DeTAILS: Deep Thematic Analysis with Iterative LLM Support

by

Ansh Sharma

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2025

© Ansh Sharma 2025

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 thesis includes first-authored peer-reviewed material that has appeared in conference proceedings published by The Association for Computing Machinery (ACM). The ACM's policy on reuse of published materials in a dissertation is as follows[1].

> Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included. Authors can use any portion of their own work in presentations and in the classroom (and no fee is expected).

The conference paper from which I have adapted content is the following:

- **Portions of Chapter 1, 2, 3, 8:** Ansh Sharma and James R Wallace. 2025. DeTAILS: Deep Thematic Analysis with Iterative LLM Support. In Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25). Association for Computing Machinery, New York, NY, USA, Article 28, 1–7. https://doi.org/10.1145/3719160.3735657

[1] `https://authors.acm.org/author-resources/author-rights` (Accessed on July 27, 2025).

Abstract

Reflexive thematic analysis (TA) yields rich insights but is challenging to scale to large datasets due to the intensive, iterative interpretation it requires. We present DeTAILS: Deep Thematic Analysis with Iterative LLM Support, a researcher-centered toolkit that integrates large language model (LLM) assistance into each phase of Braun & Clarke's six-phase reflexive TA process through iterative human-in-the-loop workflows. DeTAILS introduces key features such as "memory snapshots" to incorporate the analyst's insights, "redo-with-feedback" loops for iterative refinement of LLM suggestions, and editable LLM-generated codes and themes, enabling analysts to accelerate coding and theme development while preserving researcher control and interpretive depth. In a user study with 18 qualitative researchers (novice to expert) analyzing a large, heterogeneous dataset, DeTAILS demonstrated high usability. The study also showed that chaining LLM assistance across analytic phases enabled scalable yet robust qualitative analysis. This work advances Human-LLM collaboration in qualitative research by demonstrating how LLMs can augment reflexive thematic analysis without compromising researcher agency or trust.

Acknowledgements

Territorial Acknowledgements

During my Master's research, I spent most of my time living and working in the traditional territory of the Attawandaron (also known as Neutral), Anishinaabe and Haudenosaunee peoples. My rental home (Columbia Lake Village) and the University of Waterloo are situated on the Haldimand Tract, the land promised to the Six Nations in 1784 that includes six miles on each side of the Grand River.

Funding Acknowledgements

My gratitude to all sources of my funding including:

- Graduate Research Assistant Positions at the University of Waterloo, David R. Cheriton School of Computer Science
- Graduate Teaching Assistant Positions at the University of Waterloo, David R. Cheriton School of Computer Science
- International Master's Award of Excellence, University of Waterloo, David R. Cheriton School of Computer Science
- Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants
- MITACS Globalink Graduate Fellowship
- Google Cloud Research Credits Program

Personal Acknowledgements

To Professor James R. Wallace, my supervisor: your support and guidance have been invaluable throughout my master's journey. From brainstorming project ideas to refining the final thesis draft, your unwavering encouragement at every stage provided me with the confidence and resources necessary to succeed. I am deeply grateful for the opportunities you have afforded me and for the mentorship you have so generously offered.

To Professor Anamaria Crisan and Professor Karen Cochrane: thank you for taking the time to read my thesis and for offering your insights and suggestions to improve my work.

To Professor Robin Cohen, Professor Paulo Alencar, Professor Jian Zhao, and Professor Wenhu Chen: thank you for teaching graduate-level courses in such a captivating and enjoyable manner. Your classes were not only intellectually stimulating but also enhanced my presentation skills, making every interaction with my classmates something I eagerly anticipated.

To my parents, Mrs. Kanta Sharma and Mr. Bijender Kumar Sharma, and to my sister, Dr. Aarushi Sharma: your constant support has been the bedrock of my success. Your love and encouragement have sustained me through every challenge. Thank you for always being there.

To all my dear friends and labmates: thank you for the coffee chats.

Table of Contents

Author's Declaration **ii**

Statement of Contributions **iii**

Abstract **iv**

Acknowledgements **v**

List of Figures **xi**

List of Tables **xii**

1 Introduction **1**

1.1 Thesis Statement 2

1.2 Contributions 3

2 Background and Related Work **4**

2.1 Computational Aids for Thematic Analysis: Topics vs. Themes 5

2.2 Large Language Models as Coding Assistants 6

2.3 Prompt Engineering and Chain-of-Thought Reasoning 8

2.4 Human-AI Collaboration: Control, Reliability, and "Chaining LLMs" . . . 9

2.5 Where Current Research Falls Short 12

3 DeTAILS: Workflow and Technical Implementation **13**
3.1 Background Research . . . 14
3.2 Loading Data . . . 17
3.3 Coding . . . 19
3.4 Reviewing Codes . . . 19
3.5 Generating Themes . . . 21
3.6 Report . . . 21
3.7 Technical Implementation . . . 23
3.7.1 Services & Data Stores . . . 23
3.7.2 LLM Providers . . . 24
3.7.3 Prompts . . . 24
3.7.4 Verification & Consistency . . . 25

4 Evaluation: Usability Study **28**
4.1 Participants and Recruitment . . . 28
4.2 Materials and Apparatus . . . 29
4.3 Study Procedure . . . 29
4.4 Data Collection and Analysis . . . 31

5 Results **35**
5.1 Quantitative Results . . . 35
5.1.1 F1 Scores . . . 35
5.1.2 Time Analysis . . . 36
5.1.3 NASA-TLX . . . 38
5.1.4 Perceived Usefulness Scale . . . 40
5.1.5 AttrakDiff . . . 40
5.2 Qualitative Results . . . 42
5.2.1 Phase 1: Background Research . . . 42
5.2.2 Phase 3: Coding . . . 46
5.2.3 Phase 4: Reviewing Codes . . . 47
5.2.4 Phase 5: Generating Themes . . . 49

6 Discussion **52**
6.1 LLM-assisted Coding Accelerates Analysis 52
6.2 LLMs Help Transform Topics into Themes 53
6.3 Integrating Qualitative Research and Chain-of-Thought 54
6.4 Agency and Trust 55

7 Limitations and Future Work **57**

8 Conclusion **59**

References **60**

APPENDICES **74**

A Prompts **75**
A.1 Phase-wise Prompts 75
A.1.1 Related Concepts Prompts 75
A.1.2 Concept Outline Prompts 79
A.1.3 Intial Coding Prompts 81
A.1.4 Initial Codebook Prompts 84
A.1.5 Final Coding Prompts 88
A.1.6 Reviewed Codes Prompts 92
A.1.7 Generating Themes Prompts 102
A.2 Refine Code Prompt 110
A.3 Common Prompts 112
A.3.1 Summarization Prompts 112
A.3.2 Topic Clustering Prompts 112

