]. Let $\sigma(x)$ be the visual salience of x represented as a positive scalar value. After normalizing this value over all images to sum to one, we iteratively sample an image from this multinomial *without replacement*, generating a sequence of images (i.e. π). This process can be executed efficiently by using the Gumbel-softmax method [56]. Figure 5.6 displays a general overview of our proposed approach for evaluating TTI systems.

Table 5.4: Examples of prompts from MS-COCO ,LN-COCO and Prompts dataset as well as average and standard deviation σ of length of prompts in each dataset.

Dataset #	#Words	σ	Example Prompt
MS-COCO500	10.3	2.2	A herd of cows standing on a grass covered hillside.
LN-COCO500	42.8	19.2	In this picture we can see three cows standing on the grass. There is a tree and few mountains are visible in the background.
Prompts500	11.6	9.4	Origami cow flying over the moon.

5.2.4 Experimental Setup

TTI Systems

To run experiments and validate how our proposed evaluation metrics can distinguish between the two sets of generated images, we took three state-of-the-art TTI systems [339, 452].

To comply with intellectual-property agreements signed with our industrial partner, we may not reveal the commercial identities of the underlying models. We therefore anonymise them as systems A, B and B'. B' is the smaller version of system B, with fewer number of parameters. In prior work, systems A and B have shown competitive performance and consequently, the smaller version of system B, system B', has shown relatively worse performance compared to system A. The assumption that a model with fewer parameters has worse performance compared to a larger model of the same class has previously been used for evaluation purposes [308].

Datasets

To conduct our analysis, we require sets of prompts with associated prior information. As discussed in Section 5.2.3, each prompt is linked to at least one relevant target image. We utilized the COCO dataset [240], which provides two distinct representations of information needs: 1) COCO captions (MS-COCO) [86] and 2) Localized Narratives (LN-COCO) [315]. Additionally, we collected a third dataset, Prompts, comprising of prompts entered by users of a TTI demo.

The examples in Table 5.4 highlight the difference between prompts used for these experiments for the same image. As shown in the first two rows of this Table, the same image was described with 10 words when using MS-COCO captions and 24 words when using LN-COCO as prompts. We demonstrate how a same image was described with 10 words when using MS-COCO captions and 24 words when using LN-COCO as prompts. We are interested in investigating whether our proposed metrics and baselines are robust with respect to the lengths of the prompts for the same prior information. While the average number of words in Prompts and MS-COCO seem close (11.6

vs. 10.3), the standard deviation of the number of words in the prompts of these two datasets differs significantly. The standard deviation of the number of terms in **Prompts** is 9.41, while this number for MS-COCO is only 2.21. From Table 5.4, we can see that prompts in the MS-COCO dataset are consistently short, while in the **Prompts** dataset, the length of the prompts varies more. We conclude that prompts in the MS-COCO dataset are consistently short, while in the **Prompts** dataset, the length of the prompts varies more.

COCO Captions The COCO dataset is widely used to evaluate various deep learning and computer vision tasks, including TTI systems [452, 339, 127, 361, 85, 468, 240]. It is a large-scale dataset for object detection, segmentation, and captioning [240]. The MS-COCO captions dataset contains over 200K images, each annotated with five captions. Here, we randomly sampled 500 images from the MS COCO 2017 validation set. For each image, we randomly selected one of the captions as the input prompt for experiments and used the associated image as the target image.

Localized Narratives Recent work has raised concerns that COCO captions might not be realistic surrogates for human prompts and that they lack sufficient detail for TTI task evaluation [452]. Therefore, to evaluate TTI systems, we move beyond the simple COCO captions benchmark and use more detailed text found in the Localized Narratives (LN) dataset [315]. Localized Narratives dataset which are a subset of the MS-COCO dataset, and their associated text is four times longer than COCO captions on average [315]. Inspired by previous work [212, 468, 457], we also conducted experiments with generated images from more detailed descriptions of images (localized narrations) and examined how the longer version of prompts affects the performance of TTI systems.

Prompts Describing an image through captions or detailed narrations is different from issuing a prompt to address an information need in various contexts. To ensure that our evaluation metric is representative of prompts that appear in practice, we also analyzed performance on human-generated prompts. We collected a third dataset including 1,500 prompts from real users, referred to as **Prompts**, consisting of prompts entered by users of a TTI demo. This dataset is more representative of the types of prompts that TTI systems will encounter in practice, and they cover a wide range of topics and styles. Some examples of prompts in **Prompts** are "*A futuristic city with lots of greenery*", "*A small cabin in the woods*", and "*A cozy reading nook in a library*". The TTI demo was used for generating images for various purposes, including research, design, and presentations. We also recorded the generated grids from two TTI systems, A and B, and collected feedback on individual images, including positive and negative signals, such as thumbs up and down, as well as implicit signals, such as downloads, copies, or shares. We used both implicit and explicit positive feedback to create a set of target images for each prompt, selecting only those prompts that were tried on both TTI systems and received positive feedback on both. This resulted in a set of target images, consisting of at least one image from each of the

Table 5.5: Different variations of our proposed metrics

Metric	User Model	Relevance	Pr(π)	Metric	User Model	Relevance	Pr(π)
RBP	Position-based	$f^*(x)$	Saliency	ERR	Cascade-based	$f^*(x)$	Saliency
RBP_η^u	Position-based	$f^*(x) \times \eta(i, \pi)$	Uniform	ERR_η^u	Cascade-based	$f^*(x) \times \eta(i, \pi)$	Uniform
RBP_η	Position-based	$f^*(x) \times \eta(i, \pi)$	Saliency	ERR_η	Cascade-based	$f^*(x) \times \eta(i, \pi)$	Saliency

two TTI systems, for each prompt. We randomly sampled 500 such prompts that satisfied these requirements.

Metric Variations

We present our six different metric variations in Table 5.5. RBP and ERR metrics are defined in Equation 5.1 and 5.2 respectively. The subscripts η discounts the relevance function as explained in Equation 5.4. When π is sampled based on uniform distribution the metric has a superscript u otherwise the π is sampled using the normalized image saliency and unless specified, we adopt a value of γ equal to 0.9, as suggested in the literature [65]. We compute the average of the metric over 100 sample trajectories for each prompt. The similarity between generated images and target images is calculated using the cosine similarity of the embedded vectors of images obtained from Inception V3 which has been widely used in downstream vision tasks and evaluation methodologies [381, 327, 433, 410, 233, 43, 295].

Baseline TTI Evaluation Metrics

In addition to FID, we develop a simple diversity-based metric, *Diversity*, to assess the set-based diversity of TTI output. *Diversity* measures the average pairwise similarity between all generated images. A higher pairwise similarity rate within a grid indicates *lower* diversity. We compute pairwise similarity using cosine similarity of the embedded vectors of images obtained from Inception V3.

Modeling Saliency

Although there are no existing saliency models or gaze data available for a grid of images, we can use models trained on webpages, which usually contain multiple images (or even a grid of images sometimes). There are several existing free viewing gaze data on webpages [74, 362], and among them, [74] is the latest and largest, with gaze data on 450 webpage screenshots, collected from 41 people with eye trackers. 450 webpages are a relatively small data set, and insufficient to train the webpage saliency model. We follow the typical training paradigm in saliency modeling area: first train the model with Salicon data [184], a large scale saliency data set on 10K natural

images, then fine-tune it with webpage gaze data. The loss functions we use is a combination of KLD and NSS [62], both of which are popular loss function for saliency models.

Our webpage saliency architecture essentially follows the SimpleNet idea in [329] with some modification and simplification. More specifically it takes a MobileNet V3 (pretrained on ImageNet classification data) as backbone, and extracts embeddings from 4 layers of conv 2, 4, 6, 8. On each embedding, we apply two conv layers, where the first layer has a kernel 3×3 , number of channels matching the input embedding, max pooling 3 and relu operator and the second layer has a kernel 1×1 , 1 channel and relu operator. Then the output of each branch is converted back to the resolution of the input with a bilinear resize operator. The outputs of the 4 branches are then summed to 1 channel, followed up by a sigmoid function to generate the output saliency map. Finally, we average the predicted saliency of individual images within the grid and use it as initial distribution to sample π^5 .

Figure 5.7 represents 4 grids of generated images for the prompt ‘a cute puppy is painting’ with different levels of saliency and diversity.

Data Annotation Process

We curated a dataset of human preferences over sets of generated images for prompts from the three above-mentioned datasets. An annotation study was conducted using an internal platform, where annotators selected their preferred set of images for each prompt. Based on prior work in task-based evaluation of information retrieval systems [53], we grounded the annotation in a simulated work task provided to annotators as,

“ You and your coworkers are trying to come up with an image for a project presentation. Together, you all have come up with a description of the image, which we will refer to as a ‘prompt’ (shown in blue below). Two designers, X and Y, have sketched possible images for the group to decide on the images to use. Which designer’s sketches would you prefer to present to the group to decide on the presentation image?”

The instructions, along with a prompt, were displayed at the top of the assessment interface. Two sets of eight images were presented in a 2×4 grid view, one from TTI system X on the left and one from another TTI system Y on the right, immediately below the guidelines and the prompt. Annotators rated which of the grids they preferred on a five-point Likert scale:

1. X is much better than Y,
2. X is somewhat better than Y,
3. X and Y are the same,

⁵We release our code at <https://github.com/Narabzad/Set-Based-Text-to-ImageGeneration>

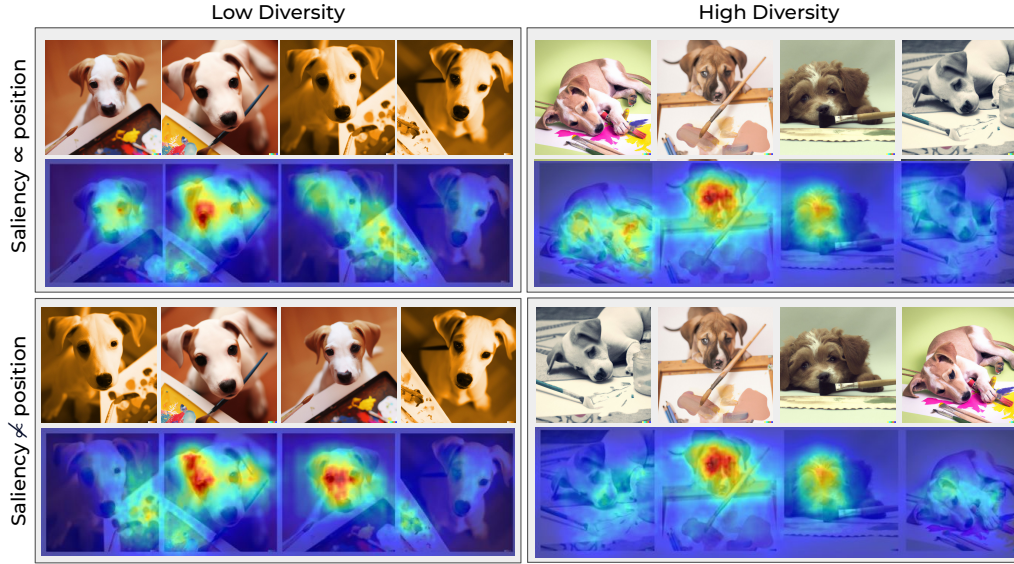


Figure 5.7: Four image grids and associated saliency maps for the prompt ‘a cute puppy is painting’. Images grids in the left column have lower diversity than those in the right column as reflected by the different positions and breeds on puppies. In image grids in the top row, images tend decrease in saliency from left to right; in the bottom row, there is less of a relationship between position and saliency.

4. Y is somewhat better than X, and
5. Y is much better than X.

To ensure the quality of the collected annotations, each pair of sets of generated images was annotated three times with different annotators. We used the consensus agreement among all three judgments to generate the final annotation results. We ran preference-based annotation between each pair of systems, i.e., A vs. B, A vs. B’, and B vs. B’, on 500 prompts from each of the MS-COCO, LN-COCO, and Prompts datasets.

In total, we collected 13,500 annotations for our dataset. Additionally, we conducted 100 quality check tests in which we presented a generated set of images for a given prompt and a random set of generated images from a different prompt. We assumed that the randomly generated set of images would always be less relevant. We are pleased to report that all of the quality tests were passed with the consensus results. Furthermore, the crowd sourcing platform we utilized had their own internal quality checks and qualification tests which were passed at the most satisfactory level. On average, each annotation took 106 seconds to complete and although we paid the raters hourly, on average we paid 0.50 USD per annotation.

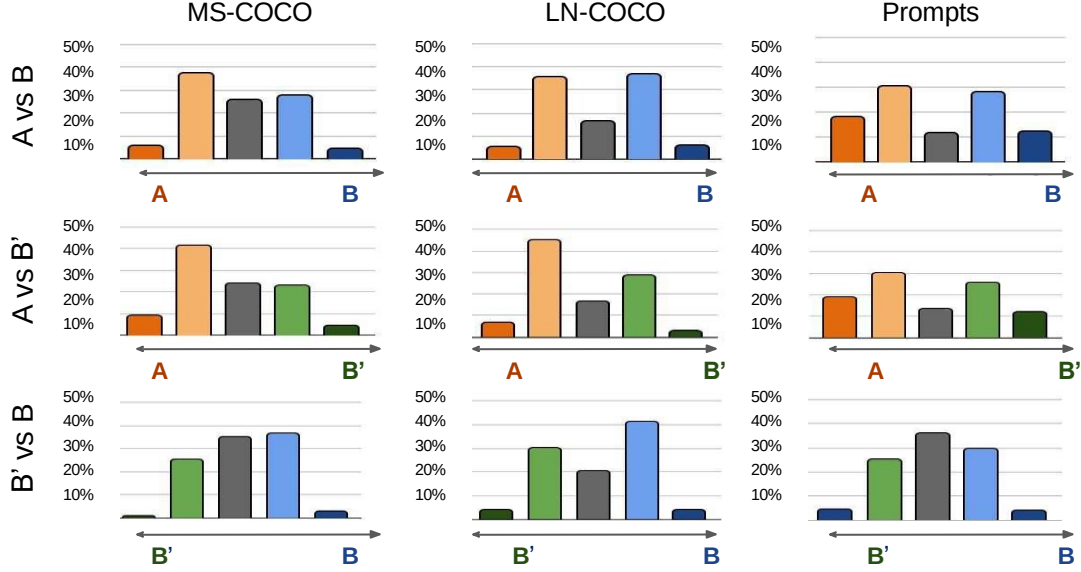


Figure 5.8: Annotation results of TTI systems A (orange), B (blue) and B' (green) on the three datasets. Given the TTI systems X/Y are on the left/right side of the arrows beneath a sub-figure, the bars in the sub-figure present if "X is much better than Y", "X is somewhat better than Y", "X and Y are the same", "Y is better than X" and "Y is much better than X" consecutively.

Agreement Rate

Three different annotators judged each pair of sets of generated images. We computed Fleiss' κ independently for each dataset. In Table 5.6, we provide the agreement rate between the annotators when annotating different pairs of systems on the three datasets separately, as well as the total agreement on all the annotated pairs of grids. We also report the agreement rate on a three-point scale where we collapse ratings in the same *direction* (e.g., considering 'A is much better than B' and 'A is somewhat better than B' as one option 'A is better than B'). As shown in this Table, the agreement rate between the three annotators was quite promising, and the majority of the prompts received at least two out of the three annotator's agreement. Even on a 5-scale consensus result, which is harder to achieve higher agreement on, only 13.6% of the prompts from MS-COCO, 12.3% of prompts from LN-COCO, and 12.6% of prompts from Prompts dataset did not have any consensus agreement between the three annotators. We qualitatively observed some of the prompts that received 0 agreement rate between the three annotators and noted that in these special cases, both of the models generated a relevant set of images, and thus the annotator's selection became less robust.

When looking at agreement across annotators for grades on our Likert scale, we observe moderate agreement **MS-COCO** ($\kappa = 0.57$) and **LN-COCO** ($\kappa = 0.58$) and fair agreement for **Prompts** ($\kappa = 0.33$). This suggests that systems are easier to precisely distinguish in **MS-COCO** and **LN-COCO** while the **Prompts** dataset is more challenging for annotators. We also computed agreement in *direction*, collapsing the bottom two and top two grades. When computing the directional agreement, we observe substantial agreement for **LN-COCO** ($\kappa = 0.62$) and moderate agreement for **MS-COCO** ($\kappa = 0.56$) and **Prompts** ($\kappa = 0.41$). We also observed that, despite perhaps being more challenging when evaluating these specific systems, the **Prompts** dataset has equal or better majority agreement compared to **MS-COCO** and **LN-COCO**. From an evaluation perspective, the difficulty in distinguishing systems could arise from, for example, harder to distinguish pairs of image sets or more ambiguous prompts. If we had observed substantially lower agreement, we would question whether these systems were in fact different in terms of the prompts in the **Prompts** dataset.

Additionally, by comparing the agreement rate on the three datasets, we observe the agreement rate on the **Prompts** dataset is significantly lower than the other two datasets. For instance, while 57.8% of the prompts in the **MS-COCO** dataset and 61.8% of prompts in the **LN-COCO** dataset had 3 out of 3 consensus agreement on a 5-scale annotation result, only 28.5% of the prompts in the **Prompts** dataset had full agreement rate between the 3 annotators. Obviously, the agreement rate increases when we only consider the direction of the options and neglect the magnitude. For instance, the 3/3 agreement rate increases from 57.8% to 63.5% for **MS-COCO**, 61.8% to 68.6% for **LN-COCO**, and from 28.5% to 45.8% in the **Prompts** dataset. However, we still see a notable difference between the three datasets in the agreement rate. We hypothesize that this difference could be because the prompts in the **Prompts** dataset are more challenging and point to more specific concepts and design patterns, such as the vibes, the quality, style, etc. However, the prompts in the **LN-COCO** and **MS-COCO** datasets do not elaborate on such specifications, and thus they obtained a higher agreement rate between the three annotators.

5.2.5 Results and Validation

Annotation Results

We present the results of our annotation experiments on the five-point Likert scale in Figure 5.8 for the three datasets. These results demonstrate that the preference rate between each pair of systems is highly dependent on the dataset (i.e., the composition of input prompts). We note that the **MS-COCO** prompts are the shortest and simplest ones for TTI systems to address. In contrast, the **Prompts** dataset and the **LN-COCO** prompts are longer and more challenging, as shown in previous work [452, 339]. Observing the first row of Figure 5.8 (A vs B), we note that when prompts are simpler (i.e., **MS-COCO**), both systems A and B perform well. In fact, in 26% of the comparisons, annotators rated both systems as “A and B are the same.” However, for

Table 5.6: Agreement rate between the three annotators on each of the datasets and between each pairs of TTI systems on a 5-level and 3-level Likert scale. Fleiss’ κ is reported on both 3-levels and 5-levels consensus.

		5-levels Consensus				3-levels Consensus			
	Dataset	0/3	2/3	3/3	κ	0/3	2/3	3/3	κ
A	MS-COCO	15.6%	24.5%	59.9%	-	13.7%	21.7%	64.6%	-
vs	LN-COCO	10.6%	26.9%	62.6%	-	9.3%	18.9%	71.8%	-
B	Prompts	11.8%	60.7%	27.5%	-	4.0%	44.3%	51.7%	-
A	MS-COCO	12.6%	25.7%	61.7%	-	10.8%	23.8%	65.4%	-
vs	LN-COCO	12.7%	25.4%	61.8%	-	10.1%	21.5%	68.4%	-
B'	Prompts	12.2%	59.5%	28.3%	-	4.6%	43.7%	51.7%	-
B	MS-COCO	12.6%	25.7%	61.7%	-	18.2%	21.4%	60.4%	-
vs	LN-COCO	13.7%	25.2%	61.1%	-	12.4%	22.1%	65.5%	-
B'	Prompts	13.6%	56.7%	29.7%	-	8.0%	58.1%	33.9%	-
	MS-COCO	13.6%	25.3%	61.1%	0.57	14.2%	22.3%	63.5%	0.56
All	LN-COCO	12.3%	25.8%	61.8%	0.58	10.6%	20.9%	68.6%	0.62
	Prompts	12.6%	59.0%	28.5%	0.33	5.5%	48.7%	45.8%	0.41
Average Consensus		12.8%	36.7%	45.1%	-	10.1%	30.6%	59.3%	-

more challenging datasets, we observed ties amongst 17% (LN-COCO) and 12% (Prompts) of the comparisons. Consequently, the percentage of extreme choices (i.e., when one system is preferred “much” better than the other) is lowest in MS-COCO, which confirms our previous observation that both systems perform similarly, making it difficult for annotators to distinguish between them. On the other hand, in the Prompts dataset, annotators tended to select extreme choices (i.e., one model is “much better” than the other one) more frequently than the other two datasets. Therefore, when the prompts become more challenging, it is easier for annotators to discriminate between two models, as it is more likely that one of the systems fails to satisfy the complex prompt. When the prompts are simple, we hypothesize that both models perform well, and annotators do not perceive many differences between the two sets of generated images.