B Interview Materials **115**

Interview Questionnaire And Survey Flowchart 116

Interview Questions 117

AttrakDiff 135

NASA-TLX 138

PU Scale 141

Feedback letter 143

C Human Ethics Protocol **144**

REB #47221 145

Protocols 147

Information Letter 174

Poster 177

Participant Consent Form 178

D Additional Plots and Tables **180**

D.1 Plots 180

D.1.1 Metrics by expertise level 180

D.2 Tables 184

D.2.1 Metrics by expertise level 184

List of Figures

3.1 DeTAILS workflow for reflexive thematic analysis 14
3.2 DeTAILS' interactive radial map in Background Research phase 15
3.3 DeTAILS' editable Concept Outline table in Background Research phase . 16
3.4 DeTAILS' Edit Mode interface in Coding phase 18
3.5 DeTAILS' Edit View with code clusters in Reviewing Codes phase 20
3.6 DeTAILS' theme generation interface with theme buckets 22

5.1 F1 Scores of participants while using DeTAILS 36
5.2 Participant timings (by phase) while using DeTAILS 37
5.3 NASA-TLX Responses (All Participants) after using DeTAILS 38
5.4 Perceived Usefulness Responses (All Participants) after using DeTAILS . . 39
5.5 AttrakDiff Responses (All Participants) after using DeTAILS 41
5.6 Related Concepts generated by DeTAILS and edited by P18 45
5.7 Themes generated by DeTAILS for P15 51

D.1 F1 Scores while using DeTAILS (by Expertise Level) 180
D.2 Participant Timings (by phase) while using DeTAILS (by Expertise Level) 181
D.3 NASA-TLX Responses (by Expertise Level) 181
D.4 Perceived Usefulness Responses (by Expertise Level) 182
D.5 AttrakDiff Responses (by Expertise Level) 183

List of Tables

3.1 Prompt Structure used for DeTAILS 27

4.1 Study participants demographics and subreddit 34

D.1 Precision, Recall, and F1 at Background Research Phase 184
D.2 Precision, Recall, F1, and Avgerage Cosine Similarity at Coding Phase . . 185
D.3 Precision, Recall, and F1 at Reviewing Codes and Generating Themes Phase 186

Chapter 1

Introduction

Reflexive Thematic Analysis (TA) is a powerful but time-consuming method for identifying patterns of meaning in qualitative data [14]. Analysts iteratively code and interpret data to develop themes, an intensive process that becomes difficult to sustain as datasets grow larger or more diverse [16, 54]. Maintaining consistent analytical depth across large open-ended survey responses or social media posts (e.g., from Reddit) is extremely challenging, often requiring significant time and effort to avoid superficial analysis [7, 54, 121]. The growing scale of qualitative datasets, such as social media posts, has spurred demand for approaches to analyze them efficiently [20]. To tackle this issue researchers have used different methods, including team-based coding and AI-driven tools, to maintain analytical depth across large datasets [20].

Computational text analysis techniques aim to accelerate this process, but they fall short of replacing manual TA. Topic modeling algorithms like Latent Dirichlet Allocation (LDA) [10] and BERTopic [53] can cluster documents into coarse "topics," yet these lack the rich context and interpretive depth of true TA themes [25, 85]. Moreover, these methods operate as "black boxes" that are hard for researchers to trust [18, 42, 64]. Qualitative analysts often prefer the intensive manual approach over opaque tools due to concerns about whether they can adequately perform the interpretive work required in reflexive TA [15, 42].

Recently, large language models (LLMs) have shown promise in assisting thematic analysis. Unlike static algorithms, LLMs like GPT-4 [94] can engage in flexible, context-aware dialogue — for example, suggesting codes for a given excerpt or summarizing a set of texts in natural language. Despite their promise, current LLM-based coding assistants often generate generic or decontextualized codes and themes lacking mechanisms for continuous researcher guidance and context embedding [36, 116]. One-off prompting systems can

speed initial analysis but produce superficial outputs that require extensive manual refinement, undermining time savings [57, 101]. Commercial CAQDAS tools like NVivo offer only partial automation, automating code suggestions (using LLMs) but leaving theme development entirely to users, while collaborative platforms such as CollabCoder rely on time-intensive consensus workflows that limit individual agility and reflexivity [47, 82]. Thus, existing methods lack an integrated, end-to-end LLM-assisted workflow that embeds researcher context, facilitates iterative human-in-the-loop feedback, and accelerates thematic analysis without sacrificing depth or trust.

To address these issues, we developed DeTAILS: Deep Thematic Analysis with Iterative LLM Support [109]. DeTAILS is a researcher-centered toolkit that integrates LLM assistance into the established six-phase TA workflow (following Braun and Clarke's framework [16]) while preserving interpretive control. At each phase, the tool provides AI-suggested insights (for example, initial open-coding suggestions or theme drafts), but always allows the human analyst to edit, refine, or redo the results via natural-language feedback or interactive manual intervention.

To evaluate DeTAILS, we ran a usability study with 18 qualitative researchers varying in TA expertise (6 novices, 6 proficient, 6 experts). Participants analyzed datasets of 30 Reddit posts (with comments). The study included a guided tutorial, a 90-minute interactive task following Braun and Clarke's reflexive TA phases, a review of pre-computed results of a large-scale dataset, post-task surveys (NASA-TLX, AttrakDiff, Perceived Usefulness Scale), and semi-structured interviews. Key quantitative findings showed that the AI's suggested codes aligned closely with the researchers' own decisions. Users also experienced very low mental and physical demands, rated the system's overall usefulness near the top of the scale, and assigned high scores for both practical utility and engaging design. These outcomes remained consistent across all expertise groups. Qualitatively, participants noted faster coding (hours to minutes), built trust via editable suggestions, and deeper interpretation through feedback, proving DeTAILS scales reflexive TA while maintaining human control.

1.1 Thesis Statement

Our research bridges human–computer interaction and qualitative methodology by introducing DeTAILS, a toolkit that enables rigorous yet scalable thematic analysis through an iterative, transparent LLM workflow. We demonstrate how thoughtfully integrating an AI assistant into each phase of reflexive TA can accelerate analysis on large or complex datasets without sacrificing interpretive depth or researcher agency. In essence, we show

that:

> It is possible to harness LLMs to augment qualitative analysis in a way that amplifies human insight rather than replaces it.

1.2 Contributions

This thesis makes three primary contributions to the field of Human-computer Interaction:

1. **The DeTAILS system and workflow:** We introduce a novel interactive application for thematic analysis that integrates LLM assistance into the six phases of reflexive TA.
2. **Empirical evaluation with diverse researchers:** We present findings from a usability study with 18 qualitative researchers from varying research backgrounds, who used DeTAILS to analyze large, heterogeneous datasets of Reddit posts. We show how qualitative researchers used DeTAILS to perform qualitative analysis at scale.
3. **Implications for trust and analytic depth:** We discuss design implications for AI-assisted qualitative research and offer insight into designing AI tools that amplify human analysts' capabilities without compromising reflexivity.