When comparing the first row with the second row in Figure 5.8 and keeping in mind that B’ is the version of B with fewer parameters, we observe that system A was preferred more frequently in the A vs B’ comparison rather than in the A vs B comparison across all three datasets. For example, in the MS-COCO dataset, A was rated as much or somewhat better than B in 43% of the prompts, while A was preferred over B’ in more than 50% of the prompts. This increase in the preference rate of A from the comparison to B versus the comparison to B’, indicates that, as expected, the larger model B was more competitive with A than the smaller model B’. It is also interesting that the option “X and Y are the same” received the highest rate when comparing TTI

Table 5.7: Agreement rate of the metrics with human annotations. Statistically significant agreement with Wilcoxon paired test and p-value < 0.05 are shown with * symbol. Bold: highest statistically significant agreement in the column.

	MSCOCO			LNCOCO			prompts		
	A vs B	A vs B'	B vs B'	A vs B	A vs B'	B vs B'	A vs B	A vs B'	B vs B'
Diversity	47.2%	41.9%	53.1%	45.3%	42.1%	45.7%	45.9%	46.3%	47.1%
RBP	53.7%	55.3%*	55.5%	53.1%*	64.1%*	52.1%	56.9%*	60.2%*	54.2%
RBP _{η}	53.8%	61.2%*	51.7%	52.5%*	58.8%*	54.2%	58.5%*	60.7%*	54.2%
RBP _{η^u}	53.3%	58.9%*	52.9%	54.2%*	60.0%*	53.4%	57.7%*	60.2%*	49.5%
ERR	52.2%	55.4%*	51.7%	50.0%*	60.1%*	54.2%	57.7%*	61.5%*	53.5%
ERR _{η}	52.2%	60.4%*	50.9%	53.1%*	60.8%*	56.2%	59.3%*	63.1%*	54.2%
ERR _{η^u}	52.2%	58.4%*	52.9%	55.2%*	62.6%*	52.7%	58.1%*	61.5%*	52.3%

system B with B'. Especially for the **Prompts** dataset, the two systems are almost indistinguishable and were labeled as the same in over 36% of the prompts. In comparison, between systems B vs B', we also note the smallest difference in preference between the two sides, with less than a 4% difference between when B was preferred over B' and when B' was preferred over B.

Agreement with Human preferences

Given two sets of images generated by systems s and s' , we expect $\mu(s(q), u_q) > \mu(s'(q), u_q)$ if human annotators preferred the set of generated images from s over s' . To ensure the reliability of agreement between annotators, we only report the agreement rates of the evaluation metrics on prompts where all three annotators fully agreed on the annotations. The agreement rates in Table 5.7 indicate the percentage of the prompts where annotators rate system X as better than system Y and our metric also computes $\mu(X(q), u_q) > \mu(Y(q), u_q)$. Since FID is not capable of quantifying the quality of a single set of generated images for a given prompt, it is impossible to report the agreement rate between annotators on a per-sample basis. We also report the results of Wilcoxon paired statistical significance tests with p-values ≤ 0.05 between the annotations and the measured performance by μ . As shown in Table 5.7, all six of our proposed metrics are able to show statistically significant agreement rates with annotations when comparing A vs B' on all three datasets. However, they do not show statistically significant agreement with annotators when comparing B vs B'. We hypothesize that since B' is a version of B only varying by the number of parameters, their performance may indeed be quite similar. Therefore, the non-significant agreement rate in this comparison could be the result of actual comparable performance between the two systems. The third row of Figure 5.8 confirms this since the two systems have the highest "B and B' are the same" rate (the gray bars) across all three datasets.

We additionally note that on the two most challenging datasets (LN-COCO and Prompts), we see statistically significant agreement rates with annotators when comparing TTI model A vs B.

We note that **MS-COCO** prompts are the most simple prompts among the three datasets, and the **MS-COCO** dataset have been widely used for training TTI systems. As such, TTI models are more familiar with the structure of simple prompts and images. However, for more complex prompts, it is more likely that one of the models fails to generate a relevant set of images, making it easier to assess their performance. As shown in Table 5.7, our proposed metric ERR_η outperforms the other evaluation metrics on the **Prompts** dataset. We achieved a 59.3% agreement rate with annotators when comparing **A** vs **B** and a 63.1% agreement rate when comparing **A** vs **B'**. This metric also shows favorable agreement rates on the other two datasets. Additionally, the position-based metric RBP_η , which considers variety and novelty, shows the highest agreement on the **MS-COCO** dataset. In general, novelty-based metrics were able to outperform other metrics on all three datasets, confirming the impact of considering variety and novelty in TTI evaluation metrics.

These results indicate that our evaluation pipeline is highly effective at separating dissimilar models on challenging prompts, yet it acts as a comparatively coarse instrument when models are very similar or the prompts are easy. To sharpen fine-grained sensitivity, future work should (i) stratify and report results across clearly defined prompt-difficulty tiers (easy / medium / hard) so that performance variation is explicit, (ii) over-sample harder prompts to magnify subtle quality gaps, and (iii) provide query-level confidence intervals or bootstrap tests to detect small but consistent differences.

Ablation Study

We studied the impact of $\Pr(\pi)$ (Section 5.2.3) on the agreement rates by comparing RBP_η with RBP_η^u and similarly comparing ERR_η with ERR_η^u . We observe in Table 5.7 that ERR_η shows a superior agreement rate compared to ERR_η^u on the majority of datasets when the agreement rate is statistically significant. Additionally, RBP_η also shows a greater agreement rate compared to RBP_η^u on two out of the three datasets. By comparing the impact of $\Pr(\pi)$ across the three datasets, we hypothesize that sampling π from the saliency distribution of the grid works best as the prompt sets become more challenging and it has the most positive effect on the **Prompts** dataset and less positive effect on the easier prompt sets. We have also included the agreement rate of the *Diversity* metric. However, it failed to show any statistically significant agreement with the human annotators. Nonetheless, by comparing the performance of metrics with and without the variety and novelty components, we can conclude that incorporating variety and novelty in TTI evaluation metrics improves the agreement rates of both the position-based and cascade-based metrics with the human annotations [25].

Table 5.8: Distinguishing images from the Preferred (P) and Not-preferred (N) set of grids in the preference-based comparisons. The lower FID in each comparison indicates the better set. For ERR and RBP, higher value indicates better predicted performance. For the sake of easier comparisons, we multiply the output of our metrics by 100. Bold: The better predicted.

Dataset	Metric	A vs B		A vs B'		B vs B'	
		Preferred (P)	Not-preferred (N)	Preferred (P)	Not-preferred (N)	Preferred (P)	Not-preferred (N)
MS-COCO	FID	27802.9	28991.4	23647	26586.6	26673.1	28757.2
	ERR	98.75	98.69	98.79	98.75	98.72	98.71
	ERR $_{\eta}$	91.23	90.99	91.54	91.35	91.35	91.25
	RBP	89.01	88.60	89.36	88.96	88.77	88.69
	RBP $_{\eta}$	80.73	80.40	81.04	80.75	80.64	80.54
LN-COCO	FID	24191.6	21983.9	22671.9	21237.3	23647.0	26586.6
	ERR	96.56	96.34	96.85	96.38	96.23	96.08
	ERR $_{\eta}$	90.8	90.27	91.43	90.35	90.08	89.76
	RBP	89.58	89.05	90.39	89.04	88.66	88.35
	RBP $_{\eta}$	64.65	64.27	65.04	64.29	64.2	63.96
Prompts	FID	3451.5	4748.6	3372.1	4792.3	9012.0	9133.8
	ERR	96.55	96.33	96.48	96.29	96.43	96.38
	ERR $_{\eta}$	90.91	90.41	90.76	90.27	90.55	90.49
	RBP	89.47	88.86	89.29	88.76	89.1	88.99
	RBP $_{\eta}$	64.85	64.5	64.73	64.39	64.63	64.61

Comparison with FID

Although we cannot measure prompt-level agreement for FID, we can measure the population-level agreement. If comparing two systems X and Y , we can use the set of all images in the Preferred grids *regardless of the source system*; call this set (P). Similarly, we have the set of images in Non-preferred grids (N). We measure FID between the following sets: 1) set of target images and all the generated images from the Preferred grids (P) and 2) set of target images and all the generated images in Non-preferred grids (N). We expect that the set of Non-preferred grids to show higher FID (lower similarity) with the target images. Our results (Table 5.8) show that, while FID is able to measure better performance (lower distance with target images) on the simplest set of prompts (i.e., MS-COCO), it fails to detect the better-performed set for challenging prompt sets. In fact, for LN-COCO, FID indicated that the Non-preferred sets of generated images had performed better when comparing both A vs B and A vs B'. However, it correctly measured a lower FID score for the preferred set in the comparison of B and B'.

We can compare the population-level consistency of FID with that of our proposed metrics. Similar to the FID analysis, we report the average of our metrics on both P and N sets of generated images. Unlike FID, all of our metrics were able to assign higher values when comparing any pairs of the TTI systems and on all the three datasets. All four of our metrics consider the generated images from Preferred grids (P) as the better set compared to the images from Non-preferred grids (N).

5.2.6 Takeaway

Because reliable offline evaluation metrics for TTI systems is critical for benchmarking purposes in the community [298], we proposed a set of evaluation metrics for set-based TTI inspired by grounding ideation criteria in traditional IR evaluation methods. We use existing and novel approaches to capture fluency, variety, novelty, and quality, while incorporating visual salience as a fundamental feature of image layouts. Moreover, to the best of our knowledge, this is the first work to explore assessing generated images as a *set*. We demonstrated the effectiveness of our family of metrics against the popular FID metric, showing that ideation-based measures were better aligned with human preferences. In fact, we showed that modeling sequential user browsing, image visual saliency as well as considering novelty and variety increases consistency with human preferences.

Although our results are positive, we believe there are many directions to continue offline TTI evaluation work. First, our models of visual salience and browsing, while validated on human preferences, could benefit from models customized for ideation tasks. Second, ideation and generation tends to be highly interactive, consisting of multiple turns where state can be carried across individual sessions and even between users. This complex interaction suggests that single turn metrics provide an incomplete picture of system effectiveness. Finally, our results may

provide insight into how to design *online* evaluation metrics for TTI based on more elaborate behavior and interaction than found in classic information retrieval settings.

5.3 Evaluation of LLM-based Applications

One of the long-lasting goals for intelligent agents [429] is for them to seamlessly interact with humans in natural language and help their end-users with their tasks, such as completing household tasks, math tutoring, and so on. The rapid development of open-source libraries [431, 222] helps that goal by simplifying the development of LLM-powered agentic applications for various user-centered tasks [230, 175, 382, 22, 281]. To ensure that the application’s behavior meets the requirements of the application developers, it is also crucial to assess its potential utility to end users [125], as this can significantly impact its improvement journey. Taking into account a range of applications, it is unrealistic to assume benchmarking for every domain, including but not limited to code generation [242], health care [18], and many others whose development we witness every day [431]. To be useful, these applications must be able to understand the expectations and requirements of their end-users, and to modify their actions and interactions to better align with user preferences. Thus, assessing their *utility* is a critical step in their improvement journey.

Moreover, directly evaluating agentic applications poses challenges, as current approaches predominantly rely on end-to-end success metrics i.e., whether the application accomplishes tasks [364, 363, 284]. However, understanding a user’s interactions with an application involves much more than success alone [207, 208, 456]. Consider math problem solving, although it is important that the application solves the problem correctly, its ability to present and explain solutions based on various *criteria*, such as completeness, conciseness, and clarity, is crucial. In another task, code completion, even an incomplete code suggestion can be useful when it provides substantial boilerplate code or proposes a framework to solve a task [125].

Furthermore, success is not always clearly defined for a task. Recognizing such criteria and being able to quantify them is essential to assess whether developer requirements are being satisfied and if the application brings utility to the end-users. Given the objective of assessing arbitrary applications, relying solely on end-to-end success metrics is untenable, due to the expansive range of tasks requiring automation. The question is *how to design a flexible methodology to assess the task utility for diverse set of applications?*

To bridge this gap, we introduce *AgentEval*, a framework to gauge the utility of LLM-powered applications. Its goal is to assess the utility by providing application developers with insights into how the current flow can be characterized. *AgentEval* builds on recent work showing that LLMs can be a scalable and cost-effective alternative to human evaluation for open-ended tasks [223]. *AgentEval* as illustrated in Figure 5.9, consists of the three following agents, formally defined in Section 5.3.1:

1. *CriticAgent* suggests the list of criteria based on the task description and a pair of solutions, where one is preferred over the other one (e.g., successful and failed examples). For instance, for math problems, the criteria could be efficiency and clarity of the proposed solution;

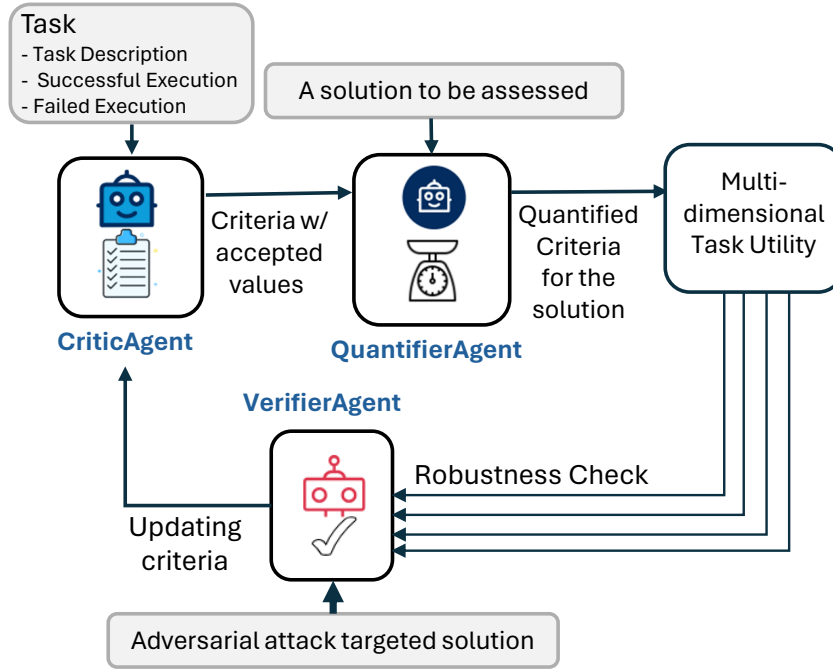


Figure 5.9: An overview of the AgentEval framework: *CriticAgent* creates a set of criteria and suggested values; *QuantifierAgent* quantifies the criteria for a considered application; and *VerifierAgent* verifies the criteria based on its robustness. The output of the *QuantifierAgent* is a multi-dimensional assessment of the utility of the application based on a suggested list of criteria and their evaluations.

2. *QuantifierAgent* quantifies how the solution performs for each criterion and returns the utility function, e.g. for math problems, if the *clarity* is ‘not clear’, ‘moderately clear’, or ‘very clear’;
3. *VerifierAgent* verifies the quality of the assessment of the suggested criteria to make sure the criteria are essential, robust, informative and have high discriminative power.

In summary, this section introduces *AgentEval*, a novel framework designed for assessing the utility of LLM-powered applications through collaborative agents. The core objective is to offer a scalable, robust, and flexible alternative to conventional evaluation approaches, particularly for open-ended and interactive tasks where traditional success metrics may not fully capture user-perceived value. To achieve this objective, we explore the following specific research questions:

- How effectively can LLM-powered agents identify meaningful evaluation criteria from task descriptions and execution examples?

- How discriminative and sensitive is the *AgentEval* framework in distinguishing between successful and failed task executions?
- How can we validate the evaluation conducted by the *AgentEval* framework, ensuring that the selected criteria are robust, stable, and effectively discriminate between varying levels of application performance?

In summary, in this section we introduce , a novel framework that leverages LLM-powered agents as a scalable and cost-effective alternative to human evaluations, to produce task utility through the collaboration of *CriticAgent* , *QuantifierAgent* and *VerifierAgent* ; and we provide in-depth analysis of *AgentEval* robustness for two applications across different solutions, that can be replicated on an unseen domain.

5.3.1 AgentEval for Task Utility

Figure 5.10 outlines a taxonomy of target tasks for LLM-powered applications, in terms of success metrics. At a high level, these tasks can be categorized into:

- 1) *Success is not clearly defined* — Users use the system in an assistive manner, seeking suggestions from it, rather than expecting it to solve the task. For example, a user can request the system to generate an email. The user usually uses the system’s response as a template, which can later be edited. Directly evaluating assistive tasks like these is hard, particularly for online evaluation, or when dealing with less well-defined tasks. One potential approach is to directly ask users how useful the help was, but this is not well-calibrated [51], hard to quantify [358], and expensive.
- 2) *Success is clearly defined* — It is clear whether the system solved the task or not, for example, assisting with household tasks, where success is clear and measurable. This category can be further divided into two subcategories:
 - *an optimal solution exists* — only one successful outcome is possible. For example, when asking an assistant to turn on a light, success is clearly defined, as there is only one way to do it.
 - *multiple solutions exist* — Increasingly, we observe situations where multiple trajectories of agent behavior can lead to success. For example, when asking an agent to suggest a food recipe, success could be multiple cuisines tasting good, but perhaps the recipe should not be expensive.

AgentEval is currently focused on tasks where *success is clearly defined and multiple successful solutions* may exist.

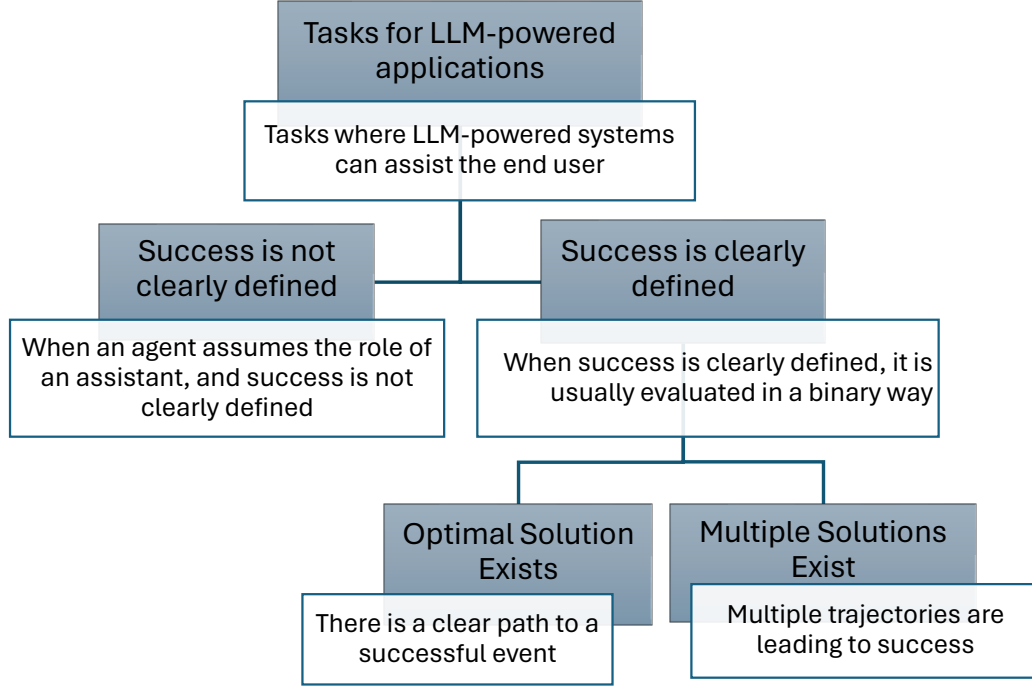


Figure 5.10: The taxonomy of tasks assessment.

Previous research on assistive agents suggests human pairwise preferences as one of the most optimal assessments, i.e. when the annotator is presented with two agents side by side and asked for their preferences [208]. In this setup of side-by-side pairwise comparison, humans tend to suggest a list criteria, explaining why they prefer one agent over the other. For instance, ‘the first agent was faster’ or ‘the second agent converses more naturally’. This comparative setup can guide humans to come up with a list of criteria that helps to infer the utility of the task. With this in mind, we designed *AgentEval* (Fig. 5.9), by employing LLMs to help us understand, verify, and assess task utility, namely:

- *CriticAgent* : The goal of this agent is to suggest a set of criteria that can be used to assess task utility. The *CriticAgent* is given a task description, as well as optionally several pairs of solutions, where preferably some are preferred over the other ones, for instance, successful and failed examples. *CriticAgent* would return a set of criteria $C = \{c_1, \dots, c_n\}$, where each criterion c_i is accompanied by a set of accepted values ω as $c_i : \{\omega_j\}_{j=1}^m$. For example, for solving math problems, the *CriticAgent* generated accepted values and criteria such as clarity, efficiency, and more - see Table 5.9.
- *QuantifierAgent* : The goal of *QuantifierAgent* is to quantify each of the suggested

criterion, to access the task utility of the system U_t , for the end user. We define the Utility for task t as: $U_t(s) = \{Q_i(s|c_i)\}_{i=1}^n$. where s represents the task sample and $Q(s|c_i)$ is the quantifier output for sample s based on the criterion c_i . For example, for math problem solving, given the generated criteria shown in Table 5.9, the solution’s Accuracy could be quantified as “Incorrect”, “partially correct” or “correct”. Eligible quantified values for quantification process are shown in “Accepted values” column in Table 5.9

- *VerifierAgent* : There might be cases where not all the criteria suggested by *CriticAgent* help assess utility. Some criteria might be redundant, while others may not aid in distinguishing performance. *VerifierAgent* validates the quality of the criteria in terms of robustness and their distinguishability of noisy samples. Essentially, it checks (1) if the criteria can be quantified robustly over repeated samples, and (2) if *QuantifierAgent* can identify the adversarial attacked targeted samples from the original ones. If the sanity checks do not pass, *VerifierAgent* will update the list of criteria, to end up with a set of robust, stable, informative and distinguishable criteria for assessment.

Finally, we note that *AgentEval* allows for incorporating a human in the loop in the role of a domain expert. For instance, *CriticAgent* could be replaced by a human expert who either comes up with the relevant criteria or helps *VerifierAgent* verify the useful criteria and filter out the unessential ones. We release all the code and data available at

5.3.2 Experimental Setup

This section provides an overview of the datasets utilized in *AgentEval* study i.e., Math problem solving and ALFWorld household task. The math dataset is chosen for its widespread usage and complex problem-solving scenarios that are fundamental in evaluating the effectiveness. ALFWorld dataset offers a scenario involving multi-turn interactions within a moderately approximated multi-modal environment. Each dataset plays a critical role in evaluating different aspects of *AgentEval*’s capabilities, from handling complex theoretical problems to navigating real-world scenarios. In both tasks, although success is clearly defined, multiple solutions exist for accomplishing the objectives. For example, when solving a math problem, there are various approaches one can take. Similarly, in the ALFWorld dataset, which involves household tasks, there are multiple ways to complete them based on how you search for objects and the thinking strategies you employ, among other factors. An example of Math problem solving and ALFWorld task is shown in Figure 5.11 and 5.12 respectively.

MATH Problem Solving

Dataset: For math problem solving, although success is clearly defined, multiple solutions exist for accomplishing the objectives. The MATH dataset is a substantial collection of 12,500 challenging mathematics problems from high school competitions [169]. Each problem comes with a step-by-step solution, enabling models to learn how to generate both derivations and explanations. The dataset covers a wide range of mathematical subjects and is tagged by difficulty levels, offering a nuanced measure of model performance across various aspects of mathematical problem-solving.

This dataset is particularly suitable for testing multi-agent systems for several reasons including:

1. The problems in the MATH dataset are not simple computations but require a deep understanding of mathematical concepts, heuristics, and problem-solving strategies.
2. Since the dataset includes step-by-step solutions, it allows for the assessment of an agent’s ability to learn and reason through a problem, not just its ability to arrive at the correct answer.
3. The variety of subjects and difficulty levels in the MATH dataset enables a comprehensive evaluation of a system’s versatility and adaptability in different mathematical domains which is crucial for multi-agent systems that are expected to operate across a range of scenarios.

Similar to math problem experimental setup in Wu et al. [431], we carry out two experimental evaluations which involves 120 problems from the most challenging category, and includes 20 problems each from six different categories, of number theory, counting and probability, prealgebra, algebra, intermediate algebra, and precalculus.

Solutions: In establishing solutions for this task to assess, we draw inspiration from the experiments showcased in [431]. We evaluate the proposed methodology by AutoGen [431], as well as Langchain ReAct [447]⁶ and a Vanilla solver that employs GPT-4 to tackle the task. These solutions have previously demonstrated promising and competitive performance [431]. In Section 5.4.1, we explore how the measured performance with *AgentEval* correlates with the ground truths.

ALFWorld Household Task

Dataset: ALFWorld, presents a set of language-based interactive decision-making tasks within simulated household environments [364]. This benchmark is distinguished by its

⁶<https://python.langchain.com/>

```

“problem”: “How many positive and negative integers is 12 a multiple of?”
“level”: “Level 5”
“type”: “Prealgebra”
“response_with_ans”:
“messages”: [
{
“content”: “How many positive and negative integers is 12 a multiple of?”,
“role”: “user”
},
{
“content”: “To determine how many positive and negative integers 12 is a multiple
of, we need to find the divisors of 12. Then we will double the number of divisors, as
each positive divisor has a corresponding negative divisor. Let’s use a Python code
snippet to find all the divisors of 12 and count them.”,
num = 12
divisors = [i for i in range(1, num + 1) if num \% i == 0]
positive_divisors_count = len(divisors)
negative_divisors_count = positive_divisors_count
total_divisors_count = positive_divisors_count + negative_divisors_count
print(f"Positive_divisors_of_12:_{divisors}")
divisors_negative = [-x for x in divisors]
print(f"Negative_divisors_of_12:_{divisors_negative}")
print(f"Total_divisors_count:_{total_divisors_count}")

Execute this code snippet to find the divisors of 12 and count how many positive
and negative integers 12 is a multiple of.
“role”: “assistant”
}
],
“time”: 17.360238790512085

```

Figure 5.11: An example of Math Problem solving with solution generated by AutoGen

diversity of tasks, offering a comprehensive platform for testing AI and multi-agent systems. This benchmark is particularly suited for such evaluations because first, ALFWorld is

```

{ "content": "Perform actions and interact with a household to solve a task.
  At the beginning of your interactions, you will be given the detailed
  description of the current environment and your goal to accomplish. For
  each of your turn, you should choose from two actions: "think" or
  "Action". If you choose "think", you should first think about the current
  condition and plan for your future actions, and then output your action
  in this turn. Your output must strictly follow this format:"think: your
  thoughts."
  "Action: your next action\\n"; If you choose "ACTION", you should directly
  output the action in this turn. Your output must strictly follow this
  format:"ACTION: your next action".
  After each of your turn, the environment will give you immediate feedback
  based on which you should plan your next few steps. if the environment
  output "Nothing happened", that means the previous action is invalid and
  you should try more options.
  Reminder:
  1. the action must be chosen from the given available actions. Any actions
  except provided available actions will be regarded as illegal.
  2. Take the target household immediatly after you find it.
  3. Reply 'TERMINATE' only under two circumstances: a). The task has given you
  clear instructions to return 'TERMINATE' b). The task seems insolvable
  .\\n Here are two examples.\\nYou are in the middle of a room. Looking
  quickly around you, you see a cabinet 13, a cabinet 12, a cabinet 11, a
  cabinet 10, a cabinet 9, a cabinet 8, a cabinet 7, a cabinet 6, a cabinet
  5, a cabinet 4, a cabinet 3, a cabinet 2, a cabinet 1, a coffeemachine
  1, a countertop 1, a diningtable 1, a drawer 1, a fridge 1, a garbagecan
  1, a microwave 1, a shelf 3, a shelf 2, a shelf 1, a sinkbasin 1, a
  stoveburner 4, a stoveburner 3, a stoveburner 2, a stoveburner 1, and a
  toaster 1.\\nYour task is to: put a hot apple in fridge.\\n",
  "role": "user"},
...
...

{ "content": "You open the microwave 1. The microwave 1 is open. In it, you
  see a egg 1.",
  "role": "user"
},
{ "content": "ACTION: heat mug 2 with microwave 1",
  "role": "assistant"},
{ "content": "You heat the mug 2 using the microwave 1.Task success, now
  reply TERMINATE\\n",
  "role": "user"},
{ "content": "TERMINATE",
  "role": "assistant"},
{ "content": "TERMINATE",
  "role": "user"}

```

Figure 5.12: An example of ALFWorld Householding task from “Pick, Heat, then Place” category.

the first interactive parallel environment that aligns text descriptions and commands with physically embodied robotic simulation. It extends two prior works: TextWorld, an engine for interactive text-based games, and ALFRED, a large-scale dataset for vision-language instruction following in embodied environments [363, 109]. The cross-modality framework of this benchmark allows for a variety of embodied tasks with corresponding text-based counterparts, allowing agents to be trained and evaluated in both the language and embodied world. In addition, ALFWorld supports the development of agents that can reason both abstractly and execute actions concretely, mimicking human-like decision-making processes in varying contexts. Finally, the dataset’s inclusion of a wide range of tasks, from household chores to more intricate problem-solving scenarios, provides a comprehensive testbed for evaluating the adaptability and problem-solving capabilities of AI and multi-agent systems. In general, the dataset allows agents to explore, interact, and learn in an abstract language environment before dealing with the complexities of an embodied environment.

Solutions: As for the solutions to assess for ALFWorld Household tasks, similar to [431], we consider ReAct [447] as well as AutoGen with two agents and AutoGen with three agents [431]. ReAct is an agent that operates within the ALFWorld environments and is responsible for suggesting plans and executing actions. On the other hand, AutoGen Two-Agent System consists of an LLM-backed assistant agent responsible for suggesting plans, and an executor agent responsible for executing actions in the ALFWorld environments. Both ReAct and this solution occasionally struggles with leveraging basic commonsense knowledge about the physical world, which can lead to repetitive errors and getting stuck in loops. In AutoGen with three agents, a grounding agent is provided just for the sake of critical common sense knowledge whenever the system exhibits early signs of recurring errors. We assess and compare the performance of these three solutions using *AgentEval*.

Implementation Details

For all experiments, we use GPT-4 version 0613, accessed through Azure OpenAI services, as the LLM model and the temperature of 0. *AgentEval* utilizes AutoGen [431] for implementation, since it provides a versatile environment where agents can be finely tuned and customized based on specific application needs. This is crucial for maintaining the flexibility to handle a wide range of applications. We tried to avoid much prompt engineering and tried to keep each agent’s instructions as if we are instructing human annotators. Moreover, another advantages of using AutoGen for implementation of *AgentEval* is that it has the flexibility to involve human in the loop. Each agent could be replaced by a human annotator. We further provide all the prompts used in our experiments in our Git repository.

5.4 AgentEval Results

Table 5.9: Verification Criteria for MathProblems

Criteria	Description	Accepted Values
Clarity	The ease of understanding the steps, explanations, and language used in the solution.	– Not Clear (0) – Moderately Clear (1) – Very Clear (2)
Efficiency	The use of optimal methods or approaches to solve the math problem.	– Inefficient (0) – Moderately Efficient (1) – Efficient (2)
Error Analysis	The identification and description of possible errors or misconceptions in the math problem-solving process.	– Not Addressed (0) – Partially Addressed (1) – Well Addressed (2)
Completeness	Quality of code in terms of efficiency and elegance	– Incomplete (0) – Mostly Complete (1) – Complete (2)

5.4.1 AgentEval for Math Problems

When executing the *CriticAgent* for Math problem solving, we first obtain a set of criteria as presented in Tab. 5.9. Then, the *QuantifierAgent* is tasked with quantifying each criterion, based on the accepted values. We present the outcome of *QuantifierAgent* measuring performance of three solutions on this task in Figure 5.13. Notably, we see that *Agenteval* does not quantify the three solutions as if they perform equally well across the different criteria. For instance, while all three solutions leverage GPT-4 as the underlying language model, Autogen outperforms ReAct and Vanilla GPT-4 in terms of accuracy. This observation, while confirmed by previous studies [431], extends to solution completeness and efficiency as well.

As depicted in Figure 5.13, the error analysis range of quantified values differs from other metrics. We scrutinize the results by categorizing them into successful and failed cases. AutoGen, Vanilla Solver and ReAct solutions are each presented in orange, blue and green respectively, where the darker bars represent the performance on successful cases and lighter bars represent the failed cases. The difference between the dark and light bar of each color,

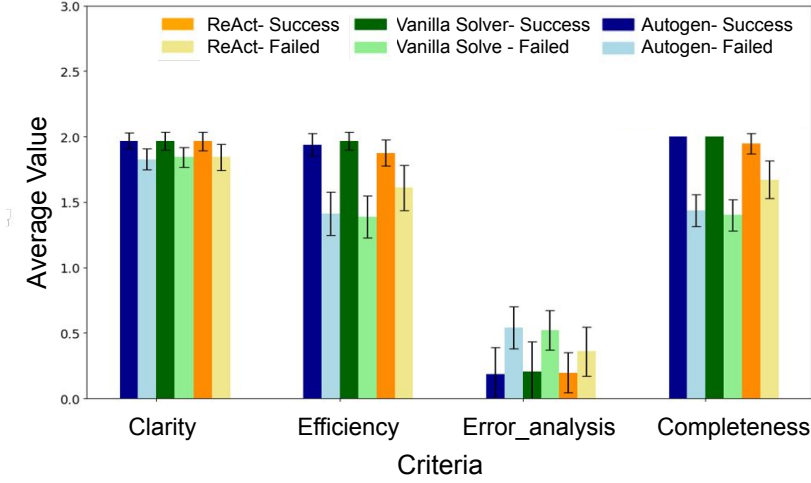


Figure 5.13: *AgentEval* assessment of three solutions on math problems categorized by success and failed cases.

verify *AgentEval*’s performance, as we expect that each positive criteria should be quantified higher for successful cases compared to their failed cases. We observe that in most cases, the successful and failed cases are distinguished, even with 95% interval confidence on all the success and failed cases.

When examining the differences between successful and failed cases among the three solutions, we note that not all successful cases are assessed identically, nor are all failed cases quantified with the same performance. This can be interpreted to mean that even though two solutions might both be successful, one might perform better or worse in certain criteria, such as clarity or efficiency. This observation provides us with valuable additional insights, especially for the developers of the proposed solutions, and goes beyond reporting the effectiveness of a application by one scalar value e.g., success rate.

AgentEval Results for ALFWorld

To study the generalizability of *AgentEval*, we repeat the experiments in 5.4.1 for *AlfWorld*, in which real-world household environments are emulated through textual interfaces [364]. We provide the criteria created for this task as well as the results for three solutions of this task in Tab. 5.10 and 5.14, respectively. Following the extraction of a set of criteria as detailed in Tab. 5.10, these criteria are passed to the *QuantifierAgent* for quantification on each sample. Figure 5.14 presents the results for three introduced solutions: AutoGen with two agents, AutoGen with three agents, and ReAct, on the 134-test set from [431]. Further, we also verify the measured performance of *AgentEval*, by comparing the failed/success

cases in the right plot of Figure 5.14. As confirmed here and in math problem solving, AgentEval is clearly able to distinguish between different solutions in terms of different criterion.

We provide an example of AgentEval applied to the ALFWorld house-holding task, as mentioned in Sec. 5.4.1 in which real-world household environments is emulated through textual interfaces [364]. When running the *CriticAgent* on this task, it identified specific criteria such as “Task understanding”, “Plan making” and “Response to Feedback” as outlined in Tab. 5.10. We consulted researchers deeply involved with these tasks, and their expertise confirmed that these criteria are critically relevant and significant similar to [223].

For example, given that these tasks are language-based and require interactive decision-making, an agent in ALFWorld is tasked with high-level objectives, such as placing a hot apple in the fridge, and must navigate and interact with a simulated household environment to achieve these objectives. Therefore, criteria displayed in Tab. 5.10 satisfy the assessment of this task. While the criteria are pretty self-descriptive, about the criterion “Use of TERMINATE” we note that the agent is prompted to use the term “TERMINATE” upon task completion, which is closely correlated with task success.

Following the extraction of a set of criteria as detailed in Tab 5.10, these criteria are passed to the *QuantifierAgent* for quantification on each sample. Figure 5.14 presents the results for three introduced solutions: AutoGen with 2 agents, AutoGen with 3 agents, and ReAct, on the 134-test set from [431]. It is important to note that all criteria, except “Use of TERMINATE” and “Correctness of Action” employ a five-level grading system, while these two criteria are binary. From this figure, it is evident that ReACT performs notably worse across all criteria, while AutoGen with 2 agents and 3 agents demonstrates competitive performance.

Table 5.10: Verification Criteria for ALFWorld Housholding Tasks.

Criteria	Description	Accepted Values
Task Understanding	How well the participant was able to comprehend the problem set and follow the task instructions	– Excellent (4) – Good (3) – Average (2) – Poor (1) – Terrible (0)
Plan Making	The ability of the participant to strategize and make a plan for tackling the task.	– Excellent (4) – Good (3) – Average (2) – Poor (1) – Terrible (0)
Action Decision	The participant’s decision-making skills in choosing the right action to perform.	– Excellent (4) – Good (3) – Average (2) – Poor (1) – Terrible (0)
Action Execution	How effectively the participant is able to execute the chosen action.	– Excellent (4) – Good (3) – Average (2) – Poor (1) – Terrible (0)
Response to Feedback	How well the participant adapts his/her next steps based on the feedback from the environment	– Excellent (4) – Good (3) – Average (2) – Poor (1) – Terrible (0)
Correctness of Action	The correctness of the action performed by the participant with respect to the available actions and the current context	– Correct (1) – Incorrect (0)
Use of Terminate	Whether the participant uses the ‘TERMINATE’ command appropriately	– Appropriate (1) – Inappropriate (0)

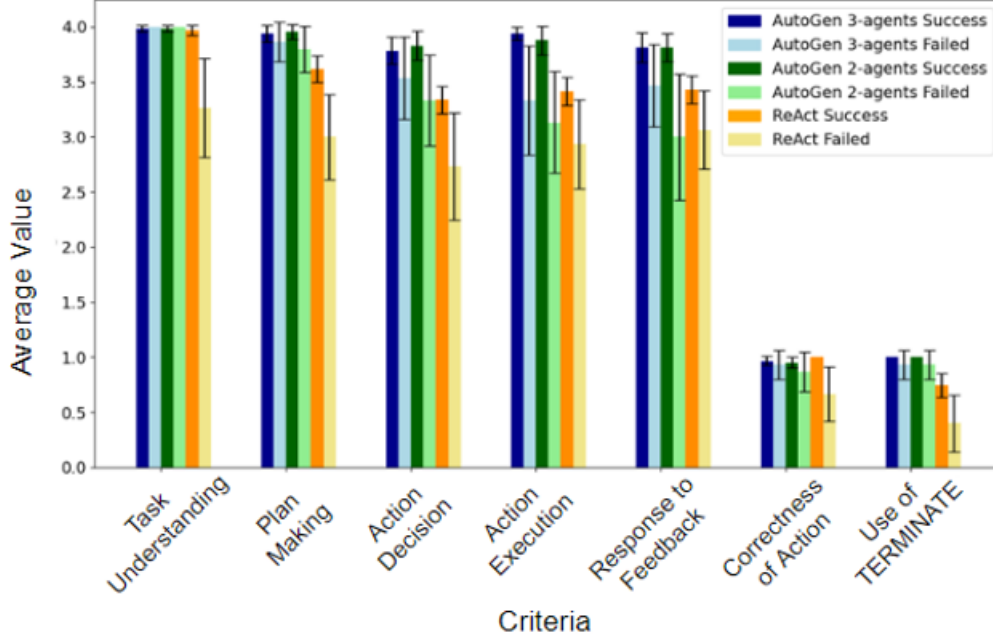


Figure 5.14: AgentEval assessment of three different solutions on ALFWorld house-holding Tasks categorized by success and failed cases.

We also categorizes the 134 games into groups of failed and successful ones. Similar to 5.13, darker colors represent performance in successful cases for each solution, while lighter colors represent performance in failed cases. AutoGen 3-agent, AutoGen 2-agent, and ReAct are represented by blue, green, and orange, respectively. For most criteria, the distinction between failed and successful cases is clear, even within a 95% confidence interval. However, for certain criteria, such as “Task understanding” all solutions, whether they failed or succeeded, exhibit very similar performance. This could be interpreted as either (1) all solutions have a good understanding of the task, even if they fail to complete it, (2) this criterion may be redundant, as it does not provide additional information among these three solutions or (3) the *QuantifierAgent* is unable to score the criterion in a meaningful way. We refrain from concluding which criteria are most suitable for this specific task. Instead, we emphasize the importance of conducting a more in-depth analysis of performance beyond success rates, tailored to one’s goals and application requirements. Later, we show that how using *VerifierAgent* could be helpful in identifying criteria with higher discriminative power and more robustness.