Chapter 2

Background and Related Work

Thematic Analysis (TA) is a foundational qualitative method for identifying patterns of meaning ("themes") within textual data [3, 12, 14]. In the reflexive approach to TA developed by Braun and Clarke, the researcher plays an active, interpretive role in theme development rather than merely applying a fixed codebook or seeking inter-coder reliability [15, 119]. This approach is intentionally flexible and iterative — codes and themes are not pre-defined but emerge through deep engagement with the data and ongoing critical reflection by the analyst [14, 89, 103]. Braun and Clarke outline TA as a multi-phase process (familiarizing with data, generating initial codes, searching for themes, reviewing themes, defining/naming them, and producing a report) [14]. Crucially, reflexive TA recognizes the subjectivity of qualitative analysis: meanings are constructed by the researcher ("the researcher as an instrument") and can vary based on one's theoretical lens or positionality [80, 86]. This interpretive freedom allows rich insights but also demands rigour in how researchers document decisions and remain self-reflexive about their influence on the analysis [43, 80, 89].

While powerful, traditional TA can be time- and labour-intensive, especially as qualitative datasets grow in size and complexity [54, 88, 89]. Ensuring rigour requires the analyst to develop deep familiarity with often voluminous data and to carefully track analytic decisions over repeated readings [54, 89, 103]. If multiple coders collaborate, managing divergent perspectives and reaching consensus on theme definitions adds further overhead [19, 54, 83]. Moreover, the sheer volume of contemporary qualitative data — from long interview transcripts to thousands of open-ended survey responses or social media posts — often makes a purely manual TA approach impractical (or extremely difficult) [16, 54]. For example, online platforms like Reddit can yield massive text corpora [7] that defy manual coding in a reasonable time frame. Researchers without large teams or ample time may

struggle to fully analyze such data, raising the risk of superficial themes or overlooked patterns [54, 121]. In sum, reflexive TA offers analytical depth but faces scalability challenges in the face of big qualitative datasets.

2.1 Computational Aids for Thematic Analysis: Topics vs. Themes

Given these challenges, there is growing interest in computational tools to assist human qualitative analysis [13, 48, 51]. One common avenue has been topic modeling algorithms — such as Latent Dirichlet Allocation (LDA) [10] or more recent transformer-based models like BERTopic [53] — which automatically cluster words or documents into "topics." Tools like the Computational Thematic Analysis (CTA) Toolkit integrate such algorithms to help qualitative researchers identify candidate themes or explore data patterns. For instance, CTA originally incorporated LDA [48] and has since been extended to include BERTopic [71], which uses contextual embeddings to find more granular topic groupings. By applying these models, the toolkit can suggest groups of related texts or keywords in a large corpus, potentially pointing the analyst to prevalent themes that merit investigation.

However, algorithmic topics differ fundamentally from themes in reflexive TA. Topic models rely on statistical co-occurrence or semantic similarity of terms, and thus their "topics" are essentially clusters of words or documents that frequently appear together. These may correspond to surface-level themes (e.g., repeated keywords), but they lack the contextual depth and interpretive framing that defines a TA theme [25, 49, 85]. A TA theme captures a pattern of meaning, often requiring the researcher's interpretation of why certain ideas co-occur and what they signify in relation to the research question [14, 15]. In contrast, an LDA topic is a set of terms that often co-appear, which might be internally coherent yet still descriptive rather than explanatory. As a result, automated topic outputs often need extensive human interpretation and refinement before they become meaningful analytic themes. Prior studies have noted that directly using topic model results in place of human coding can lead to loss of depth — important context may be missing, and spurious word patterns can be mistaken for significant insights [85]. While topic models could rapidly cluster text, the resultant groups might require substantial re-interpretation and editing by researchers to align with true themes of the data.

Another limitation is the "black box" nature of many computational methods. Algorithms like LDA or neural topic models operate via complex statistical processes that are not always transparent to researchers, which can engender skepticism [18]. Qualitative analysts may be understandably wary of trusting an opaque model to surface themes,

especially given the importance of context in qualitative meaning-making. Indeed, interviews with social scientists indicate a tension: researchers find these tools promising for handling scale, yet express doubts about the trustworthiness and interpretability of machine-generated results [42, 64]. They worry that important but subtle content might be missed, or that algorithmic bias could skew the findings. This skepticism, along with practical barriers (many qualitative researchers lack programming or data science expertise), means that in practice a large number of researchers still rely on manual methods or simple tools (e.g., spreadsheets) for coding [71]. General-purpose software like Excel or Google Sheets, or traditional CAQDAS tools without advanced AI integration, remain common because they are familiar and give the researcher full control, even if they offer little assistance with scale [127]. Researchers often prefer the devil they know — intensive manual coding — over a complex new tool that might be hard to use or whose outputs are hard to trust [64]. Thus, despite the potential of computational methods to speed up analysis, their adoption in reflexive TA has been slow due to these usability and trust concerns [15, 42].

2.2 Large Language Models as Coding Assistants

LLMs offer a new avenue for assisting qualitative analysis. Unlike traditional topic models such as LDA, which are based on statistical co-occurrence patterns, LLMs employ deep neural architectures with self-attention mechanisms to support flexible, context-driven dialogue and reasoning. LLMs like GPT-4 are trained on vast text corpora and can generate human-like text, summarize documents, classify content, and even simulate analytical reasoning [40, 94, 139]. This has led researchers to explore LLMs as coding assistants that could help scale up thematic analysis without entirely sacrificing interpretive depth [31, 34, 36, 37, 46, 47, 59, 61, 70, 115, 138]. Unlike topic models that output clusters of words, an LLM can engage with data in a conversational manner — for example, by suggesting potential codes for a given excerpt along with an explanation, or by summarizing a theme across multiple texts in coherent prose. Such capabilities align surprisingly well with tasks in qualitative research that involve interpreting and synthesizing unstructured text.

Early studies have demonstrated that LLMs can support both deductive and inductive coding approaches [27, 34, 45, 72, 73, 84, 99, 111, 131]. In deductive coding, the model is given a predefined codebook or example-coded texts and asked to classify new data accordingly. For instance, Xiao et al. [131] showed that GPT-3 could apply existing codes to interview transcripts with moderate success when provided with clear code definitions and example passages for each code. Few-shot prompting techniques – where the prompt includes one or more manually coded examples – have been effective in guiding LLMs

to mimic a coding scheme. Researchers have fed models with domain-specific codebooks (e.g., a set of mental health categories) and found that the LLM's assignments often align with human coders on obvious cases. However, these approaches also highlight the need for caution: LLMs sometimes over-generalize the examples or apply codes too liberally, missing subtle distinctions that a human would catch. Ensuring the model truly understands the code definitions (rather than just matching keywords) remains a challenge.

For inductive coding, where the goal is to generate new codes or themes from the data, LLMs have shown both promise and pitfalls. Several research teams have experimented with using LLMs like ChatGPT to generate initial codes or theme summaries from qualitative data in an interactive fashion. For example, Zhang et al. [137] prompted ChatGPT to perform thematic analysis on survey responses; they found that detailed prompts which describe the analysis procedure and expected output (e.g., instructing the model on how to identify distinct concepts and group them) yielded more useful results. The model was able to suggest plausible codes for each text segment and even draft descriptions for potential themes. Other work by De Paoli [34] proposed a multi-stage workflow where an LLM first generates a set of codes for each interview, then consolidates unique codes across interviews to form an emergent codebook. The workflow in [70] similarly has the LLM summarize clusters of ideas and iteratively refine them into themes and sub-themes. In these inductive use cases, the LLM is treated as a brainstorming partner that can rapidly surface patterns, which the human researcher can then verify or refine.

Across both inductive and deductive modes, a consistent finding is that LLMs perform best as assistants rather than autonomous analysts [31, 36, 116]. Multiple studies emphasize that LLM-generated coding requires human oversight to ensure quality, context sensitivity, and trustworthiness [23, 65, 69, 104, 106]. For example, Dai, Xiong, and Ku [31] report that while an LLM could suggest reasonable open-codes from raw text, human reviewers needed to merge redundancies, correct misinterpretations, and provide context that the model lacked. Similarly, Tai, Lin, Chen, et al. [116] found that GPT-4 could classify text into high-level categories with high recall, but it sometimes confidently produced incorrect justifications for its choices, which a human had to catch. In an evaluation by Deiner, Zhang, Williams, et al. [36] on social media data, experts judged many LLM-generated themes to be plausible and relevant, yet also noted that the depth and specificity of insight did not fully match a human analyst's work. The LLM tended toward broad or generic themes and occasionally missed the more subtle subthemes that domain experts would identify. These outcomes underscore the notion that LLMs augment but do not replace human interpretation. When used wisely, they can accelerate coding by handling routine text processing and offering draft interpretations, thereby freeing human researchers to apply their expert judgment to refining and contextualizing the results.

2.3 Prompt Engineering and Chain-of-Thought Reasoning

A key factor in successfully leveraging LLMs for qualitative analysis is effective prompt engineering – i.e., crafting the instructions and context given to the model [33, 38, 46, 61, 75, 123, 131, 137, 138]. Simple prompts (e.g., "What themes are in this text?") often yield shallow answers, since the model may default to generalizations or even fabricate patterns (a form of "hallucination") [31, 34]. Recent work has therefore explored more structured prompting techniques that guide the model through the complex reasoning required in thematic analysis. One especially promising approach is Chain-of-Thought (CoT) prompting, which encourages the LLM to generate a step-by-step rationale before giving a final answer [9, 75, 126, 135]. Instead of asking directly for a code or theme, the prompt might instruct the model to first list key ideas in the text, then reflect on how those ideas relate, and finally decide on an appropriate code label — essentially simulating the analyst's thought process.

Studies have shown that CoT prompting can significantly improve LLM performance on coding tasks. Hou et al. [61] found that extending a coding prompt with explicit reasoning steps (e.g., "first identify distinct ideas in this excerpt; next, compare these ideas to see if they form a pattern; then decide on a code label, explaining your choice") led GPT-4 to produce more coherent and transparent coding decisions. The model's output in that case included the intermediate reasoning (for instance, noting that "Idea X and Y both relate to a lack of trust in technology, so I will group them under a 'trust issues' code"), which made it easier for researchers to follow the logic and spot errors. Importantly, this interpretable, staged process reduced instances of non-sequitur or overly generic codes that often plagued one-shot prompting. Similarly, Dunivin [38] demonstrated that prompting GPT-4 to justify each coding decision with reference to the text (a form of CoT) improved its agreement with human coders. In a case study with a complex coding schema, GPT-4 without rationale often misapplied codes, but when forced to explain its choices step-by-step, its accuracy and consistency approached that of humans. Dunivin's findings were striking: GPT-4 with CoT achieved substantial inter-coder reliability on most codes (Cohen's κ ¿ 0.6 for 8 of 9 codes) and even excellent agreement ($\kappa \approx 0.8$) on a few codes, whereas GPT-3.5 or GPT-4 without CoT fell far short. This underscores that the prompt format can make the difference between an LLM that "reads" like a shallow pattern-matcher and one that more closely approximates human reasoning.

Building on this idea, researchers have begun to design multi-turn or multi-step prompting workflows for qualitative analysis. Instead of a single prompt that tries to do everything, a workflow breaks the analysis into a sequence of LLM interactions, each accomplishing a

subtask. For example, a pipeline might involve: (1) an LLM prompt that summarizes or clusters raw data segments into preliminary topics; (2) another prompt that takes those clusters and generates initial codes or categories; (3) a third prompt that drafts theme definitions from the codes; and so on. This approach mirrors the codes-to-theory progression described by Saldana [103] – data → codes → categories → themes – aligning each step with a dedicated AI assist. By chaining prompts in a logical sequence, the LLM can be contextualized with outputs from prior steps, reducing the cognitive load on the model at each stage and potentially improving overall coherence of the final analysis [129]. Several prototypes have implemented such pipelines. For instance, De Paoli [34] chained LLM operations so that after code generation, a subsequent prompt merged and refined those codes, and a further prompt helped organize them into themes. Similarly Gao, Shu, and Yeo [46] also followed the codes-to-theory progression using LLMs to generate themes, which were used to build reports and hierarchical visualizations of codes to themes. Their results suggest that prompt-chaining yields more stable and diverse codebooks compared to a one-shot approach, though human refinement was still needed at each juncture. Stability in codebooks ensures consistent application across the dataset, reducing ambiguity in theme development, while diversity captures a broader range of patterns, preventing the oversight of minority voices or subtle differences in reflexive TA. The general insight is that decomposing the qualitative analysis into smaller reasoning tasks can harness the strengths of LLMs more effectively, much like a human analyst tackles complex analysis in stages. This stepwise use of CoT across an entire workflow is a natural extension of CoT prompting: rather than just a chain of thoughts within one answer, the analyst orchestrates a chain of LLM calls that progressively build up the analysis.