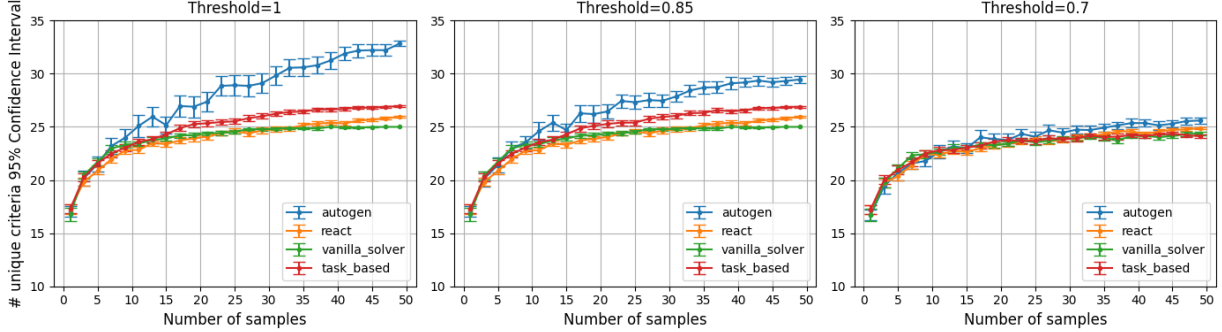


Figure 5.15: Task-based vs solution-based criteria for Math problems. Error bar show the 95% confidence interval.

5.5 Criteria Diversity

Here, our main goal is to study the diversity of the suggested criteria. We investigate the extent inputs to *AgentEval* (Figure 5.9 such as ‘Task Description’ and ‘Successful/Failed Executions’) contribute to *CriticAgent* for creating a more diverse set of criteria. To do so, we use two distinct methods, with *CriticAgent* generating (1) “task-based” criteria solely from the task description, and (2) “solution-based” criteria, derived from both the task and execution examples. For example, a solution to a mathematical problem, might satisfy criteria such as ‘Accuracy’ and ‘Clarity’, independent of the solution. However, when additional tools such as coding are used to solve the problems, additional criteria like ‘Code Efficiency’ may be introduced to the set of criteria. This makes sense, since the application leveraged coding to solve math problems.

Fig. 5.15 displays the number of unique criteria extracted for mathematical problem solving in task-based mode, and three different solution-based approaches. To keep the balance between computational costs and analyzing the robustness, we conducted 50 runs of the *CriticAgent* with different seeds. Subsequently, for $N = 50$ iterations, we randomly select $M \leq 50$ samples, as shown on the x-axis of Fig. 5.15, and present the average number of unique extracted criteria, along with its 95% confidence interval after repeating this process 50 times. We note that because the total pool of criteria includes 50 iterations in total, the confidence intervals become smaller when M get closer to the maximum number of samples i.e., 50

To gain deeper insights into diversity of criteria, we took a closer look at them to study if they are truly unique or to what extent they have similarities. This is important to determine if *CriticAgent*, when continually generating criteria, will always produce new criteria, or if it will eventually converge to a set. We noted that some criteria are similar

Table 5.11: Example pairs of similar criteria.

- Problem Difficulty: The complexity of the math problem that has been solved.
- Problem Complexity: The level of difficulty of the problem.
- Innovativeness: The novelty and creativity in the approach to solve the problem
- Innovation: The ability to solve a problem using a unique or creative method not commonly known.
- Time Taken: The time taken to solve the problem.
- Time to Completion: The amount of time taken to solve the problem completely
- Understandability: The clarity and ease of comprehension of the solution provided.
- Readability: How easy it is to comprehend the provided solution.

but worded differently. For example, ‘Problem Complexity’ vs. ‘Problem Difficulty’ or ‘Time Taken’ vs. ‘Time to Completion’. Tab. 5.11 lists such instances. To consolidate the similar criteria and reduce noise in the number of unique criteria and redundancy, inspired from previous work [245, 399, 331], we employ a pre-trained language model fine-tuned for paraphrasing⁷, to measure the semantic similarity of criteria descriptions. Using a threshold τ , we classify pairs with cosine similarity greater than τ as semi-identical ones and select one of them as the representative of the pair. Figure 5.15 illustrates the impact of different τ values (0.7, 0.85, 1) on the diversity of criteria. A threshold of 1 means no filtering occurs. This analysis shows that the solution-based approach has potential to produce more diverse criteria than the task-based approach, although this varies by the creativity of the model. For example, while the AutoGen solution demonstrates the highest diversity, task-based methods yield more unique criteria than ReAct and Vanilla Solver. Another interesting observation is that repeating the *CriticAgent* will eventually lead to a convergence in the number of criteria. This suggests that the *CriticAgent*’s ability to create new criteria will diminish, converging to an almost finite list of criteria, which will reduce the cost as well.

5.6 Verification

As outlined in Section 5.3.1 and illustrated in Fig. 5.9, the *VerifierAgent*’s primary role is to ensure the selected criteria are effective toward evaluating the utility for the end-user, while

⁷<https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2>

Algorithm 1 *VerifierAgent*

```
1: for  $i = 1, 2, \dots, 50$  do
2:   Run CriticAgent with  $seed = i$  to obtain a set of criteria  $C_i$ 
3: end for
4: Obtain summarized_criteria by using another LLM agent to summarize  $C_1, C_2, \dots, C_{50}$ .
5: for  $i = 1, 2, \dots, 18$  do
6:   for all  $s$  in  $S$  do
7:     Run QuantifierAgent with  $seed = i$  on sample  $s$ 
8:   end for
9: end for
10: for all  $crit$  in summarized_criteria do
11:   for all  $s$  in  $S$  do
12:     Compute the coefficient of variation of  $s$ 's quantified result with respect to  $crit$ 
        across all  $seed$ 
13:   end for
14:   Compute mean coefficient of variation by averaging all sample's coefficient of vari-
        ation
15: end for
16:  $final\_criteria \leftarrow []$ 
17: for all  $crit$  in summarized_criteria do
18:   if  $crit$  has a mean coefficient of variation within a certain range, and  $crit$  has decent
        adversarial testing performance then
19:     Add  $crit$  to final_criteria
20:   end if
21: end for
22: To evaluate future tasks, use final_criteria with QuantifierAgent.
```

maintaining robustness and high discriminative power. To achieve this, the *VerifierAgent* undertakes two main actions:

1. *Criteria Stability*: The criteria should be essential and robust, meaning they should not be redundant and we should be able to quantify them stably if we repeatedly quantify it for an individual solution, showing no divergence. As such, *VerifierAgent* enhances the criteria by iterating over the generation and quantification phases. It then consolidates these criteria by identifying and eliminating redundancies, followed by evaluating the dispersion of the distribution of the quantified criteria. This step modifies the criteria, ensuring that only the most robust criteria are retained.
2. *Discriminative Power*: A reliable evaluation should detect and withstand noise. To

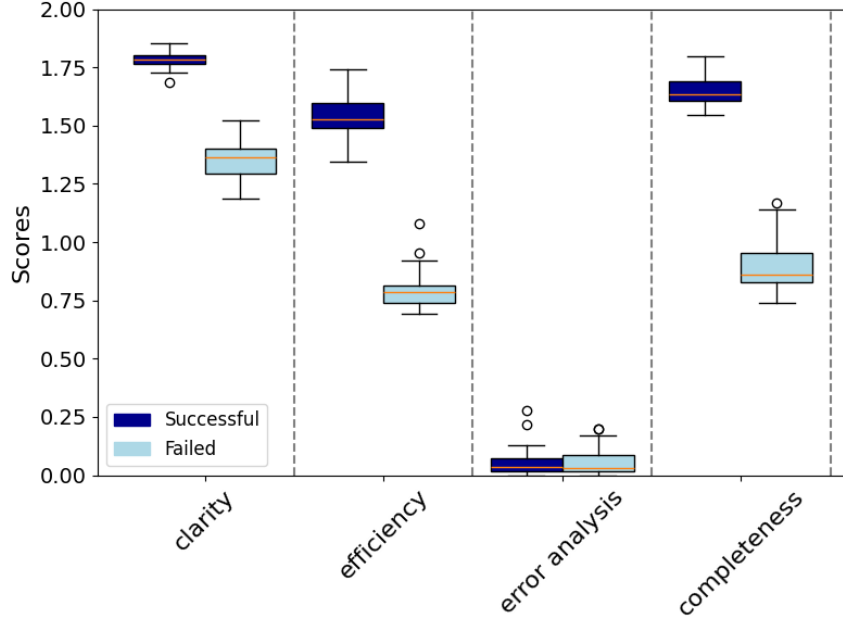


Figure 5.16: Distribution of *QuantifierAgent* output on AutoGen results on successful (dark blue) and failed (light blue) cases on different criteria.

test that, we propose to use adversarial examples and then assess the system’s ability to differentiate between these compromised examples and standard cases. Should the system fail to distinguish effectively, it indicates that the criteria are insufficient for reliable assessment under varied conditions.

We note that both steps involve a tunable threshold that can be adapted based on application needs, ensuring flexible criteria validation. The proposed methodology for *VerifierAgent* is summarized in Algorithm 1. We note that, due to the high computational costs associated with these experiments, we did not perform ablation studies on the number of seeds or the repetition counts for criteria iteration. Consequently, we defer the analysis of sensitivity regarding iteration counters (*i*) in Algorithm 1 to future research.

5.6.1 Criteria Stability

Our goal here is to explore the stability of criteria and robustness of the quantifier for having a more essential, robust and stable set of criteria. We specifically evaluate the *QuantifierAgent*’s robustness using criteria for mathematical problems (Table 5.9), conducting 50

repeats of runs with different seeds on 120 problems (Section 5.3.2). Ideal expected outcomes include consistent performance across all criteria on all the repeats. Figure 5.16 illustrates the distribution of quantifier values for both failed (dark blue) and successful cases (light blue) across all criteria through box plots. The more robust a criterion, the narrower the range of quantified performance (narrower box plots). Also, the less overlap between the successful and failed boxes, the higher the distinguishability of the criteria. We observe that all four criteria, except ‘error analysis’ allow for easy differentiation between successful and failed cases. Additionally, some criteria prove to be more robust compared to others. We believe that such an analysis of the quantifier agent’s performance will yield valuable insights for enhancing reliability, trustworthiness, and explainability in performance evaluation.

Further, to refine and expand the criteria set without redundancy, we operate the *CriticAgent* multiple times i.e., we execute *CriticAgent* 50 times with varied seeds. The criteria are then summarized into one list of useful criteria using the LLM. Additionally, as explained in Section 5.5, we remove similar and redundant criteria using pre-trained language models, thus obtaining a comprehensive list of criteria.

Now, we aim to determine the stability of these criteria through repeated quantifications. Our goal is to identify criteria that maintain consistent results without significant divergence, even when quantified multiple times. Using this consolidated list, we measure the dispersion of quantified results using the coefficient of variation, a standardized metric that facilitates comparison across various test cases when *QuantifierAgent* quantifies them. Given the consolidated list of criteria, we use the *QuantifierAgent* to quantify various test cases and report the coefficient of variation as a measure of the dispersion of the *QuantifierAgent*’s outputs with respect to each criterion across different seeds and report the mean coefficient of variation across all samples. we run *QuantifierAgent* with 50 seeds and plot the change (Δ) in the sum of mean coefficient of variation across all criteria against the number of seeds, in Figure 5.17. For each criterion, we compute the absolute difference with the mean coefficient of variation calculated when using $n - 1$ seeds, summing up the absolute differences across all criteria. According to the plot, after approximately 18 seeds, the magnitude of mean coefficient of variation stabilizes and becomes rather trivial. In almost all cases, the mean coefficient of variation is around or below 0.5, which is relatively small, suggesting that *QuantifierAgent* is quite robust.

5.6.2 Quantification Robustness

To study the robustness of the *QuantifierAgent*, we selected a specific subset of criteria related to mathematical problems, as detailed in Table 5.9, and conducted 50 runs of the quantifier agent on the 120 problems described in Section 5.3.2. Our expectation is to observe consistent quantified performance for each of the criteria. In Figure 5.18, we

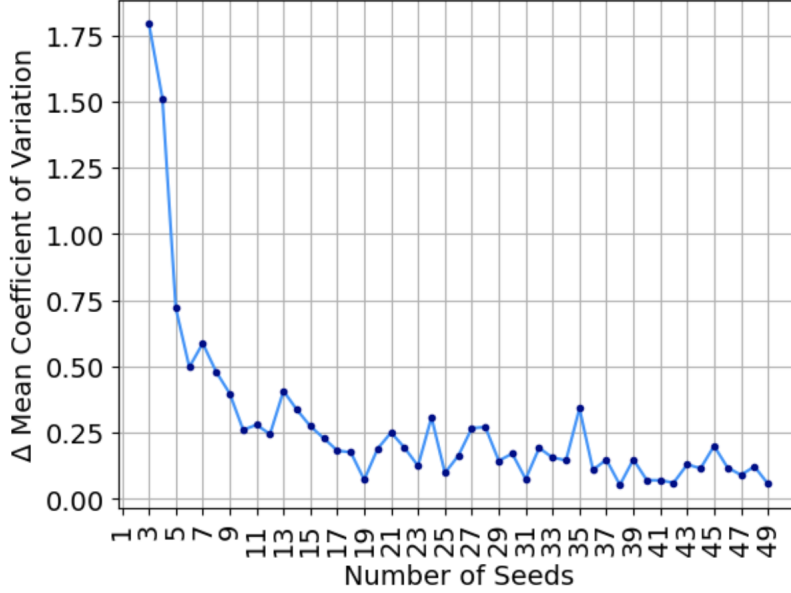


Figure 5.17: Δ sum of mean coefficient of variation across all criteria with increasing number of seeds.

present the distribution of quantified performance across 50 runs for both successful and failed cases, focusing on the five selected criteria. A consistently horizontal performance trend indicates greater robustness in the quantifier, whereas more fluctuations in the figure suggest less robustness and a noisier performance of the agent.

As shown in the results, for four out of the five generated criteria, we consistently observe steady performance. Not only do the success cases consistently outperform the failed cases, but their performance also falls within a similar range across runs. However, when it comes to the “error analysis” criterion, we observe a more variable performance of the quantifier. It does not consistently predict one group (success or failed) to perform better than the other, and the quantifier’s performance varies across different runs. This suggests that the AgentEval tool may not exhibit promising robustness for this particular criterion. The underlying issues could be either the criterion itself lacks clarity and appropriateness for the task, or the *QuantifierAgent* struggles to quantify this criterion effectively. In either case, it is advisable to either modify or eliminate this criterion to enhance trustworthiness and reliability. We further show that *VerifierAgent* is designed to take care of such criteria.

We recognize the importance of thoroughly investigating the robustness of each criterion in quantification studies. This analysis is crucial as it sheds light on the stability of each criterion. Moreover, when ground truths are available, such as in cases of success versus

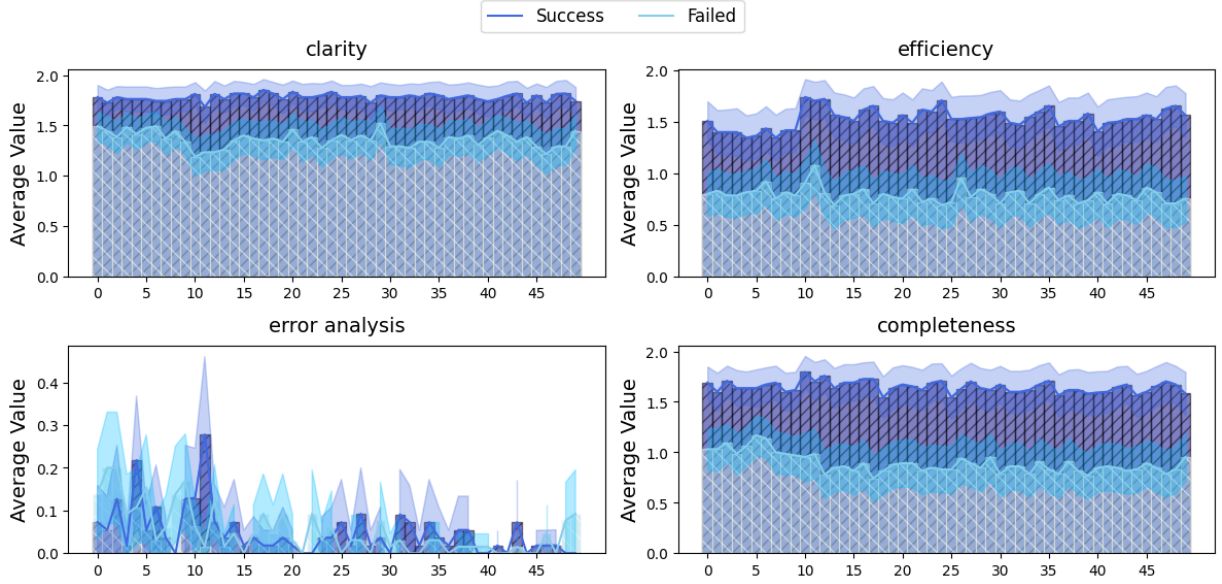


Figure 5.18: Quantifier Robustness on criteria of Math Problem Solving problem. Each bar represent the average performance of success (dark blue “//”) and failed (light blue “\\”) cases and 95% interval on each set is shaded across the average point. The two plots are overlaid.

failure, they provide a benchmark to validate our assessments. Additionally, it is important to acknowledge that not all criteria exhibit the same level of robustness. This variability demands careful consideration during evaluations, especially given the non-deterministic nature of LLMs. Such awareness is essential to ensure the reliability and accuracy of our assessments in the dynamic field of LLMs.

5.6.3 Discriminative Power

It is crucial to ensure the quality of quantification of each criterion. Ideally, this validation would involve comparisons with known pairwise samples, where sample S^+ is definitively superior to S^- for a given criterion. If the evaluator also confirms superiority of S^+ w.r.t S^- , it has robust quantification. However, due to rapid expansion of LLM-powered applications, obtaining annotated data for many tasks is often unfeasible. Therefore, we propose using synthetically altered versions of samples for verification. Let us assume we have an alternative disturbed version of sample S , which is called S' . Assuming sample S is more likely to outperform its disturbed version S' , our assessment should confirm this assumption by assigning better quantified performance S in comparison to S' . In experiments

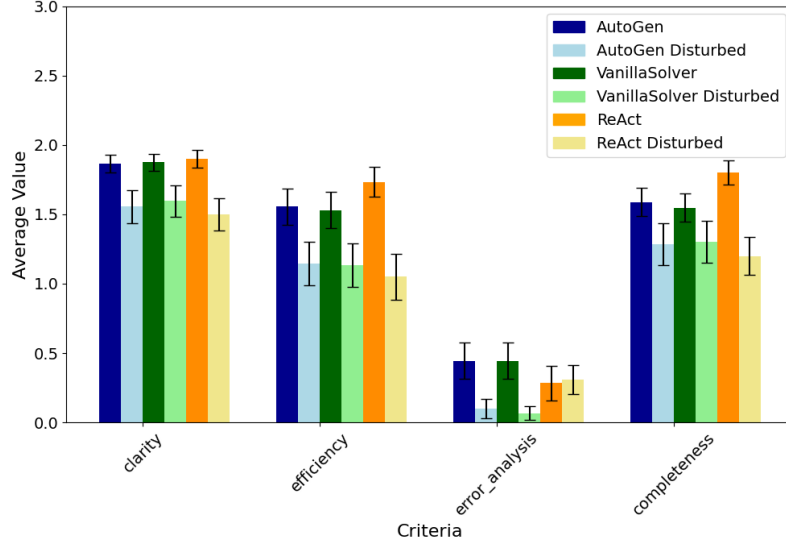


Figure 5.19: Assessment of original and disturbed solutions on Math dataset (discriminative power study).

with mathematical problems, we introduced random noise by removing portions of the solution sentences from AutoGen, VanillaSolver, and ReAct’s results respectively, expecting that criteria like ‘Completeness’ or ‘Clarity’ would show be higher in S than in S' . We disturbed solutions by removing 25% of the sentences and assessed the *QuantifierAgent*’s performance. As shown in Fig. 5.19, criteria measuring aspects like ‘Clarity’ and ‘Completeness’ were lower in disturbed solutions (lighter bars), confirming *QuantifierAgent*’s high discriminative power and effectiveness. We defer a detailed analysis of sensitivity with respect to varying noise levels to future work.