2.4 Human-AI Collaboration: Control, Reliability, and "Chaining LLMs"

While advanced prompting techniques and pipelines can improve LLM outputs, maintaining human control over the process and results is paramount in research settings [23, 104]. Prior work cautions that LLMs are prone to hallucination – they can produce confident-sounding yet incorrect statements – which poses the risk of misleading researchers if the AI's suggestions are taken at face value [114, 133]. There is also the danger of overtrust: users may become overly reliant on the AI and accept its outputs uncritically, especially when the model is fluent and authoritative in tone [47, 50, 74, 113]. Studies have documented instances of LLMs "cheating" or shortcutting tasks in ways that a naive user might not catch [4, 11, 81]. For example, an LLM might pick up on subtle wording in a

prompt and guess a desired answer without truly analyzing the data (a form of reward hacking) [24, 117]. In an evaluation of an AI coding assistant Nyaaba et al. [90] found that although the AI achieved high overlap with human-generated codes, the explanations it provided for those codes were often less specific and less grounded in the data than human interpretations. This lack of specificity indicates that the model may not fully capture the contextual richness that a human would bring, reinforcing the need for human analysts to remain in the loop and inject interpretive depth where the AI cannot. Moreover, even a CoT reasoning chain can contain subtle errors or biases in intermediate steps, which, if not corrected, could compound in later stages of analysis [4]. Researchers therefore argue that iterative human oversight and correction mechanisms are necessary to validate AI-generated codes and themes before they are finalized.

To address these concerns, HCI researchers have explored interface designs that keep humans actively involved at each step of an AI-assisted qualitative analysis [132]. Long, Gero, and Chilton [81] advocate for "AI workflow design" principles that give users guided support while still requiring them to vet and approve outputs at each stage. Instead of a one-shot automation, the idea is to structure the tool such that the user can review and intervene in a stepwise fashion — much like a checklist — to ensure nothing egregious slips through. For example, CollabCoder [47] is a system that integrates LLM suggestions into a collaborative coding workflow, but with rigorous user control. In CollabCoder, two or more analysts begin with independent open-coding of data (with the LLM offering code suggestions they may accept or ignore), then compare and discuss their codes with support from the system, which highlights disagreements and even suggests potential merges. The analysts remain responsible for deciding final codes and organizing them, while the AI aids with tedious tasks like clustering similar codes or retrieving supporting quotes. A user study found this approach improved the efficiency of coding without sacrificing the reliability of the final codebook, precisely because human judgment was exercised at every key juncture. This exemplifies a human-in-the-loop paradigm: the AI handles scalable pattern detection and summarization, but humans guide the analysis direction, make interpretative calls, and correct the AI when it goes astray.

Recently, the notion of "Chaining LLMs" has been introduced as a design strategy to enhance human control in complex AI tasks. Chaining LLMs means linking together a sequence of LLM operations such that the output of one step becomes the input to the next, forming a pipeline or chain [2, 129, 130, 140]. Crucially, users can intervene at each link in the chain — adjusting the inputs, editing the prompt for the next step, or even altering the sequence of steps — to steer the overall process. In the context of qualitative coding, a chained LLM workflow might have separate steps for clustering data, labeling clusters with codes, grouping codes into themes, and writing analytic memos. At each step,

the user can inspect what the LLM did (e.g., the clusters it formed or the code names it generated) and provide feedback or modifications before proceeding. This design ensures that the researcher's intentions and interpretations remain central, with the AI acting as a responsive tool rather than an uncontrollable black box. Gao, Shu, and Yeo [46] describe how their MindCoder system embodies this concept. MindCoder employs a four-step LLM reasoning chain for TA (topic clustering → code labeling → conceptualizing → reporting) and offers an interactive UI where users can view the input and output at each step side-by-side. If a cluster looks odd or a code label is misapplied, the user can "troubleshoot" by editing or rephrasing the prompt and rerunning that step, or by adjusting parameters (like how many clusters to form) and propagating those changes down the chain. Global structural changes — for instance, deciding to add an extra coding step or to skip a summarization — can also be made, and the system will re-run subsequent LLM prompts accordingly. This flexible chaining approach has two major benefits: transparency (the analyst sees each intermediate result the AI produces, rather than only a final mysterious output) and control (the analyst can correct errors early and guide the AI toward the interpretations that matter for their research). Early evaluations of systems like MindCoder indicate that users appreciated the balance of automation and control – they could get a "first draft" thematic analysis in minutes, then iteratively refine it with the system's help, which both saved time and increased confidence in the results.

Despite these encouraging developments, it's important to acknowledge and address ongoing concerns that researchers have about integrating AI into qualitative analysis. One is the potential for AI to introduce bias or flatten the richness of qualitative data. LLMs are trained on large, general internet corpora; if the qualitative data or the analytical lens involves marginalized perspectives or subtle social differences, a general LLM might reflect dominant-culture biases or miss the point entirely [23, 104]. Care must be taken to ensure the AI's involvement does not marginalize the very voices qualitative research seeks to amplify. Another concern is data privacy and ethics [124, 136]. Many LLM tools (e.g., cloud-based APIs like OpenAI's) require uploading data to external servers, which is problematic if the texts contain sensitive or confidential information. Researchers have expressed hesitation to use such services for data like interview transcripts about personal topics, due to uncertainty about how the data might be stored or used by the AI providers. This has driven interest in solutions that can be deployed locally or that allow the use of self-hosted models, ensuring that data never leaves the researcher's control.

2.5 Where Current Research Falls Short

Current research on computational aids for reflexive thematic analysis reveals persistent challenges in scaling qualitative methods while preserving interpretive depth, transparency, and researcher control. LLMs provide a more flexible alternative by simulating reasoning through prompts and supporting both inductive and deductive coding, yet they introduce risks such as hallucinations, over-generalization, and biases from training data, which can undermine trustworthiness without robust human oversight. However, existing systems often fall short in integrating seamless, iterative Human-AI collaboration: they may chain prompts effectively but lack intuitive mechanisms for real-time feedback, dynamic adjustments, or local deployment to address privacy concerns, potentially leading to rigid analyses that compromise reflexivity or exclude sensitive datasets.

To address these gaps, our work is guided by four research questions:

RQ1 Can LLMs accelerate thematic analysis of large data sets?

RQ2 Can LLMs help transform surface-level topics into meaningful themes?

RQ3 Can Chain-of-Thought prompting support iterative, reflexive sense-making?

RQ4 How can AI-enhanced qualitative data analysis tools foster researcher agency and trust?

The toolkit we present, *DeTAILS: Deep Thematic Analysis with Iterative LLM Support* [109], has been designed in light of these considerations. It emphasizes researcher control at every stage — offering suggestions and semi-automated routines, but always allowing the human analyst to query, refine, or redo steps through natural language prompting or manual intervention rather than accepting AI output blindly. By incorporating a "redo with feedback" loop, DeTAILS ensures that humans can iteratively teach or correct the AI, keeping the interpretive authority firmly in the researcher's hands. In summary, our work builds on the rich thread of prior research — from innovations in prompt engineering and CoT prompting to interaction techniques for chaining LLMs — to advance a human-centered approach to AI-assisted thematic analysis that is transparent, flexible, and above all aligned with the reflexive ideals of qualitative research.

Chapter 3

DeTAILS: Workflow and Technical Implementation

DeTAILS follows Braun and Clarke's six-phase reflexive TA framework [16], applied to Reddit data (Figure 3.1). A key innovation is a preliminary Background Research phase that builds a "memory snapshot" from the researcher's insights, giving the LLM essential domain context. Rather than relying on example-based coding [31, 37], DeTAILS uses researcher knowledge to align the LLM, making the process more contextually grounded [77]. It is a standalone application built with Electron, React, SQLite, ChromaDB, FastAPI, LangChain, and Ollama.

The DeTAILS workflow begins with data loading and initial coding of a subset through human-LLM collaboration, producing an initial codebook. This mirrors established TA practices (including those for social media data) in which an initial subset guides broader coding [100]. DeTAILS then applies the initial codebook to the full dataset, allowing for flexible refinement to handle Reddit's heterogeneity. The later phases involve code review, theme development, and report generation, with LLMs supporting each step. We now describe each of these activities, and how DeTAILS combines qualitative depth with AI support.

For demonstration purposes, we showcase several steps of the DeTAILS user interface (from Figure 3.2 to Figure 3.6) via a pilot analysis of December 2024 posts from *r/emotionalintelligence* subreddit [125], processed using the Google Vertex AI API using the gemini-2.5-pro-preview-03-25 model to study the role of AI in mental health.

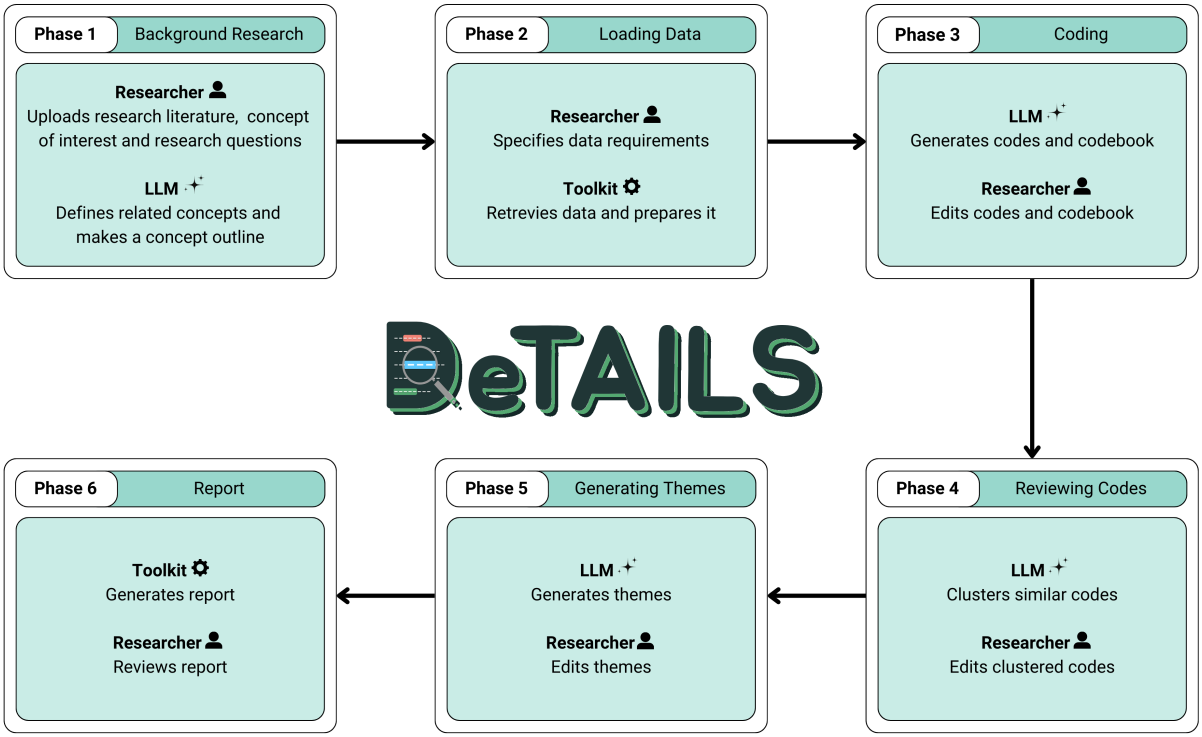


Figure 3.1: *DeTAILS workflow:* To perform reflexive thematic analysis using DeTAILS, researchers go through six phases: *1) Background Research, 2) Loading Data, 3) Coding, 4) Reviewing Codes, 5) Generating Themes, and 6) Report.* In each phase, the researcher guides the process, with the LLM/toolkit executing the necessary tasks. Similar to manual reflexive TA, this workflow is iterative, allowing users to come back to any previous steps and make changes as they go along.

3.1 Background Research

Qualitative researchers establish a robust mental framework for TA by thoroughly understanding existing research and the broader context of their research topic [54, 89]. DeTAILS supports this process by creating a shared conceptual memory that stores essential information — such as research questions, core concepts, and relevant literature — provided by the user during the project's initial phase. This memory remains accessible to both the researcher and the LLM, ensuring that coding decisions stay aligned with the defined research context.