We have already filtered out the criteria that were unstable, i.e., those that had a high mean standard deviation and dispersion when being quantified in the previous section. In most cases, the *QuantifierAgent* quantifies the disturbed output to be worse than the original task output. We believe analyzing the *QuantifierAgent*’s performance will enhance the reliability, trustworthiness, and explainability in evaluations.

5.6.4 *VerifierAgent*

After modifying the list of criteria (Section 5.6.1), we have developed a stable and robust list of criteria that the *QuantifierAgent* can reliably quantify. Further, we also proposed a method for assessing whether the criteria can distinguish between noise-adversarially

attacked samples and the original ones. These two tests will serve as input for the *VerifierAgent* (described in Algorithm 1), which can also have its threshold tuned for different applications. For instance, one might prioritize the stability of the criteria, while another may value the discriminative power of the *AgentEval* for specific applications. As such, the *VerifierAgent* will modify and update the criteria based on to what extent they pass the two tests, i.e., if the mean coefficient of variation is below a specific threshold and the percentage of adversarial testing it has passed. The *VerifierAgent* will then update the criteria if necessary. We believe that having a *VerifierAgent* would help continuously updating the criteria as needed because, by improving the systems, we may require new criteria that were not previously necessary for utility assessment.

5.6.5 Limitations

Here, we discuss some limitations of the *AgentEval* framework. Firstly, the performance of the *AgentEval* is highly dependent on the quality of the output logs of the applications. Flaws or limitations in these outputs can significantly impact the framework’s ability to accurately assess utility.

Secondly, our experiments were conducted exclusively with closed-source LLMs, specifically with GPT-4. This may limit the generalizability of our findings. Plans to include a broader array of LLMs, including open-source models, are considered for future studies to validate and possibly enhance the robustness of our conclusions. Additionally, the tests conducted were limited to specific scenarios within math problem solving and household tasks. Expanding the diversity of test scenarios could help in understanding the broader applicability of the framework.

Thirdly, while *AgentEval* employs a novel methodology leveraging LLMs to estimate utility, the absence of human evaluation in our validation process could be viewed as a drawback. Human evaluations provide unique insights, especially in subjective aspects of utility that automated systems might overlook. However, such evaluations are often cost-prohibitive and logistically challenging, restricting our ability to implement them within this study. Especially do developers of agentic LLM-powered applications who needs insights fast as they go with the deployments.

Lastly, as LLM technologies evolve, the criteria and metrics used for evaluation may need to be updated or revised. What works for assessing current LLMs may not hold as these models become more advanced. Continuous updates to the evaluation framework will be necessary to keep pace with technological advancements.

5.6.6 Takeaway

We introduced the *AgentEval* framework, designed to swiftly gauge the utility of arbitrary LLM-powered agentic applications. Our framework leverages recent findings suggesting LLMs as a scalable and cost-effective alternative to human evaluations for open-ended tasks. *AgentEval* consists of three agents: *CriticAgent* suggests criteria based on task descriptions and executions of the applications, *QuantifierAgent* quantifies how well the application flow aligns with these criteria, and *VerifierAgent* modifies the list of criteria if needed. This framework is customizable, adaptable, and can operate in various modes, employing combinations of LLMs, human inputs, and tools. We believe that suggested *AgentEval*'s utility extends beyond immediate performance. It can uncover new system capabilities over time and adapt to changes in user needs tracked by developers. *AgentEval* can also enable developers to assess the alignment between application behavior and suggested user requirements, providing them with insights into areas for improvement. In summary, our contributions include introducing the *AgentEval* framework, and conducting a robust analysis of its performance across various datasets and baselines. *AgentEval* represents a significant step towards assessing LLM-powered applications.

Chapter 6

Concluding Remarks

In this final chapter, we summarize the research presented in this thesis, highlight its core contributions, and discuss the broader implications of evaluation of information seeking systems in the generative era. We also outline promising directions for future work, reinforcing the central argument that evaluation methodologies must evolve in step with system development to remain effective.

6.1 Summary of Contributions

Dealing with Sparse labeling. We explored the effects of incomplete or sparse labeling on IR system evaluation. We demonstrated that when relevance judgments are sparse (as is common in large-scale retrieval benchmarks), evaluation outcomes can be biased. In particular, strong retrieval models that succeed in retrieving relevant documents often appeared to underperform simply because many of the documents they retrieved were unjudged (not labeled in the pool) and thus treated as non-relevant. This analysis confirmed a critical gap in traditional evaluation: current metrics underestimate systems that retrieve relevant-but-unlabeled information. This finding set the stage for developing new evaluation strategies to ensure fair comparisons among IR systems, even under labeling sparsity.

To address the bias introduced by sparse labels, we introduced a novel evaluation methodology based on the Fréchet Distance. Drawing inspiration from techniques in other domains, we proposed using Fréchet Distance to measure the divergence between the distribution of retrieved results and the distribution of known relevant items. Our experiments, conducted on large-scale benchmarks (e.g., the MSMARCO and TREC Deep Learning datasets), showed that the Fréchet Distance-based metric is more robust to sparse judgments and

provides a stable basis for comparison. Notably, this method enabled fair comparison between traditional retrieval systems and generative systems that produce answers, since it does not require exhaustive relevance labels for every possible answer.

LLM-based Relevance Judgments for IR Evaluation In this thesis, we further explored the potential of large language models (LLMs) as evaluators for information retrieval systems. Specifically, we examined the use of LLM-based relevance judgments as a means to automate and scale the evaluation process. We proposed a comprehensive methodology to assess the extent to which relevance judgments generated by different LLMs align with human judgments. In addition, we investigated the sensitivity and robustness of LLM-based relevance assessments with respect to prompt design and the choice of LLM, and we analyzed how these factors influence system rankings. Our findings underscore that while LLMs offer promising opportunities for efficient and scalable evaluation, rigorous methodology is essential to ensure that they serve as a reliable complement to human assessment, rather than an inconsistent or unpredictable substitute.

Evaluation of Generative Information. In the final substantive chapter, this thesis broadened its scope to consider evaluation techniques for applications beyond traditional information-seeking tasks, encompassing a variety of tasks and modalities. We recognized that modern information access systems—often powered by large language models—are no longer limited to returning ranked lists of documents, but are instead capable of achieving more complex objectives. To address this shift, we first adapted traditional retrieval benchmarks to evaluate generated answers in open-domain question answering scenarios. We explored methods that allow both generated and retrieved responses to be assessed on the same footing, enabling a fair comparison between retrieval-based and generation-based approaches.

We then extended evaluation concepts to cross-modal generative tasks, specifically focusing on text-to-image generation systems. Although these systems lie outside the classical domain of information retrieval, we found that IR evaluation principles—such as precision-recall analogs and distributional metrics—remain highly relevant. This interdisciplinary application demonstrated that IR-inspired evaluation methodologies can inform the assessment of generative models across domains.

Finally, moving beyond isolated queries or single-turn outputs, we introduced a broader framework for agent-level evaluation of LLM-powered applications. We proposed evaluating such systems based on how effectively they fulfill an end-to-end task or user goal, rather than merely assessing the accuracy of individual responses. Our AgentEval framework automatically generates a set of evaluation criteria tailored to the task and employs LLMs to quantify performance along these dimensions. Crucially, it also incorporates mechanisms to validate the evaluation’s sensitivity and soundness. This part of the thesis showcased how

evaluation can become adaptive and dynamic, evolving alongside the systems it assesses and adjusting to emerging user needs and capabilities.

6.2 Broader Implications

The findings and methods developed in this work have broad implications for the field of information retrieval and the evaluation of information access systems, particularly as we enter the generative era.

Traditional IR evaluation has long relied on the Cranfield paradigm of test collections with fixed queries, documents, and relevance judgments. While we do not say that was bad but Our work reveals that this paradigm must be updated to remain viable. The evaluation challenges revealed in this thesis indicate that rather than treating relevance in a binary fashion, we can evaluate systems by how well they approximate the idealized distribution of relevance. This shift encourages researchers to design metrics that remain stable even as the set of judged documents is partial. It also implies that benchmark creators should consider alternative judging strategies (e.g. shallow pooling, targeted judgments) that complement these new metrics to reduce bias. Ultimately, by mitigating biases from incomplete data, we ensure that truly superior models are recognized for retrieving or generating relevant information, which is crucial for driving progress in the field.

Additionally, as IR systems increasingly integrate generative components—for example, search engines that respond to queries with synthesized answers from large language models—the historical divide between “retrieval” and “generation” in evaluation is rapidly diminishing. This thesis contributes to the development of a unified evaluation framework capable of fairly assessing both types of systems. Achieving this unification requires rethinking how ground truth data is constructed. Traditional relevance judgments, for instance, tend to favor retrieval-based systems by focusing on document-level annotations, which may not fully capture the quality or utility of generated content.

By enabling direct comparison between systems that return ranked lists of documents and those that produce synthesized answers, our proposed evaluation methods support more balanced assessments of which approach better satisfies an information need. This allows practitioners to make more informed decisions about when to deploy retrieval, generation, or retrieval-augmented generation (RAG) systems. Having fair, modality-agnostic evaluation not only facilitates progress in generative IR but also supports trade-offs between system effectiveness and efficiency. Looking ahead, evaluation-focused venues such as TREC may evolve to support tasks in which systems are free to retrieve or generate content, with unified metrics designed to compare their effectiveness on equal footing.

Furthermore, this thesis highlights that large language models are not only subjects of evaluation but also powerful tools for evaluation. Their ability to generate relevance judg-

ments enables rapid experimentation without the need to wait for human labels. However, using LLMs to judge the output of other LLMs raises concerns about circularity, where AI systems critique one another. Our findings suggest this is a manageable challenge. For instance, treating the evaluator LLM as a higher-capacity or more robust model (e.g., using a more advanced version or an ensemble-based approach) can mitigate risks. Still, we emphasize the continued necessity of human-grounded evaluation. Human assessors—whether involved in calibration, adjudication, or benchmark construction—are essential for maintaining alignment between evaluation metrics and actual human utility. The implication is a hybrid future, where evaluation methodologies combine the efficiency of LLM-based judgments with the credibility and grounding of human oversight.

Traditional IR metrics have focused on document relevance, often at the individual query level. However, the generative era demands a shift toward evaluating information utility—the degree to which a system fulfills a user’s actual goal. This thesis supports this transition through the development of task-level evaluation frameworks, which assess whether a system can successfully complete complex, multi-step tasks such as answering questions, aiding decisions, or producing structured outputs.

By reframing evaluation around user-centered goals, we encourage the development of systems that are not just accurate, but genuinely useful. Our work on criteria such as clarity and completeness further suggests that future evaluation schemes will benefit from multi-dimensional rating schemas, going beyond binary relevance to capture richer aspects of answer quality. This shift is critical for ensuring that evaluation frameworks remain aligned with the evolving capabilities and roles of modern information access systems.

6.3 Limitations

While this thesis advances several aspects of evaluation for information access systems in the generative era, it is important to acknowledge its limitations. These constraints highlight areas where future work can build upon and extend the contributions presented here.

First, despite efforts to include a range of datasets, the experimental scope remains largely limited to standard benchmarks such as TREC Deep Learning and MS MARCO. Although these collections are widely used and cover diverse query and document types, they may not fully capture the breadth of domains, user intents, or content modalities encountered in real-world applications. As a result, some conclusions—particularly regarding evaluation robustness or LLM judgment behavior—may not generalize beyond these benchmarks without further validation.

Second, the work is restricted to passage-level retrieval and evaluation, which imposes certain structural constraints. Many datasets rely on fixed token limits, and the experiments

assume that all relevant content can be captured within relatively short contexts (typically under a few hundred tokens). In practice, however, modern LLMs and IR systems increasingly operate over longer contexts and full documents, suggesting that future evaluation frameworks must scale accordingly.

Third, the experiments in this thesis involve a limited set of IR systems and LLMs. Although care was taken to include strong baselines and representative models, the findings—particularly regarding the impact of sparse labeling, LLM-based relevance judgments, and prompt sensitivity—should be interpreted in the context of the specific models tested. Broader validation across more diverse architectures, including multilingual and domain-specific systems, would be necessary to fully assess the generalizability of these conclusions.

Fourth, the robustness of certain conclusions has not been exhaustively tested. For example, while we report consistent trends across multiple runs in LLM-based evaluations, we do not provide a comprehensive sensitivity analysis across all configurations or sampling seeds. This limitation is especially relevant in experiments involving non-deterministic model outputs, such as open-ended generation and LLM-based relevance assessments.

Fifth, While in general, in this thesis we tend to propose novel evaluation metrics and frameworks, it focuses primarily on offline, system-centric metrics. Online evaluation, user-centric metrics such as satisfaction, trust, or long-term engagement were not directly measured, which limits the extent to which the evaluation frameworks reflect real-world user experiences.

Finally, although all code, data, and evaluation scripts used in this thesis have been made publicly available wherever possible, some experiments—particularly those involving LLM-based applications—may be challenging to reproduce exactly due to the inherent variability and proprietary nature of commercial LLMs (e.g., GPT-4). Model updates, rate limits, and undocumented changes in API behavior can all affect reproducibility over time. Moreover, prompt-based evaluations depend on subtle variations in instruction wording, model versioning, and context length, which may yield different results even when the same codebase is used.

By clearly acknowledging these limitations, we hope to provide transparency about the scope of this work and to support future research in addressing these open challenges.

6.4 Future Research Directions

While this thesis advances the evaluation of information access systems in the generative era, it also opens up several potential future directions for research. Below, we outline some of the key avenues this work paves the way for:

Enhancing and Generalizing Distribution-Based Metrics. The use of Fréchet Distance proved effective for comparing result distributions under sparse labeling conditions. Future research could build on this foundation by exploring alternative distribution similarity measures or by incorporating learned semantic representations to better capture relevance. In the era of generative information, ensuring that generated content aligns with the distribution of real or relevant content is particularly important. For this reason, distribution-based metrics may offer a more flexible and meaningful basis for evaluation than absolute comparisons alone. Notably, such metrics enable fair comparison between generated and retrieved content within a unified framework. However, an open challenge remains: how can we combine distribution-based metrics with traditional pointwise evaluation metrics (such as NDCG) to obtain the best of both worlds? Developing hybrid evaluation measures that are both robust to label sparsity and easily interpretable is a promising direction for future work.

Standardizing LLM-Based Evaluation Protocols. Our analysis revealed that LLM-based evaluators can produce highly variable judgments depending on the specific data points, prompts, or the choice of model. A critical direction for future work is the development of standardized protocols for LLM-based relevance assessment. This includes establishing agreed-upon prompt templates, calibration strategies (such as providing clear instructions or few-shot examples), and potentially releasing community-maintained LLM judge models fine-tuned specifically for evaluation tasks. By defining best practices and incorporating validation mechanisms, the field can move toward treating LLM-based judgments as a reliable and reproducible component of the evaluation pipeline. Shared task initiatives could also introduce dedicated tracks for automated evaluation, encouraging further innovation and fostering consensus on the use of LLMs as evaluators.

Evaluating Truthfulness and Hallucination in GenIR. Generative IR systems introduce unique challenges, particularly regarding hallucination—where models produce inaccurate or fabricated content. While traditional IR evaluation often assumed that retrieved documents contained authentic information, this assumption has been increasingly challenged—particularly by prior work on misinformation, such as the TREC Health Misinformation Track. However, the emergence of LLMs introduced a new class of issues, notably hallucinated content, which differs from earlier forms of inaccuracy and necessitates fresh evaluation strategies.

Future research should focus on developing metrics and methodologies that can assess the truthfulness and factual correctness of generated answers. Building on the nugget-based approach, one promising direction involves leveraging knowledge bases or verified reference collections to automatically validate statements within a generated response. Designing evaluation strategies that effectively penalize hallucinations while rewarding informative-

ness is essential for the responsible deployment of generative IR systems, and remains a rich area for further exploration.

User-Centric and Interaction-Based Evaluation. As systems evolve into interactive assistants—such as in conversational search or multi-turn dialogue—evaluation must account for the dynamic nature of user interaction. Similar to traditional IR metrics that rely on user modeling and implicit signals (e.g., click-through data), generative systems also require user modeling frameworks, but adapted to their specific interaction patterns and capabilities. Future work should explore evaluation methods that simulate user behavior over a session and assess how effectively systems respond, adapt, and guide users toward successful outcomes. Additionally, incorporating real user feedback, such as implicit signals or behavioral traces from system logs, can enrich evaluation strategies and make them more reflective of real-world usage.

Adaptive Evaluation Criteria for New Applications. The notion of adaptive evaluation—introduced through our AgentEval framework—can be significantly extended as new types of LLM applications emerge, including coding assistants, educational tutors, and creative writing tools. Each of these domains requires tailored evaluation criteria that reflect task-specific goals and success measures. Moreover, as these systems improve, evaluation itself must adapt to remain informative and discriminative. Future research should explore how evaluation frameworks can evolve alongside systems, providing continuous feedback for iterative development and refinement. An adaptive evaluation loop—where criteria evolve in response to system improvements—ensures that evaluation remains a guiding force in the advancement of LLM-powered applications.

By pursuing these and other related directions, researchers will continue to enrich the field’s understanding of evaluation in an ever-changing landscape. Each direction aims to ensure that as our information access systems advance, the means by which we evaluate them remain rigorous, relevant, and reflective of what truly matters to end users.

6.5 Closing Remarks

As information access systems continue to evolve—from traditional document retrieval engines to complex, LLM-powered agents—our evaluation methodologies must evolve in parallel. This thesis has aimed to address that challenge by proposing robust, adaptable frameworks that account for the increasing complexity, generative nature, and human-centered goals of modern systems. From detecting flaws in evaluation with sparse labels to leveraging large language models as evaluators and designing task-level assessments,

the contributions made here reflect a commitment to making evaluation not just an afterthought, but a first-class concern in system development. I strongly believe that as we build increasingly capable models, our evaluation strategies must be just as progressive, rigorous, and forward-thinking. Without meaningful and trustworthy evaluation, we risk mistaking scale for progress. By continuing to innovate in this space, we can ensure that information access technologies genuinely serve the needs of users.

References

- [1] Zahra Abbasiantaeb and Mohammad Aliannejadi. Generate then retrieve: Conversational response retrieval using llms as answer and query generators, 2024.
- [2] Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. Can we use large language models to fill relevance judgment holes?, 2024.
- [3] Heba Abdel-Nabi, Arafat Awajan, and Mostafa Z Ali. Deep learning-based question answering: a survey. *Knowledge and Information Systems*, 65(4):1399–1485, 2023.
- [4] Eneko Agirre, Giorgio Maria Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. Clef 2008: Ad hoc track overview. In Carol Peters, Thomas Deselaers, Nicola Ferro, Julio Gonzalo, Gareth J. F. Jones, Mikko Kurimo, Thomas Mandl, Anselmo Peñas, and Vivien Petras, editors, *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 15–37, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-04447-2.
- [5] JM Ah-Pine, CM Cifarelli, SM Clinchant, GM Csurka, and Jean-Michel Renders. Xrce’s participation to imageclef 2008. In *9th Workshop of the Cross-Language Evaluation Forum (CLEF 2008)*, 2008.
- [6] Ali Ahmadvand, Negar Arabzadeh, Julia Kiseleva, Patricio Figueroa Sanz, Xin Deng, Sujay Jauhar, Michael Gamon, Eugene Agichtein, Ned Friend, et al. Supporting complex information-seeking tasks with implicit constraints. *arXiv preprint arXiv:2205.00584*, 2022.
- [7] Azzah Al-Maskari and Mark Sanderson. A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology*, 61(5):859–868, 2010.
- [8] Marwah Alaofi, Luke Gallagher, Dana McKay, Lauren L. Saling, Mark Sanderson, Falk Scholer, Damiano Spina, and Ryen W. White. Where do queries come from?

- In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2850–2862. ACM, 2022.
- [9] Marwah Alaofi, Negar Arabzadeh, Charles LA Clarke, and Mark Sanderson. Generative information retrieval evaluation. *arXiv preprint arXiv:2404.08137*, 2024.
 - [10] Marwah Alaofi, Paul Thomas, Falk Scholer, and Mark Sanderson. Llms can be fooled into labelling a document as relevant: best café near me; this paper is perfectly relevant. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, page 32–41, 2024.
 - [11] Eloi Alonso, Bastien Moysset, and Ronaldo Messina. Adversarial generation of handwritten text images conditioned on sequences. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 481–486. IEEE, 2019.
 - [12] Helmut Alt. The computational geometry of comparing shapes. *Efficient Algorithms: Essays Dedicated to Kurt Mehlhorn on the Occasion of His 60th Birthday*, pages 235–248, 2009.
 - [13] Helmut Alt and Maike Buchin. Can we compute the similarity between surfaces? *CoRR*, abs/cs/0703011, 2007.
 - [14] Helmut Alt and Michael Godau. Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5 (01n02):75–91, 1995.
 - [15] Helmut Alt, Christian Knauer, and Carola Wenk. Matching polygonal curves with respect to the fréchet distance. In *STACS 2001: 18th Annual Symposium on Theoretical Aspects of Computer Science Dresden, Germany, February 15–17, 2001 Proceedings 18*, pages 63–74. Springer, 2001.
 - [16] Enrique Amigó, Jorge Carrillo-de Albornoz, Mario Almagro-Cádiz, Julio Gonzalo, Javier Rodríguez-Vidal, and Felisa Verdejo. Evall: Open access evaluation for information access systems. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’17, page 1301–1304, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350228.