When starting a new project, users begin on the 'Context' Page, where they input

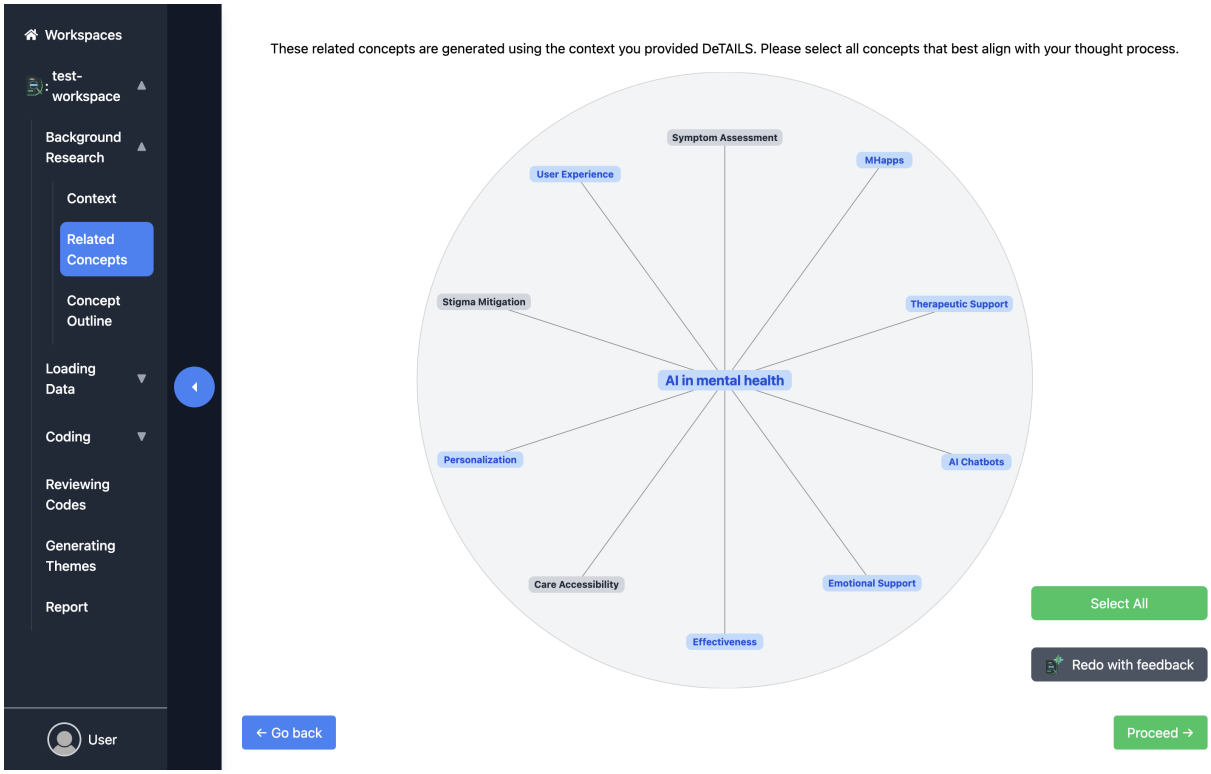


Figure 3.2: In DeTAILS' *Background Research* (phase 1), an interactive radial map visualizes related concepts. The map centers on the core concept and displays related concepts derived from research questions and uploaded literature. Researchers can select concepts, edit them, or regenerate them by clicking the *'Redo with feedback'* button in the bottom right, which opens a feedback modal. Users can also revisit and modify their uploaded research, which updates the map. When ready, users can proceed to the concept outline step.

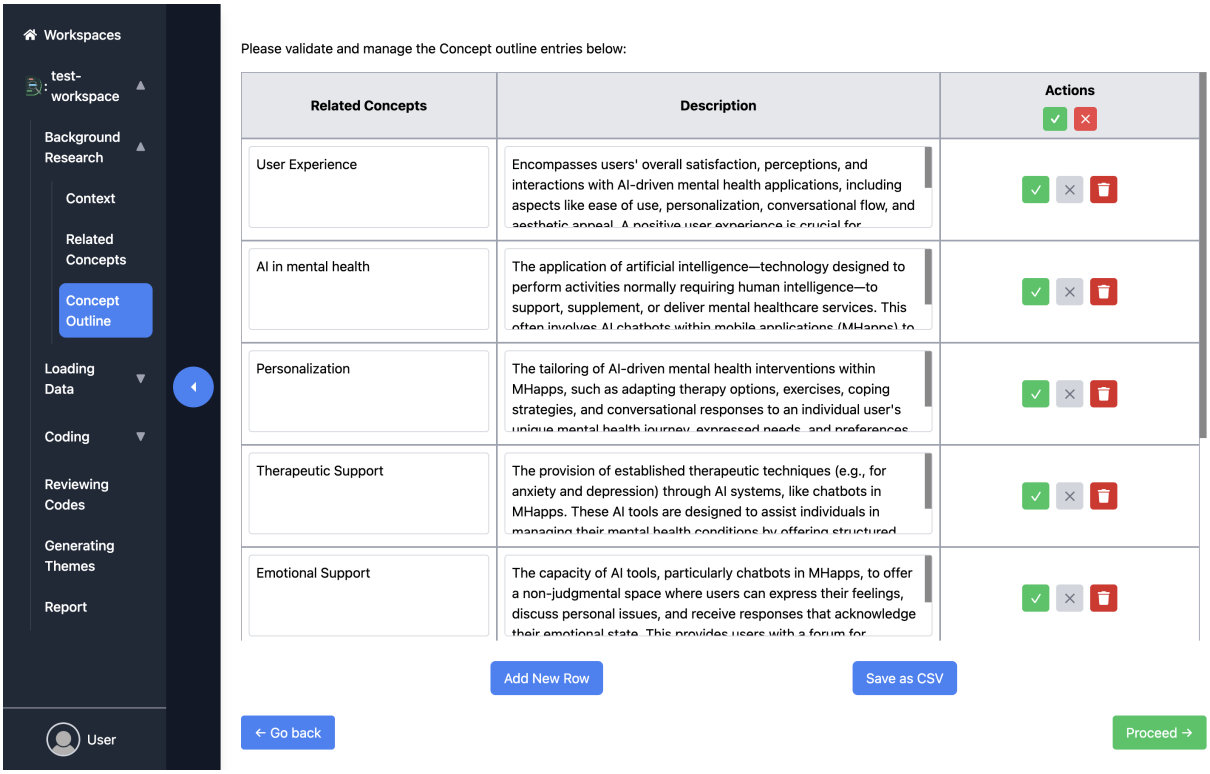


Figure 3.3: In the final step of the *Background Research* (Phase 1) of DeTAILS, the *Concept Outline* is presented as an editable table containing related concepts, and their descriptions. Researchers can edit, add, or delete entries using the Actions as needed. Once finalized, the table integrates with the core concept and research questions to form a *"memory snapshot"*. This snapshot guides future coding and theme generation and can be updated by revisiting and editing this page at any time. The table can be downloaded in CSV format for sharing. Clicking the *Proceed* button finalizes the outline and creates the memory snapshot in the backend.

the main concept under investigation, articulate specific research questions, and upload background literature. This literature is processed and indexed into a ChromaDB vector database, facilitating Retrieval-Augmented Generation (RAG) so that later LLM responses are grounded in theories from the user's sources. Building on this foundation, the LLM suggests related concepts based on the provided context along with its extensive training data, while interactive options such as 'Redo with feedback' and direct editing allow researchers to refine these suggestions (Figure 3.2).

The LLM then assists in structuring the background research by generating a 'Concept Outline' table that itemizes each concept, and offers working definitions for data segments during coding (Figure 3.3). Researchers retain full authority to modify this outline, thereby ensuring that the conceptual boundaries precisely match their objectives. Ultimately, these steps culminate in a detailed "memory snapshot" that encapsulates the research context, questions, and conceptual outline, serving as a consistent reference point throughout the analysis. The 'memory snapshot' grounds all subsequent coding in researcher expertise, making LLM-generated codes contextually anchored — an approach that deepens reflexivity beyond what manual TA alone can achieve.