- [17] Maryam Amirizani, Jihan Yao, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. Developing a framework for auditing large language models using human-in-the-loop. *arXiv preprint arXiv:2402.09346*, 2024.
- [18] Albert Andrew. Potential applications and implications of large language models in primary care. *Family Medicine and Community Health*, 12(Suppl 1), 2024.
- [19] Negar Arabzadeh and Charles LA Clarke. A comparison of methods for evaluating generative ir. *arXiv preprint arXiv:2404.04044*, 2024.
- [20] Negar Arabzadeh and Charles LA Clarke. Fréchet distance for offline evaluation of information retrieval systems with sparse labels. *arXiv preprint arXiv:2401.17543*, 2024.
- [21] Negar Arabzadeh, Xinyi Yan, and Charles L. A. Clarke. Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection. In Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong, editors, *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 2862–2866. ACM, 2021.
- [22] Negar Arabzadeh, Ali Ahmadvand, Julia Kiseleva, Yang Liu, Ahmed Hassan Awadallah, Ming Zhong, and Milad Shokouhi. Preme: Preference-based meeting exploration through an interactive questionnaire. *arXiv preprint arXiv:2205.02370*, 2022.
- [23] Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles LA Clarke. Shallow pooling for sparse labels. *Information Retrieval Journal*, 25(4):365–385, 2022.
- [24] Negar Arabzadeh, Amin Bigdeli, Radin Hamidi Rad, and Ebrahim Bagheri. Quantifying ranker coverage of different query subspaces. In *46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2298–2302, 2023.
- [25] Negar Arabzadeh, Oleksandra Kmet, Ben Carterette, Charles LA Clarke, Claudia Hauff, and Praveen Chandar. A is for adele: An offline evaluation metric for instant search. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 3–12, 2023.
- [26] Negar Arabzadeh, Amin Bigdeli, and Charles LA Clarke. Adapting standard retrieval benchmarks to evaluate generated answers. In *European Conference on Information Retrieval*, pages 399–414. Springer, 2024.

- [27] Negar Arabzadeh, Siqing Huo, Nikhil Mehta, Qingyun Wu, Chi Wang, Ahmed Hassan Awadallah, Charles L. A. Clarke, and Julia Kiseleva. Assessing and verifying task utility in LLM-powered applications. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21868–21888, Miami, Florida, USA, November 2024. Association for Computational Linguistics.
- [28] Negar Arabzadeh, Julia Kiseleva, Qingyun Wu, Chi Wang, Ahmed Awadallah, Victor Dibia, Adam Fourney, and Charles Clarke. Towards better human-agent alignment: Assessing task utility in llm-powered applications. *arXiv preprint arXiv:2402.09015*, 2024.
- [29] Ioannis Arapakis, Xiao Bai, and Berkant Barla Cambazoglu. Impact of response latency on user behavior in web search. In Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin, editors, *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 103–112. ACM, 2014.
- [30] Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. Ask me anything: A simple strategy for prompting language models, 2022.
- [31] Javed A Aslam and Robert Savell. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 361–362, 2003.
- [32] Javed A Aslam, Virgil Pavlu, and Emine Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 541–548, 2006.
- [33] Leif Azzopardi, Paul Thomas, and Nick Craswell. Measuring the utility of search engine result pages: an information foraging based measure. In *The 41st International ACM SIGIR conference on research & development in information retrieval*, pages 605–614, 2018.
- [34] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

- [35] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. Relevance assessment: are judges exchangeable and does it matter. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, pages 667–674. ACM, 2008.
- [36] Peter Bailey, Nick Craswell, Ryen W. White, Liwei Chen, Ashwin Satyanarayana, and Seyed M. M. Tahaghoghi. Evaluating whole-page relevance. In Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy, editors, *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*, pages 767–768. ACM, 2010.
- [37] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. Retrieval consistency in the presence of query variations. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 395–404. ACM, 2017.
- [38] Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20041–20053, 2023.
- [39] Krisztian Balog and ChengXiang Zhai. User simulation for evaluating information access systems, 2023. arXiv:2306.08550.
- [40] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.
- [41] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023.
- [42] Muneera Bano, Didar Zowghi, and Jon Whittle. Exploring qualitative research using llms. *arXiv preprint arXiv:2306.13298*, 2023.
- [43] Shane Barratt and Rishi Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018.

- [44] Anja Belz and Ehud Reiter. Comparing automatic and human evaluation of nlg systems. In *11th conference of the european chapter of the association for computational linguistics*, pages 313–320, 2006.
- [45] Yaniv Bernstein and Justin Zobel. Redundant documents and search effectiveness. In Otthein Herzog, Hans-Jörg Schek, Norbert Fuhr, Abdur Chowdhury, and Wilfried Teiken, editors, *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 - November 5, 2005*, pages 736–743. ACM, 2005.
- [46] Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*, 2023.
- [47] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.
- [48] Dania Bilal and Jacek Gwizdka. Children’s query types and reformulations in google search. *Information Processing & Management (IP&M)*, 54(6):1022–1041, 2018. ISSN 0306-4573.
- [49] Miłkołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021.
- [50] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. Inpars: Data augmentation for information retrieval using large language models, 2022.
- [51] Alexey Borisov, Julia Kiseleva, Ilya Markov, and Maarten de Rijke. Calibration: A simple way to improve click models. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1503–1506, 2018.
- [52] Ali Borji. Pros and cons of gan evaluation measures: New developments, 2021.
- [53] Pia Borlund and Peter Ingwersen. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 53(3):225–250, 1997.
- [54] Horatiu S. Bota, Ke Zhou, and Joemon M. Jose. Playing your cards right: The effect of entity cards on search behaviour and workload. In Diane Kelly, Robert Capra, Nicholas J. Belkin, Jaime Teevan, and Pertti Vakkari, editors, *Proceedings of*

- the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13-17, 2016*, pages 131–140. ACM, 2016.
- [55] Andrei Broder. A taxonomy of Web search. *SIGIR Forum*, 36(2):3–10, 2002.
 - [56] Sebastian Bruch, Shuguang Han, Mike Bendersky, and Marc Najork. A stochastic treatment of learning to rank scoring functions. In *Proceedings of the 13th ACM International Conference on Web Search and Data Mining (WSDM 2020)*, 2020.
 - [57] Chris Buckley and Ellen M Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, 2004.
 - [58] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. Bias and the limits of pooling. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’06, page 619–620, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933697.
 - [59] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *22nd International Conference on Machine Learning*, pages 89–96, Bonn, Germany, August 2005.
 - [60] Stefan Büttcher and Charles L. A. Clarke. Efficiency vs. effectiveness in terabyte-scale information retrieval. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Fourteenth Text REtrieval Conference, TREC 2005, Gaithersburg, Maryland, USA, November 15-18, 2005*, volume 500-266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2005.
 - [61] Stefan Büttcher, Charles LA Clarke, Peter CK Yeung, and Ian Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 63–70, 2007.
 - [62] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand. What do different evaluation metrics tell us about saliency models? *IEEE TPAMI*, 41(3):740–757, 2018.
 - [63] Yinqiong Cai, Jiafeng Guo, Yixing Fan, Qingyao Ai, Ruqing Zhang, and Xueqi Cheng. Hard negatives or false negatives: Correcting pooling bias in training neural ranking models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 118–127, 2022.

- [64] Huanhuan Cao, Daxin Jiang, Jian Pei, Qi He, Zhen Liao, Enhong Chen, and Hang Li. Context-aware query suggestion by mining click-through and session data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 875–883, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934.
- [65] Ben Carterette. System effectiveness, user models, and user utility: a conceptual framework for investigation. In Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft, editors, *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval*, pages 903–912. ACM, 2011.
- [66] Ben Carterette and James Allan. Semiautomatic evaluation of retrieval systems using document similarities. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 873–876, New York, NY, USA, 2007. ACM.
- [67] Ben Carterette and Paul N. Bennett. Evaluation measures for preference judgments. In *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–686, Singapore, 2008.
- [68] Ben Carterette and Mark D Smucker. Hypothesis testing with incomplete relevance judgments. In *Proceedings of the sixteenth ACM conference on conference on information and knowledge management*, pages 643–652, 2007.
- [69] Ben Carterette, Paul Bennett, and Olivier Chapelle. A test collection of preference judgments. In *SIGIR 2008 Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments*, Singapore, July 2008.
- [70] Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. Here or there: Preference judgments for relevance. Computer Science Department Faculty Publication Series 46, University of Massachusetts Amherst, 2008.
- [71] Inês Carvalho and Stanislav Ivanov. Chatgpt for tourism: applications, benefits and risks. *Tourism Review*, 2023.
- [72] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*, 2020.
- [73] Deepayan Chakrabarti, Ravi Kumar, and Kunal Punera. Quicklink selection for navigational query results. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors, *Proceedings of the 18th International Conference on World*

- Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 391–400. ACM, 2009.
- [74] Souradeep Chakraborty, Zijun Wei, Conor Kelton, Seoyoung Ahn, Aruna Balasubramanian, Gregory J. Zelinsky, and Dimitris Samaras. Predicting visual attention in graphic design documents. *IEEE Transactions on Multimedia*, pages 1–1, 2022.
 - [75] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*, 2023.
 - [76] CS Richard Chan, Charuta Pethe, and Steven Skiena. Natural language processing versus rule-based text analysis: Comparing bert score and readability indices to predict crowdfunding outcomes. *Journal of Business Venturing Insights*, 16:e00276, 2021.
 - [77] Praveen Chandar, Fernando Diaz, and Brian St. Thomas. Beyond accuracy: Grounding evaluation metrics for human-machine learning systems. In *Advances in Neural Information Processing Systems*, 2020.
 - [78] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2023.
 - [79] Olivier Chapelle and Yi Chang. Yahoo! Learning to Rank Challenge overview. *Proceedings of Machine Learning Research*, 14:1–24, June 2011.
 - [80] Olivier Chapelle and Ya Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*, pages 1–10, 2009.
 - [81] Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 621–630, 2009.
 - [82] Anwoy Chatterjee, HSVNS Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. Posix: A prompt sensitivity index for large language models. *arXiv preprint arXiv:2410.02185*, 2024.
 - [83] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. Continual learning for generative retrieval over dynamic

- corpora. In Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos, editors, *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 306–315. ACM, 2023.
- [84] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yiqun Liu, Yixing Fan, and Xueqi Cheng. A unified generative retriever for knowledge- intensive language tasks via prompt learning. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1448–1457. ACM, 2023.
 - [85] Wenhui Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.
 - [86] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
 - [87] Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *arXiv preprint arXiv:2401.16788*, 2024.
 - [88] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.
 - [89] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.
 - [90] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.
 - [91] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054, 2023.
 - [92] Min Jin Chong and David A. Forsyth. Effectively unbiased FID and inception score and where to find them. *CoRR*, abs/1911.07023, 2019.

- [93] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.
- [94] Peng Chu, Anita Komlodi, and Gyöngyi Rózsa. Online search in english as a non-native language. In *Information Science with Impact: Research in and for the Community - Proceedings of the 78th ASIS&T Annual Meeting, ASIST 2015, St. Louis, Missouri, Missouri, USA, October 6-10, 2015*, volume 52, pages 1–9. Wiley, 2015.
- [95] Karen Church, Barry Smyth, Paul Cotter, and Keith Bradley. Mobile information access: A study of emerging search behavior on the mobile internet. *ACM Transactions on the Web (TWEB)*, 1(1):4, 2007.
- [96] Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. Sentence mover’s similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy, July 2019. Association for Computational Linguistics.
- [97] Charles L Clarke, Nick Craswell, and Ian Soboroff. Overview of the trec 2009 web track. Technical report, WATERLOO UNIV (ONTARIO), 2009.
- [98] Charles L. A. Clarke and Laura Dietz. Llm-based relevance assessment still can’t replace human relevance assessment, 2024.
- [99] Charles L. A. Clarke, Eugene Agichtein, Susan T. Dumais, and Ryen W. White. The influence of caption features on clickthrough patterns in web search. In Wessel Kraaij, Arjen P. de Vries, Charles L. A. Clarke, Norbert Fuhr, and Noriko Kando, editors, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 135–142. ACM, 2007.
- [100] Charles L. A. Clarke, J. Shane Culpepper, and Alistair Moffat. Assessing efficiency–effectiveness tradeoffs in multi-stage retrieval systems without using relevance judgments. *Information Retrieval*, 19(4):351–377, aug 2016. ISSN 1386-4564.
- [101] Charles L. A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. Offline evaluation by maximum similarity to an ideal ranking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM ’20*, page 225–234, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368599.
- [102] Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. Assessing top- k preferences. *ACM Transactions on Information Systems*, 39(3), July 2021.

- [103] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 659–666, 2008.
- [104] Charles LA Clarke, Alexandra Vtyurina, and Mark D Smucker. Offline evaluation without gain. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pages 185–192, 2020.
- [105] Cyril W Cleverdon. The significance of the cranfield tests on index languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, 1991.
- [106] Paul Clough and Mark Sanderson. Evaluating the performance of information retrieval systems using test collections. 2013.
- [107] Alexis Conneau and Douwe Kiela. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*, 2018.
- [108] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 282–289, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130155.
- [109] Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, pages 41–75. Springer, 2019.
- [110] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In Marc Najork, Andrei Z. Broder, and Soumen Chakrabarti, editors, *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94. ACM, 2008.
- [111] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the TREC 2020 Deep Learning Track. In *29th Text REtrieval Conference*, Gaithersburg, Maryland, November 2020.

- [112] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*, 2020.
- [113] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the trec 2020 deep learning track, 2021.
- [114] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. MS MARCO: Benchmarking ranking models in the large-data regime. In *44th Annual International ACM SIGIR Conference on Research and Development in Infomation Retrieval*, pages 1566–1576, Montreal, Canada, July 2021.
- [115] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. Overview of the trec 2021 deep learning track. In *Text REtrieval Conference (TREC)*. NIST, TREC, May 2022.
- [116] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. Overview of the TREC 2022 deep learning track. In Ian Soboroff and Angela Ellis, editors, *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022*, volume 500-338 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2022.
- [117] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Hossein A. Rahmani, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. Overview of the TREC 2023 deep learning track. In *Text REtrieval Conference*, Gaithersburg, Maryland, 2023.
- [118] W. Bruce Croft and David J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, 1979.
- [119] J. Shane Culpepper, Guglielmo Faggioli, Nicola Ferro, and Oren Kurland. Topic difficulty: Collection and query formulation effects. *ACM Transactions on Information Systems (TOIS)*, 40(1):19:1–19:36, 2022.
- [120] Zhuyun Dai and Jamie Callan. Context-aware sentence/passage term importance estimation for first stage retrieval. *arXiv preprint arXiv:1910.10687*, 2019.
- [121] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. How to prompt? opportunities and challenges of zero-and few-shot learning for human-ai interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*, 2022.
- [122] Tirth Dave, Sai Anirudh Athaluri, and Satyam Singh. Chatgpt in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*, 6:1169595, 2023.

- [123] Niklas Deckers, Maik Fröbe, Johannes Kiesel, Gianluca Pandolfo, Christopher Schröder, Benno Stein, and Martin Potthast. The infinite index: Information retrieval on generative text-to-image models. *arXiv preprint arXiv:2212.07476*, 2022.
- [124] Fernando Diaz, Ryen W. White, Georg Buscher, and Dan Liebling. Robust models of mouse movement on dynamic web search results pages. In *Proceedings of the 22nd ACM conference on Information and knowledge management (CIKM 2013)*, pages 1451–1460, New York, NY, USA, 2013. Association for Computing Machinery.
- [125] Victor Dibia, Adam Fourney, Gagan Bansal, Forough Poursabzi-Sangdeh, Han Liu, and Saleema Amershi. Aligning offline metrics and human judgments of value for code generation models, 2023.
- [126] Panagiotis Dimitrakopoulos, Giorgos Sfikas, and Christophoros Nikou. Wind: Wasserstein inception distance for evaluating generative adversarial network performance. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3182–3186. IEEE, 2020.
- [127] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34: 19822–19835, 2021.
- [128] Bradley Efron. Second thoughts on the bootstrap. *Statistical science*, pages 135–140, 2003.
- [129] Thomas Eiter and Heikki Mannila. Computing discrete fréchet distance. 1994.
- [130] Björn Engelmann, Timo Breuer, Jana Isabelle Frieze, Philipp Schaer, and Norbert Fuhr. Context-driven interactive query simulations based on generative large language models. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II*, volume 14609 of *Lecture Notes in Computer Science*, pages 173–188. Springer, 2024.
- [131] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR '23*, page 39–50. ACM, August 2023. doi: 10.1145/3578337.3605136. URL <http://dx.doi.org/10.1145/3578337.3605136>.

- [132] Naghmeh Farzi and Laura Dietz. An exam-based evaluation approach beyond traditional relevance judgments, 2024.
- [133] Naghmeh Farzi and Laura Dietz. Pencils down! automatic rubric-based evaluation of retrieve/generate systems. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 175–184, 2024.
- [134] Zhangyin Feng, Weitao Ma, Weijiang Yu, Lei Huang, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications. *arXiv preprint arXiv:2311.05876*, 2023.
- [135] Marco Ferrante, Nicola Ferro, and Norbert Fuhr. Towards meaningful statements in IR evaluation. Mapping evaluation measures to interval scales. *IEEE Access*, 9: 136182–136216, September 2021.
- [136] Nicola Ferro and Maria Maistro. Evaluation of ir systems. In *Information Retrieval: Advanced Topics and Techniques*, pages 111–191. 2024.
- [137] Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320, 2016.
- [138] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. Splade v2: Sparse lexical and expansion model for information retrieval, 2021.
- [139] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, 2021.
- [140] Luke Friedman, Sameer Ahuja, David Allen, Terry Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*, 2023.
- [141] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel. Adversarial text-to-image synthesis: A review. *Neural Networks*, 144:187–209, 2021.
- [142] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*, 2023.
- [143] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, pages 89–106, 2022.

- [144] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. *Neural approaches to conversational information retrieval*, volume 44. Springer, 2023.
- [145] Yiqi Gao, Xinglin Hou, Yuanmeng Zhang, Tiezheng Ge, Yuning Jiang, and Peng Wang. Caponimage: Context-driven dense-captioning on image. *arXiv preprint arXiv:2204.12974*, 2022.
- [146] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023.
- [147] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.
- [148] Amin Gheibi, Anil Maheshwari, Jörg-Rüdiger Sack, and Christian Scheffer. Minimum backward fréchet distance. In *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 381–388, 2014.
- [149] Lukas Gienapp, Harris Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. Evaluating generative ad hoc information retrieval. *CoRR*, abs/2311.04694, 2023. *arXiv preprint*.
- [150] Michael Gordon and Praveen Pathak. Finding information on the world wide web: the retrieval effectiveness of search engines. *Information processing & management*, 35(2):141–180, 1999.
- [151] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.
- [152] J. P. Guilford. Creativity. *American Psychologist*, 5(9):444–454, 1950.
- [153] Shengbo Guo and Scott Sanner. Probabilistic latent maximal marginal relevance. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 833–834, 2010.
- [154] Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*, 2023.

- [155] Kevin Haas, Peter Mika, Paul Tarjan, and Roi Blanco. Enhanced results for web search. In Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft, editors, *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 725–734. ACM, 2011.
- [156] Jutta Haider and Olof Sundin. *Invisible search and online search engines: The ubiquity of search in everyday life*. Taylor & Francis, 2019.
- [157] Kailash A Hambarde and Hugo Proenca. Information retrieval: recent advances and beyond. *IEEE Access*, 11:76581–76604, 2023.
- [158] Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. Learning-to-rank with BERT in TF-ranking. *CoRR*, abs/2004.08476, 2020.
- [159] Michael Hanna and Ondřej Bojar. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, 2021.
- [160] Donna Harman. User-friendly systems instead of user-friendly front-ends. *Journal of the American Society for Information Science (JASIST)*, 43(2):164–174, 1992.
- [161] Donna K Harman. *The first text retrieval conference (TREC-1)*, volume 500. US Department of Commerce, National Institute of Standards and Technology, 1993.
- [162] GM Harshvardhan, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285, 2020.
- [163] Morgan Harvey and Matthew Pointon. Searching on the go: The effects of fragmented attention on mobile web search tasks. In Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White, editors, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 155–164. ACM, 2017.
- [164] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. Antique: A non-factoid question answering benchmark. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 166–173. Springer, 2020.
- [165] Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*, 2019.

- [166] Helena Häußler. Do users trust search engines? and if so, why? developing a trust measure and applying it in an experiment. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 471–474, 2023.
- [167] David Hawking, Ellen Voorhees, Nick Craswell, Peter Bailey, et al. Overview of the trec-8 web track. In *TREC*, 1999.
- [168] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
- [169] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- [170] William Hersh. Search still matters: information retrieval in the era of generative ai. *Journal of the American Medical Informatics Association*, 31(9):2159–2161, 2024.
- [171] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [172] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1552–1565, 2020.
- [173] Sebastian Hofstätter, Navid Rekabsaz, Carsten Eickhoff, and Allan Hanbury. On the effect of low-frequency terms on neural-IR models. In *42nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1137–1140, Paris, France, July 2019.
- [174] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. *CoRR*, abs/2010.02666, 2020.
- [175] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.
- [176] David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. Twenty years of confusion in human evaluation: Nlg needs evaluation

- sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, pages 169–182. Association for Computational Linguistics, 2020.
- [177] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023.
 - [178] Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*, 2023.
 - [179] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.
 - [180] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.
 - [181] Siqing Huo, Negar Arabzadeh, and Charles LA Clarke. Retrieving supporting evidence for llms generated answers. *arXiv preprint arXiv:2306.13781*, 2023.
 - [182] Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. Multi-dimensional evaluation of text summarization with in-context learning. *arXiv preprint arXiv:2306.01200*, 2023.
 - [183] Youngseung Jeon, Seungwan Jin, Patrick C. Shih, and Kyungsik Han. Fashionq: An ai-driven creativity support tool for facilitating ideation in fashion design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI ’21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966.
 - [184] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.
 - [185] Minghui Jiang, Ying Xu, and Binhai Zhu. Protein structure–structure alignment with discrete fréchet distance. *Journal of bioinformatics and computational biology*, 6(01):51–64, 2008.

- [186] Xue Jiang, Yihong Dong, Lecheng Wang, Qiwei Shang, and Ge Li. Self-planning code generation with large language model. *arXiv preprint arXiv:2303.06689*, 2023.
- [187] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.
- [188] Thorsten Joachims and Filip Radlinski. Search engines that learn from implicit feedback. *IEEE Computer*, 40(8):34–40, August 2007.
- [189] Roger W Johnson. An introduction to the bootstrap. *Teaching statistics*, 23(2):49–54, 2001.
- [190] K Sparck Jones. A look back and a look forward. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 13–29, 1988.
- [191] Karen Sparck Jones and Julia R Galliers. Evaluating natural language processing systems: An analysis and review. 1995.
- [192] Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. Evaluating open-domain question answering in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [193] Chandresh S Kanani, Sriparna Saha, and Pushpak Bhattacharyya. Improving diversity and reducing redundancy in paragraph captions. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.
- [194] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Souichiro Hidaka. Overview of ir tasks at the first ntcir workshop. In *Proceedings of the first NTCIR workshop on research in Japanese text retrieval and term recognition*, pages 11–44, 1999.
- [195] Pegah Karimi, Jeba Rezwana, Safat Siddiqui, Mary Lou Maher, and Nasrin Dehbzorgi. Creative sketching partner: An analysis of human-ai co-creativity. In *Proceedings of the 25th International Conference on Intelligent User Interfaces, UII '20*, page 221–230, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371186.

- [196] Gabriella Kazai, Emine Yilmaz, Nick Craswell, and S.M.M. Tahaghoghi. User intent and assessor disagreement in web search evaluation. In *22nd ACM International Conference on Information and Knowledge Management*, pages 699–708, San Francisco, California, 2013.
- [197] Jaana Kekäläinen. Binary and graded relevance in ir evaluations—comparison of the effects on ranking of ir systems. *Information processing & management*, 41(5): 1019–1033, 2005.
- [198] Jaana Kekäläinen and Kalervo Järvelin. Using graded relevance assessments in ir evaluation. *Journal of the American Society for Information Science and Technology*, 53(13):1120–1129, 2002.
- [199] Andruid Kerne, Andrew M. Webb, Steven M. Smith, Rhema Linder, Nic Lupfer, Yin Qu, Jon Moeller, and Sashikanth Damaraju. Using metrics of curation to evaluate information-based ideation. *ACM Transactions on Computer-Human Interaction*, 21(3), June 2014.
- [200] Muhammad Zeeshan Khan, Saira Jabeen, Muhammad Usman Ghani Khan, Tanzila Saba, Asim Rehmat, Amjad Rehman, and Usman Tariq. A realistic image generation of face from text description using the fully trained generative adversarial networks. *IEEE Access*, 9:1250–1260, 2020.
- [201] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, Xi’an, China, July 2020.
- [202] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.
- [203] Julia Kiseleva and Maarten de Rijke. Evaluating personal assistants on mobile devices. *arXiv preprint arXiv:1706.04524*, 2017.
- [204] Julia Kiseleva, Eric Crestan, Riccardo Brigo, and Roland Dittel. Modelling and detecting changes in user satisfaction. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1449–1458, 2014.
- [205] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. Predicting user satisfaction with intelligent assistants.

In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 45–54, 2016.

- [206] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 121–130, 2016.
- [207] Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, Arthur Szlam, Yuxuan Sun, Katja Hofmann, Marc-Alexandre Côté, Ahmed Awadallah, Linar Abdrazakov, Igor Churin, Putra Manggala, Kata Naszadi, Michiel van der Meer, and Taewoon Kim. Interactive grounded language understanding in a collaborative environment: Iglu 2021. In Douwe Kiela, Marco Ciccone, and Barbara Caputo, editors, *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176 of *Proceedings of Machine Learning Research*, pages 146–161. PMLR, 06–14 Dec 2022.
- [208] Julia Kiseleva, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, Maartje ter Hoeve, Zoya Volovikova, Aleksandr Panov, Yuxuan Sun, Kavya Srinet, Arthur Szlam, Ahmed Awadallah, Seungeun Rho, Taehwan Kwon, Daniel Wontae Nam, Felipe Bivort Haiek, Edwin Zhang, Linar Abdrazakov, Guo Qingyam, Jason Zhang, and Zhibin Guo. Interactive grounded language understanding in a collaborative environment: Retrospective on iglu 2022 competition. In Marco Ciccone, Gustavo Stolovitzky, and Jacob Albrecht, editors, *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 204–216. PMLR, 28 Nov–09 Dec 2022.
- [209] Brent Kitchens, Steven L Johnson, and Peter Gray. Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS quarterly*, 44(4), 2020.
- [210] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists’ creative works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI ’23*, page 919–933, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701061.
- [211] Janin Koch, Andrés Lucero, Lena Hegemann, and Antti Oulasvirta. May ai? design ideation with cooperative contextual bandits. In *Proceedings of the 2019 CHI Con-*

- ference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702.
- [212] Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Text-to-image generation grounded by fine-grained user attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 237–246, 2021.
 - [213] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 2024.
 - [214] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.
 - [215] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fréchet inception distance. In *Proc. ICLR*, 2023.
 - [216] Philippe Laban, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. Near-negative distinction: Giving a second life to human evaluation datasets, 2022. URL <https://arxiv.org/abs/2205.06871>.
 - [217] Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. The language of prompting: What linguistic properties make a prompt successful?, 2023.
 - [218] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, 13(3):106–131, 2012.
 - [219] Mike Lewis and Angela Fan. Generative question answering: Learning to answer the whole question. In *International Conference on Learning Representations*, 2018.
 - [220] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
 - [221] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*, 2024.

- [222] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for” mind” exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*, 2023.
- [223] Quintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. *arXiv preprint arXiv:2310.19740*, 2023.
- [224] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 2024.
- [225] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.
- [226] Ziming Li, Julia Kiseleva, Alekh Agarwal, Maarten de Rijke, and Ryen W White. Optimizing interactive systems via data-driven objectives. *arXiv preprint arXiv:2006.12999*, 2020.
- [227] Ziming Li, Dookun Park, Julia Kiseleva, Young-Bum Kim, and Sungjin Lee. Deus: A data-driven approach to estimate user satisfaction in multi-turn dialogues. *arXiv preprint arXiv:2103.01287*, 2021.
- [228] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.
- [229] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023.

- [230] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate, 2023.
- [231] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18187–18196, 2022.
- [232] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [233] Chunmian Lin, Lin Li, Wenting Luo, Kelvin CP Wang, and Jiangang Guo. Transfer learning based traffic sign recognition using inception-v3 model. *Periodica Polytechnica Transportation Engineering*, 47(3):242–250, 2019.
- [234] Jimmy Lin and Xueguang Ma. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques. *arXiv preprint arXiv:2106.14807*, 2021.
- [235] Jimmy Lin, Rodrigo Frassetto Nogueira, and Andrew Yates. Pretrained transformers for text ranking: BERT and beyond. *CoRR*, abs/2010.06467, 2020.
- [236] Jimmy Lin, Daniel Campos, Nick Craswell, Bhaskar Mitra, and Emine Yilmaz. Significant improvements over the state of the art? A case study of the MS MARCO document ranking leaderboard. In *44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2283–2287, Montreal, Canada, July 2021.
- [237] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362, 2021.
- [238] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature, 2022.
- [239] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. In-batch negatives for knowledge distillation with tightly-coupled teachers for dense retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 163–173, 2021.

- [240] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.
- [241] Alex Liu and Min Sun. From voices to validity: Leveraging large language models (llms) for textual analysis of policy stakeholder interviews. *arXiv preprint arXiv:2312.01202*, 2023.
- [242] Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, and Li Zhang. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*, 2024.
- [243] Shaohui Liu, Yi Wei, Jiwen Lu, and Jie Zhou. An improved evaluation framework for generative adversarial networks. *arXiv preprint arXiv:1803.07474*, 2018.
- [244] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.
- [245] Yanchen Liu, Timo Schick, and Hinrich Schütze. Semantic-oriented unlabeled priming for large-scale language models. *arXiv preprint arXiv:2202.06133*, 2022.
- [246] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.
- [247] Zhenghao Liu, Kaitao Zhang, Chenyan Xiong, Zhiyuan Liu, and Maosong Sun. Open-Match: An open source library for Neu-IR research. *CoRR*, abs/2102.00166, 2021.
- [248] Chi-kiu Lo. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the Second Conference on Machine Translation*, pages 589–597, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
- [249] Chi-kiu Lo. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy, August 2019. Association for Computational Linguistics.
- [250] Sheng Lu, Hendrik Schuff, and Iryna Gurevych. How are prompts different in terms of sensitivity? In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5833–5856, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

- [251] Chengxi Luo, Charles L. A. Clarke, and Mark Smucker. Evaluation measures based on preference graphs. In *44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1534–1543, Montreal, Canada, July 2021.
- [252] Lorenzo Luzi, Carlos Ortiz Marrero, Nile WYNAR, Richard G Baraniuk, and Michael J Henry. Evaluating generative networks using gaussian mixtures of image features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 279–288, 2023.
- [253] Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*, 2023.
- [254] Sean MacAvaney and Luca Soldaini. One-shot labeling for automatic relevance estimation. In *SIGIR*, page 2230–2235, 2023.
- [255] Joel M. Mackenzie, Rodger Benham, Matthias Petri, Johanne R. Trippas, J. Shane Culpepper, and Alistair Moffat. Cc-news-en: A large english news corpus. In Mathieu d’Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 3077–3084. ACM, 2020.
- [256] Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. Generative relevance feedback with large language models. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2026–2031. ACM, 2023.
- [257] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. On crowd-sourcing relevance magnitudes for information retrieval evaluation. *ACM Transactions on Information Systems*, 35(3), January 2017.
- [258] Loic Maisonnasse, Philippe Mulhem, Eric Gaussier, and Jean Pierre Chevallet. Lig at imageclef 2008. In *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, Revised Selected Papers*, pages 704–711. Springer, 2009.
- [259] Mireille Makary, Michael Oakes, and Fadi Yamout. Towards automatic generation of relevance judgments for a test collection. In *ICDIM*, pages 121–126. IEEE, 2016.

- [260] Mireille Makary, Michael Oakes, Ruslan Mitkov, and Fadi Yammout. Using supervised machine learning to automatically build relevance judgments for a test collection. In *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 108–112. IEEE, 2017.
- [261] Maarit Mäkinen. Digital empowerment as a process for enhancing citizens’ participation. *E-learning and Digital Media*, 3(3):381–395, 2006.
- [262] Christopher D Manning. *An introduction to information retrieval*. 2009.
- [263] Gregory Marton. Nuggeteer: Automatic nugget-based evaluation using descriptions and judgements. 2006.
- [264] David Maxwell and Leif Azzopardi. Stuck in traffic: how temporal delays affect search behaviour. In David Elsweiler, Bernd Ludwig, Leif Azzopardi, and Max L. Wilson, editors, *Fifth Information Interaction in Context Symposium, IIX ’14, Regensburg, Germany, August 26-29, 2014*, pages 155–164. ACM, 2014.
- [265] Timothy R. McIntosh, Teo Susnjak, Nalin Arachchilage, Tong Liu, Paul Watters, and Malka N. Halgamuge. Inadequacies of large language model benchmarks in the era of generative artificial intelligence, 2024.
- [266] Michael McTear and Marina Ashurkina. *Transforming conversational AI: Exploring the power of large language models in interactive conversational agents*. Springer Nature, 2024.
- [267] Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, and Maarten de Rijke. Query performance prediction using relevance judgments generated by large language models. *arXiv preprint arXiv:2404.01012*, 2024.
- [268] Donald Metzler, Trevor Strohman, and W.Bruce Croft. Indri TREC notebook 2006: Lessons learned from three Terabyte Tracks. In *15th Text REtrieval Conference*, Gaithersburg, Maryland, November 2006.
- [269] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*, 2023.
- [270] Tomás Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.
- [271] Bhaskar Mitra. Search and society: Reimagining information access for radical futures. *arXiv preprint arXiv:2403.17901*, 2024.

- [272] Bhaskar Mitra and Nick Craswell. Neural models for information retrieval. *arXiv preprint arXiv:1705.01509*, 2017.
- [273] Bhaskar Mitra, Nick Craswell, et al. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018.
- [274] Bhaskar Mitra, Henriette Cramer, and Olya Gurevich. Sociotechnical implications of generative artificial intelligence for information access, 2024. URL <https://arxiv.org/abs/2405.11612>.
- [275] Stefano Mizzaro. How many relevances in information retrieval? *Interacting with computers*, 10(3):303–320, 1998.
- [276] Alistair Moffat. Computing maximized effectiveness distance for recall-based metrics. *TKDE*, 30(1):198–203, 2017.
- [277] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):2:1–2:27, 2008.
- [278] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, December 2008.
- [279] Alistair Moffat, Joel Mackenzie, Paul Thomas, and Leif Azzopardi. A flexible framework for offline effectiveness metrics. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 578–587. ACM, 2022.
- [280] Shrestha Mohanty, Negar Arabzadeh, Julia Kiseleva, Artem Zholus, Milagro Teruel, Ahmed Awadallah, Yuxuan Sun, Kavya Srinet, and Arthur Szlam. Transforming human-centered ai collaboration: Redefining embodied agents capabilities through interactive grounded language instructions. *arXiv preprint arXiv:2305.10783*, 2023.
- [281] Shrestha Mohanty, Negar Arabzadeh, Andrea Tupini, Yuxuan Sun, Alexey Skrynnik, Artem Zholus, Marc-Alexandre Côté, and Julia Kiseleva. Idat: A multi-modal dataset and toolkit for building and evaluating interactive task-solving agents. *arXiv preprint arXiv:2407.08898*, 2024.
- [282] Sophie Monchaux, Franck Amadieu, Aline Chevalier, and Claudette Mariné. Query strategies during information searching: Effects of prior domain knowledge and complexity of the information problems to be solved. *Information Processing & Management (IP&M)*, 51(5):557–569, 2015.

- [283] William Morgan, Warren Greiff, and John Henderson. Direct maximization of average precision by hill-climbing, with a comparison to a maximum entropy approach. In *Proceedings of HLT-NAACL 2004: Short Papers*, HLT-NAACL-Short '04, pages 93–96, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [284] Vivek Myers, Andre Wang He, Kuan Fang, Homer Rich Walke, Philippe Hansen-Estruch, Ching-An Cheng, Mihai Jalobeanu, Andrey Kolobov, Anca Dragan, and Sergey Levine. Goal representations for instruction following: A semi-supervised language interface to control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 3894–3908. PMLR, 06–09 Nov 2023.
- [285] Vidhya Navalpakkam, Ravi Kumar, Lihong Li, and D. Sivakumar. Attention and selection in online choice tasks. In Judith Masthoff, Bamshad Mobasher, Michel C. Desmarais, and Roger Nkambou, editors, *User Modeling, Adaptation, and Personalization*, pages 200–211, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
- [286] Vidhya Navalpakkam, LaDawn Jentzsch, Rory Sayres, Sujith Ravi, Amr Ahmed, and Alexander J. Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In Daniel Schwabe, Virgilio A. F. Almeida, Hartmut Glaser, Ricardo Baeza-Yates, and Sue B. Moon, editors, *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 953–964. International World Wide Web Conferences Steering Committee / ACM, 2013.
- [287] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660, December 2016.
- [288] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *CoRR*, abs/1901.04085, 2019.
- [289] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. From doc2query to doctttttquery. *Online preprint*, 6, 2019.
- [290] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with BERT. *CoRR*, abs/1910.14424, 2019.
- [291] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*, abs/1904.08375, 2019.
- [292] Rodrigo Frassetto Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *CoRR*, abs/1901.04085, 2019.

- [293] Rodrigo Frassetto Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with BERT. *CoRR*, abs/1910.14424, 2019.
- [294] Douglas W. Oard, Tetsuya Sakai, and Noriko Kando. Celebrating 20 years of NTCIR: the book. In Nicola Ferro, Ian Soboroff, and Min Zhang, editors, *Proceedings of the 9th International Workshop on Evaluating Information Access co-located with the 14th NTCIR Conference on the Evaluation of Information Access Technologies (NTCIR 2019), Tokyo, Japan, June 10, 2019*, 2019.
- [295] Artem Obukhov and Mikhail Krasnyanskiy. Quality assessment method for gan based on modified metrics inception score and fr chet inception distance. In *Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1*, pages 102–114. Springer, 2020.
- [296] Katherine Ognyanova, David Lazer, Ronald E Robertson, and Christo Wilson. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*, 2020.
- [297] Jonas Oppenlaender. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, pages 192–202, 2022.
- [298] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkil , and Shin’ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14277–14286, 2023.
- [299] Arnold Overwijk, Chenyan Xiong, and Jamie Callan. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3360–3362, 2022.
- [300] Ville Paananen, Jonas Oppenlaender, and Aku Visuri. Using text-to-image generation for architectural design ideation, 2023.
- [301] Lawrence Page. The pagerank citation ranking: Bringing order to the web. Technical report, Technical Report, 1999.
- [302] Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. In google we trust: Users’ decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication*, 12(3):801–823, 04 2007. ISSN 1083-6101.

- [303] Jiayi Pan, Yichi Zhang, Nicholas Tomlin, Yifei Zhou, Sergey Levine, and Alane Suhr. Autonomous evaluation and refinement of digital agents. *arXiv preprint arXiv:2404.06474*, 2024.
- [304] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- [305] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.
- [306] G. Parmar, R. Zhang, and J. Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11400–11410, Los Alamitos, CA, USA, jun 2022. IEEE Computer Society.
- [307] Andrew Parry, Maik Fröbe, Harrison Scells, Ferdinand Schlatt, Guglielmo Faggioli, Saber Zerhoubi, Sean MacAvaney, and Eugene Yang. Variations in relevance judgments and the shelf life of test collections. *arXiv preprint arXiv:2502.20937*, 2025.
- [308] Nikita Pavlichenko and Dmitry Ustalov. Best prompts for text-to-image models and how to find them. *arXiv preprint arXiv:2209.11711*, 2022.
- [309] Virgil Pavlu, Shahzad Rajput, Peter B Golbus, and Javed A Aslam. Ir system evaluation using nugget-based test collections. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 393–402, 2012.
- [310] Gustavo Penha, Arthur Câmara, and Claudia Hauff. Evaluating the robustness of retrieval pipelines with query variation generators. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*, volume 13185 of *Lecture Notes in Computer Science*, pages 397–412. Springer, 2022.
- [311] Gustavo Penha, Enrico Palumbo, Maryam Aziz, Alice Wang, and Hugues Bouchard. Improving content retrievability in search with controllable query generation. In Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben, editors, *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3182–3192. ACM, 2023.

- [312] Carol Peters. Information retrieval evaluation in a changing world lessons learned from 20 years of CLEF. 2019.
- [313] Vitali Petsiuk, Alexander E. Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A. Plummer, Ori Kerret, Tonio Buonassisi, Kate Saenko, Armando Solar-Lezama, and Iddo Drori. Human evaluation of text-to-image models on a multi-task benchmark, 2022.
- [314] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.
- [315] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, pages 647–664. Springer, 2020.
- [316] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. The Expando-Mono-Duo design pattern for text ranking with pretrained sequence-to-sequence models. *CoRR*, abs/2101.05667, 2021.
- [317] Ronak Pradeep, Kai Hui, Jai Gupta, Ádám D. Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Q. Tran. How does generative retrieval scale to millions of passages? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1305–1321. Association for Computational Linguistics, 2023.
- [318] Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, and Jimmy Lin. Initial nugget evaluation results for the trec 2024 rag track with the autonuggetizer framework. *arXiv preprint arXiv:2411.09607*, 2024.
- [319] Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*, 2023.
- [320] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*, 2020.
- [321] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

- [322] Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. Report on the 1st workshop on large language model for evaluation in information retrieval (llm4eval 2024) at sigir 2024, 2024.
- [323] Hossein A Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles LA Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. LLM4Eval: Large language model for evaluation in IR. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3040–3043, 2024.
- [324] Hossein A. Rahmani, Emine Yilmaz, Nick Craswell, Bhaskar Mitra, Paul Thomas, Charles L. A. Clarke, Mohammad Aliannejadi, Clemencia Siro, and Guglielmo Faggioli. Llmjudge: Llms for relevance judgments, 2024.
- [325] Shahzad Rajput, Virgiliu Pavlu, Peter B. Golbus, and Javed A. Aslam. A nugget-based test collection construction paradigm. In Craig Macdonald, Iadh Ounis, and Ian Ruthven, editors, *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1945–1948. ACM, 2011.
- [326] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.
- [327] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [328] Sri Devi Ravana, Prabha Rajagopal, and Vimala Balakrishnan. Ranking retrieval systems using pseudo relevance judgments. *Aslib Journal of Information Management*, 67(6):700–714, 2015.
- [329] Navyasri Reddy, Samyak Jain, Pradeep Yarlagadda, and Vineet Gandhi. Tidying deep saliency prediction architectures. In *IROS*, 2020.
- [330] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019.
- [331] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

- [332] Ehud Reiter and Anja Belz. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558, 2009.
- [333] Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530, 2007.
- [334] Cynthia L. Bennett Emily Denton Rida Qadri, Renee Shelby. Ai’s regimes of representation: A community-centered study of text-to-image models in south asia. In *2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’23*, 2023.
- [335] Soo Young Rieh and Hong (Iris) Xie. Analysis of multiple query reformulations on the eeb: The interactive information retrieval context. *Information Processing & Management (IP&M)*, 42(3):751–768, May 2006.
- [336] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST, January 1995.
- [337] Stephen E Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR’94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer, 1994.
- [338] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [339] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.
- [340] Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2): 1–39, 2022.

- [341] Tetsuya Sakai and Noriko Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11:447–470, 2008.
- [342] Tetsuya Sakai and Zhaohao Zeng. Good evaluation measures based on document preferences. In *43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 359–368, Xi’an, China, July 2020.
- [343] Tetsuya Sakai, Makoto P. Kato, and Young-In Song. Click the search button and be happy: Evaluating direct and immediate information access. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, page 621–630, 2011.
- [344] Alireza Salemi and Hamed Zamani. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2395–2400, 2024.
- [345] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 29, 2016.
- [346] David P Sander and Laura Dietz. Exam: How to evaluate retrieve-and-generate systems for users who do not (yet) know what they want. In *DESIRES*, pages 136–146, 2021.
- [347] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends® in Information Retrieval*, 4(4):247–375, 2010.
- [348] Mark Sanderson, Falk Scholer, and Andrew Turpin. Relatively relevant: Assessor shift in document judgements. In *Australasian Document Computing Symposium*, 2010.
- [349] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.
- [350] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *CoRR*, abs/2112.01488, 2021.
- [351] Tefko Saracevic. *The notion of relevance in information science*. 05 2019.

- [352] Falk Scholer, Andrew Turpin, and Mark Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft, editors, *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 1063–1072. ACM, 2011.
- [353] Sebastian Schultheiß and Dirk Lewandowski. Misplaced trust? the relationship between trust, ability to identify commercially influenced results and search engine preference. *Journal of Information Science*, 49(3):609–623, 2023.
- [354] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.
- [355] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*, 2019.
- [356] Elizabeth Seger, Shahar Avin, Gavin Pearson, Mark Briers, Seán Ó Heigeartaigh, Helena Bacon, Henry Ajder, Claire Alderson, Fergus Anderson, Joseph Baddeley, et al. Tackling threats to informed decision-making in democratic societies: Promoting epistemic security in a technologically-advanced world. 2020.
- [357] Mahsa Seifkar, Linh Nhi Phan Minh, Negar Arabzadeh, Charles LA Clarke, and Mark D Smucker. A preference judgment tool for authoritative assessment. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3100–3104, 2023.
- [358] Anna Sepiarskaia, Julia Kiseleva, Filip Radlinski, and Maarten de Rijke. Preference elicitation as an optimization problem. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 172–180, 2018.
- [359] Jami J. Shah, Noe Vargas-Hernandez, and Steve M. Smith. Metrics for measuring ideation effectiveness. *Design Studies*, 24(2):111–134, March 2003.
- [360] Yunqiu Shao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. On annotation methodologies for image search evaluation. *ACM Transactions on Information Systems*, 37(3), March 2019.
- [361] Shikhar Sharma, Dendi Suhubdy, Vincent Michalski, Samira Ebrahimi Kahou, and Yoshua Bengio. Chatpainter: Improving text to image generation using dialogue. *arXiv preprint arXiv:1802.08216*, 2018.

- [362] Chengyao Shen and Qi Zhao. Webpage saliency. In *ECCV*, pages 33–46, 2014.
- [363] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749, 2020.
- [364] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.
- [365] Johanna M Silvennoinen and Jussi PP Jokinen. Appraisals of salient visual elements in web page design. *Advances in Human-Computer Interaction*, 2016, 2016.
- [366] Brandon Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets, 2023.
- [367] Mark D. Smucker and Charles L. A. Clarke. Time-based calibration of effectiveness measures. In William R. Hersh, Jamie Callan, Yoelle Maarek, and Mark Sanderson, editors, *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, pages 95–104. ACM, 2012.
- [368] Ian Soboroff, Charles Nicholas, and Patrick Cahan. Ranking retrieval systems without relevance judgments. In *SIGIR*, pages 66–73, 2001.
- [369] Michael Soloveitchik, Tzvi Diskin, Efrat Morin, and Ami Wiesel. Conditional frechet inception distance. *arXiv preprint arXiv:2103.11521*, 2021.
- [370] Qiurong Song and Jiepu Jiang. How misinformation density affects health information search. In *Proceedings of the ACM Web Conference 2022*, pages 2668–2677, 2022.
- [371] Karen Sparck Jones and Cornelis Joost Van Rijsbergen. Information retrieval test collections. *Journal of documentation*, 32(1):59–75, 1976.
- [372] Karen Spark-Jones. Report on the need for and provision of an ‘ideal’ information retrieval test collection. *Computer Laboratory*, 1975.
- [373] Karen Spärck Jones and C.J. Van Rijsbergen. Report on the need for and provision of an ‘ideal’ information retrieval test collection. Technical Report 5266, British Library Research and Development Report, 1975.

- [374] Bernd Carsten Stahl and Damian Eke. The ethics of chatgpt—exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74:102700, 2024.
- [375] Moritz Staudinger, Wojciech Kusa, Florina Piroi, Aldo Lipani, and Allan Hanbury. A reproducibility and generalizability study of large language models for query generation. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pages 186–196, 2024.
- [376] Gencer Sumbul, Sonali Nayak, and Begüm Demir. Sd-rsic: Summarization-driven deep remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6922–6934, 2020.
- [377] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. Learning to tokenize for generative retrieval. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [378] Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*, pages 14918–14937, 2023.
- [379] Krysta M. Svore and Christopher J.C. Burges. A machine learning approach for improved BM25 retrieval. In *18th ACM International Conference on Information and Knowledge Management*, pages 1811–1814, Hong Kong, China, November 2009.
- [380] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.
- [381] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.
- [382] Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*, 2023.
- [383] Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint arXiv:2303.07992*, 2023.

- [384] Tao Tao and ChengXiang Zhai. An exploration of proximity measures in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 295–302, 2007.
- [385] Jaime Teevan, Kevyn Collins-Thompson, Ryen W. White, and Susan Dumais. Slow search. *Communications of the ACM*, 57(8):36–38, aug 2014. ISSN 0001-0782.
- [386] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- [387] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.
- [388] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1930–1940, 2024.
- [389] Lindia Tjauatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *arXiv preprint arXiv:2311.04076*, 2023.
- [390] Anastasios Tombros and Mark Sanderson. Advantages of query biased summaries in information retrieval. In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 2–10. ACM, 1998.
- [391] S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *CoRR*, abs/2401.01313, 2024.
- [392] Sergio Duarte Torres, Djoerd Hiemstra, and Pavel Serdyukov. Query log analysis in the context of information retrieval for children. In Fabio Crestani, Stéphane Marchand-Maillet, Hsin-Hsi Chen, Efthimis N. Efthimiadis, and Jacques Savoy, editors, *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*, pages 847–848. ACM, 2010.
- [393] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

- Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [394] Andrew Turpin and Falk Scholer. User performance versus precision measures for simple search tasks. In Efthimis N. Efthimiadis, Susan T. Dumais, David Hawking, and Kalervo Järvelin, editors, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*, pages 11–18. ACM, 2006.
 - [395] Inigo Jauregi Unanue, Jacob Parnell, and Massimo Piccardi. Berttune: Fine-tuning neural machine translation with bertscore. *arXiv preprint arXiv:2106.02208*, 2021.
 - [396] Shivani Upadhyay, Ehsan Kamalloo, and Jimmy Lin. Llms can patch up missing relevance judgments in evaluation, 2024.
 - [397] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. A large-scale study of relevance assessments with large language models: An initial look, 2024.
 - [398] Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. Umbrella: Umbrella is the (open-source reproduction of the) bing relevance assessor. *arXiv preprint arXiv:2406.06519*, 2024.
 - [399] Teemu Vahtola, Mathias Creutz, and Jörg Tiedemann. It is not easy to detect paraphrases: Analysing semantic similarity with antonyms and negation using the new semantoneg benchmark. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 249–262, 2022.
 - [400] Roelof van Zwol, Vanessa Murdock, Lluís Garcia Pueyo, and Georgina Ramirez. Diversifying image search with user generated content. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pages 67–74, 2008.
 - [401] Henriikka Vartiainen and Matti Tedre. Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity*, 34(1):1–21, 2023.
 - [402] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.
 - [403] E. Voorhees, D.K. Harman, National Institute of Standards, and Technology (US). *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge eMA MA, 2005.

- [404] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323, Melbourne, Australia, 1998.
- [405] Ellen M Voorhees. Report on trec-9. In *ACM SIGIR Forum*, volume 34, pages 1–8. ACM New York, NY, USA, 2000.
- [406] Ellen M Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–82, 2001.
- [407] Ellen M Voorhees. The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages*, pages 355–370. Springer, 2001.
- [408] Ellen M Voorhees. The evolution of cranfield. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*, pages 45–69, 2019.
- [409] Ellen M. Voorhees, Nick Craswell, and Jimmy Lin. Too many relevants: Whither cranfield test collections? In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2970–2980. ACM, 2022.
- [410] Cheng Wang, Delei Chen, Lin Hao, Xuebo Liu, Yu Zeng, Jianwei Chen, and Guokai Zhang. Pulmonary image classification based on inception-v3 transfer learning model. *IEEE Access*, 7:146533–146541, 2019.
- [411] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, 2023.
- [412] Jindong Wang, Xixu HU, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, Binxing Jiao, Yue Zhang, and Xing Xie. On the robustness of chatGPT: An adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023.
- [413] Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*, 2023.

- [414] Lidan Wang, Jimmy Lin, and Donald Metzler. A cascade ranking model for efficient ranked retrieval. In Wei-Ying Ma, Jian-Yun Nie, Ricardo Baeza-Yates, Tat-Seng Chua, and W. Bruce Croft, editors, *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 105–114. ACM, 2011.
- [415] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.
- [416] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18359–18369, 2023.
- [417] Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. Rethinking the evaluation for conversational recommendation in the era of large language models. *arXiv preprint arXiv:2305.13112*, 2023.
- [418] Yifan Wang, Lingdan Wu, Liang Luo, Yifen Zhang, and Guangheng Dong. Short-term internet search using makes people rely on search engines when facing unknown issues. *PloS one*, 12(4):e0176325, 2017.
- [419] Yufei Wang, Wanjuan Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023.
- [420] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.
- [421] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.
- [422] Ran Wei, Jing Guo, Sai Wang, and C Huang. The role of digital information accessibility in shaping the relationships of exposure to covid-19 misinformation and cognitive and attitudinal effects in asia. *Communication and Society*, 62:207–264, 2022.
- [423] Ryen W White and Chirag Shah. *Information Access in the Era of Generative AI*. Springer, 2025.

- [424] Ryen W. White, Susan T. Dumais, and Jaime Teevan. Characterizing the influence of domain expertise on web search behavior. In Ricardo Baeza-Yates, Paolo Boldi, Berthier A. Ribeiro-Neto, and Berkant Barla Cambazoglu, editors, *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009*, pages 132–141. ACM, 2009.
- [425] Norbert Wiener. Some moral and technical consequences of automation. *Science*, 131(3410):1355–1358, 1960.
- [426] Kyle Williams and Imed Zitouni. Does that mean you’re happy? rnn-based modeling of user interaction sequences to detect good abandonment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 727–736, 2017.
- [427] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabisa. Detecting good abandonment in mobile search. In *Proceedings of the 25th International Conference on World Wide Web*, pages 495–505, 2016.
- [428] Kyle Williams, Julia Kiseleva, Aidan C Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabisa. Is this your final answer? evaluating the effect of answers on good abandonment in mobile search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 889–892, 2016.
- [429] Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.
- [430] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, 2024.
- [431] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.
- [432] Zhijing Wu, Mark Sanderson, Berkant Barla Cambazoglu, W. Bruce Croft, and Falk Scholer. Providing direct answers in search results: A study of user behavior. In Mathieu d’Aquin, Stefan Dietze, Claudia Hauff, Edward Curry, and Philippe Cudré-Mauroux, editors, *The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1635–1644. ACM, 2020.

- [433] Xiaoling Xia, Cui Xu, and Bing Nan. Inception-v3 for flower classification. In *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, pages 783–787. IEEE, 2017.
- [434] Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Haitian Chen, Min Zhang, and Shaoping Ma. Preference-based evaluation metrics for web image search. In *43rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2101–2104, Xi’an, China, July 2020.
- [435] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval, 2020.
- [436] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, abs/2007.00808, 2020.
- [437] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *ICLR*, 2021.
- [438] Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pages 1–10, 2022.
- [439] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [440] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11. ACM, 1996. doi: 10.1145/243199.243202.
- [441] Xinyi Yan, Chengxi Luo, Charles LA Clarke, Nick Craswell, Ellen M Voorhees, and Pablo Castells. Human preferences as dueling bandits. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 567–577, 2022.
- [442] Li-Chia Yang and Alexander Lerch. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784, 2020.

- [443] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of Lucene for information retrieval research. In *40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1256, Tokyo, Japan, August 2017.
- [444] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Reproducible ranking baselines using Lucene. *Journal of Data and Information Quality*, 10(4):1–20, December 2018.
- [445] Tianchi Yang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, and Qi Zhang. Auto search indexer for end-to-end document retrieval. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6955–6970. Association for Computational Linguistics, 2023.
- [446] Ziyang Yang, Alistair Moffat, and Andrew Turpin. Pairwise crowd judgments: Preference, absolute, and ratio. In *23rd Australasian Document Computing Symposium*, Dunedin, New Zealand, December 2018.
- [447] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.
- [448] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roei Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szepkektor. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [449] Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, Nitish Shirish Keskar, and Caiming Xiong. Modeling multi-hop question answering as single sequence prediction. *arXiv preprint arXiv:2205.09226*, 2022.
- [450] Emine Yilmaz, Evangelos Kanoulas, and Javed A Aslam. A simple and efficient sampling method for estimating ap and ndcg. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610, 2008.
- [451] Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. Neural generative question answering. *arXiv preprint arXiv:1512.01337*, 2015.
- [452] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022.

- [453] Maia Zaharieva, Bogdan Ionescu, Alexandru-Lucian Gînsca, Rodrygo LT Santos, and Henning Müller. Retrieving diverse social images at mediaeval 2017: Challenges, dataset and evaluation. In *MediaEval*, 2017.
- [454] ChengXiang Zhai, William W Cohen, and John Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Acm sigir forum*, volume 49, pages 2–9. ACM New York, NY, USA, 2015.
- [455] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. Repbert: Contextualized text embeddings for first-stage retrieval. *arXiv preprint arXiv:2006.15498*, abs/2006.15498, 2020.
- [456] Chi Zhang, Penglin Cai, Yuhui Fu, Haoqi Yuan, and Zongqing Lu. Creative agents: Empowering agents with imagination for creative tasks. *arXiv preprint arXiv:2312.02519*, 2023.
- [457] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 833–842, 2021.
- [458] Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. Are large language models good at utility judgments? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1941–1951, 2024.
- [459] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019.
- [460] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57, 2024.
- [461] Xinyu Zhang, Sebastian Hofstätter, Patrick Lewis, Raphael Tang, and Jimmy Lin. Rank-without-gpt: Building gpt-independent listwise rerankers on open-source large language models. *arXiv preprint arXiv:2312.02969*, 2023.
- [462] Yinan Zhang, Xueqing Liu, and ChengXiang Zhai. Information retrieval evaluation as search simulation: A general formal framework for ir evaluation. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 193–200, 2017.

- [463] Yue Zhang, ChengCheng Hu, Yuqi Liu, Hui Fang, and Jimmy Lin. Learning to rank in the age of Muppets: Effectiveness–efficiency tradeoffs in multi-stage ranking. In Nafise Sadat Moosavi, Iryna Gurevych, Angela Fan, Thomas Wolf, Yufang Hou, Ana Marasović, and Sujith Ravi, editors, *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 64–73, Virtual, November 2021. Association for Computational Linguistics.
- [464] Qinghua Zhao, Yuxuan Lei, Qiang Wang, Zhongfeng Kang, and Junfeng Liu. Enhancing text representations separately with entity descriptions. *Neurocomputing*, 552:126511, 2023.
- [465] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.
- [466] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- [467] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.
- [468] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. LAFITE: towards language-free training for text-to-image generation. *CoRR*, abs/2111.13792, 2021.
- [469] Dongqing Zhu and Ben Carterette. An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, Geneva, Switzerland, July 2010.
- [470] Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.
- [471] Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Berdersky. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *arXiv preprint arXiv:2310.14122*, 2023.
- [472] Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. Through the lens of core competency: Survey on evaluation of large language models. In Jiajun Zhang, editor, *Proceedings of the*

22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum), pages 88–109, Harbin, China, August 2023. Chinese Information Processing Society of China.

- [473] Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 307–314. ACM, 1998.

Appendix A

GitHub Repositories for Reproducibility

To facilitate reproducibility of the experiments and methods described in this thesis, this appendix provides links to the corresponding GitHub repositories. Each repository contains the code, scripts, or datasets used in a specific part of the thesis. The repositories are organized below based on the chapter or section of the thesis they relate to.

Thesis Section	GitHub Repository
Section 3.1: Shallow pooling for sparse labels	https://github.com/Narabzad/Shallow-Pooling-for-Sparse-labels
Section 4.2: LLM-based relevance judgments	https://github.com/Narabzad/llm-relevance-judgement-comparison
Section 4.3: Prompt Sensitivity in LLM-based relevance judgments	https://github.com/Narabzad/prompt-sensitivity-relevance-judgements
Section 5.3.1: AgentEval	https://github.com/Narabzad/AgentEval
Section 5.2: Evaluation of text-to-image generation systems	https://github.com/Narabzad/Set-Based-Text-to-ImageGeneration