3.2 Loading Data

DeTAILS is designed to analyze Reddit text data, including posts and their associated comments, and offers two flexible methods for data loading. The first, 'Torrents' via Academic Torrents [112, 125], lets users select a subreddit and date range; DeTAILS then uses the user's Transmission BitTorrent client to download public Reddit archives (such as those from Pushshift.io) and automatically retrieves the relevant content. Alternatively, the 'Upload' method allows users to import pre-collected Reddit data, provided it matches the format used by the PASS tool [66]. Once loaded, the dataset can be finalized using filters, such as adjusting date ranges, removing empty posts, or applying keyword searches, to select a precise analysis corpus. When the dataset is ready, users can 'lock' it to ensure consistency throughout the coding process.

Automated ingestion and filtering of Reddit archives shifts effort from dataset assembly to interpretation, contrasting with manual workflows where data preparation often consumes researchers' time.

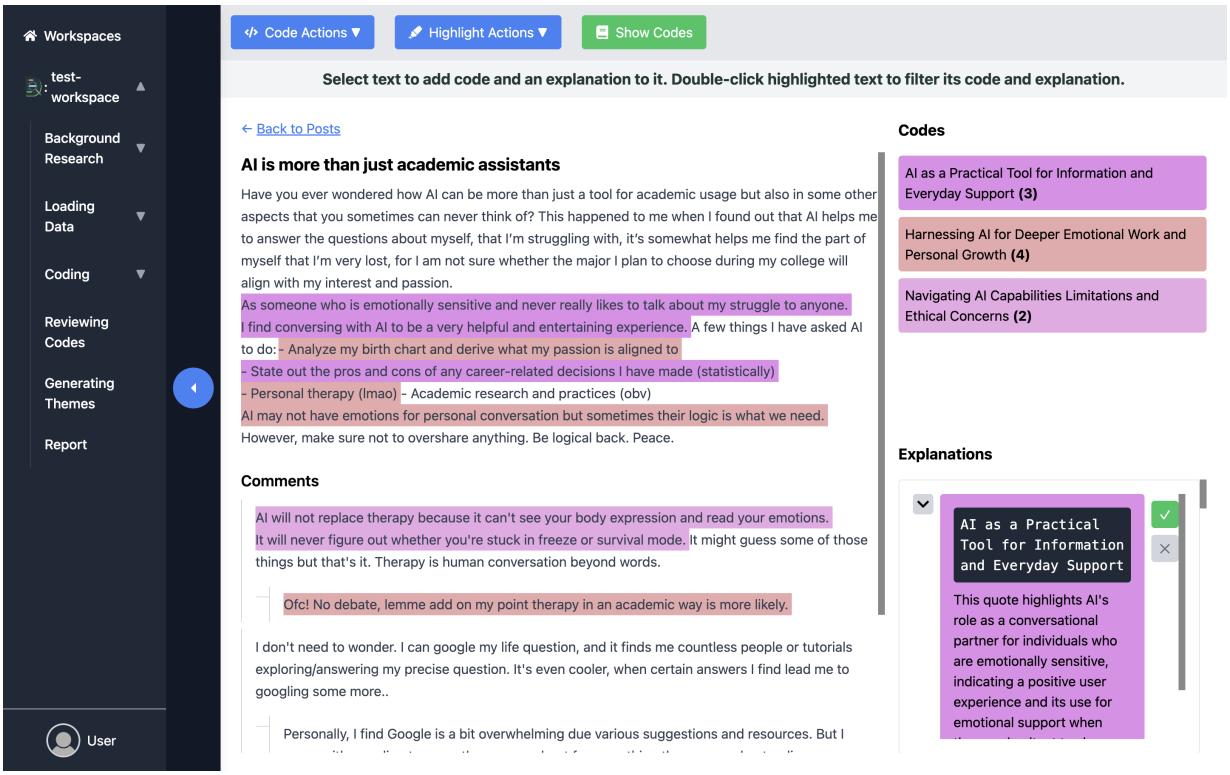


Figure 3.4: In DeTAILS' *Coding* (Phase 3), the *Edit Mode* lets researchers inspect a single post in its entirety — including nested comment threads — while iteratively refining the LLM-generated annotations. Action buttons at the top (*Code Actions*, *Highlight Actions*, and *Show Codes*) allow users to add, merge, or delete codes, adjust quote boundaries, and momentarily surface the full quote-code-explanation table for cross-reference. Within the transcript, colour-coded highlights mark every quote that has been linked to a code; double-clicking a quote filters the right-hand panels to the associated quotes. The *Codes* column lists every code already applied to this post, together with a frequency badge that shows how many quotes carry that label. Below it, stacked *Explanation* cards justify each quote-code pairing; researchers can endorse an explanation (✓) or open a chat-modal (×) to challenge the assignment, negotiate alternatives, and overwrite the code or explanation as desired. A persistent 'Back to Posts' link returns to the review table, where all edits are immediately reflected across the dataset. This granular, chat-augmented view keeps the researcher in control while using the LLM as a transparent coder

.

3.3 Coding

In the initial coding phase, DeTAILS begins by processing each post. It sends the full transcript along with the background research memory snapshot to the LLM. The model then identifies relevant quotes, assigns descriptive codes, and provides a brief explanation for each decision. These outputs are stored per post and undergo validation through regex and text matching to avoid LLM hallucinations and to ensure that the quotes actually appear in the transcript; verified quotes are subsequently highlighted in the user interface.

The LLM then periodically groups similar codes under consistent labels, streamlining the review process for the researcher. Users interact with the system through two primary views: a 'Review Mode', which offers a tabular summary of posts, quotes, codes, and explanations for quick assessment, and an 'Edit Mode', which displays the full transcript alongside highlighted quotes and color-coordinated code and explanations (Figure 3.4). In 'Edit Mode,' researchers can engage in chat-like interactions with the LLM for clarifications, make manual edits, adjust quote boundaries, or even trigger a complete 'Redo with feedback'. When invoked, this option creates a temporary state codebook based on current explanations and user feedback which guides the LLM in recoding the posts.

After reviewing and refining the codes from the first data split, the LLM unifies these into a concise codebook. This document captures the essence of each code based on its application in the data, and is presented for further review and manual adjustment by the researcher. In the subsequent processing of the second data split, the initial codebook is provided along with each post's transcript and background research memory snapshot. Here, the LLM applies the existing codes for consistency, yet still retains the flexibility to introduce new codes if necessary, all while maintaining the rigorous verification and interactive editing features established in the initial phase.

The human-LLM coding loop rapidly generates and verifies descriptive and conceptual codes at scale, acting as a consistent 'second coder' and freeing researchers from the tedium of first-pass labeling.

3.4 Reviewing Codes

After coding all posts, DeTAILS moves into a higher-level review and organization phase. In this stage, all codes and their corresponding explanations are batched and sent to the LLM for clustering. The model groups semantically similar codes and assigns each cluster a consolidated 'reviewed code'. These clusters are then visually presented via buckets, where each of them represents a reviewed code and contains blocks for the individual

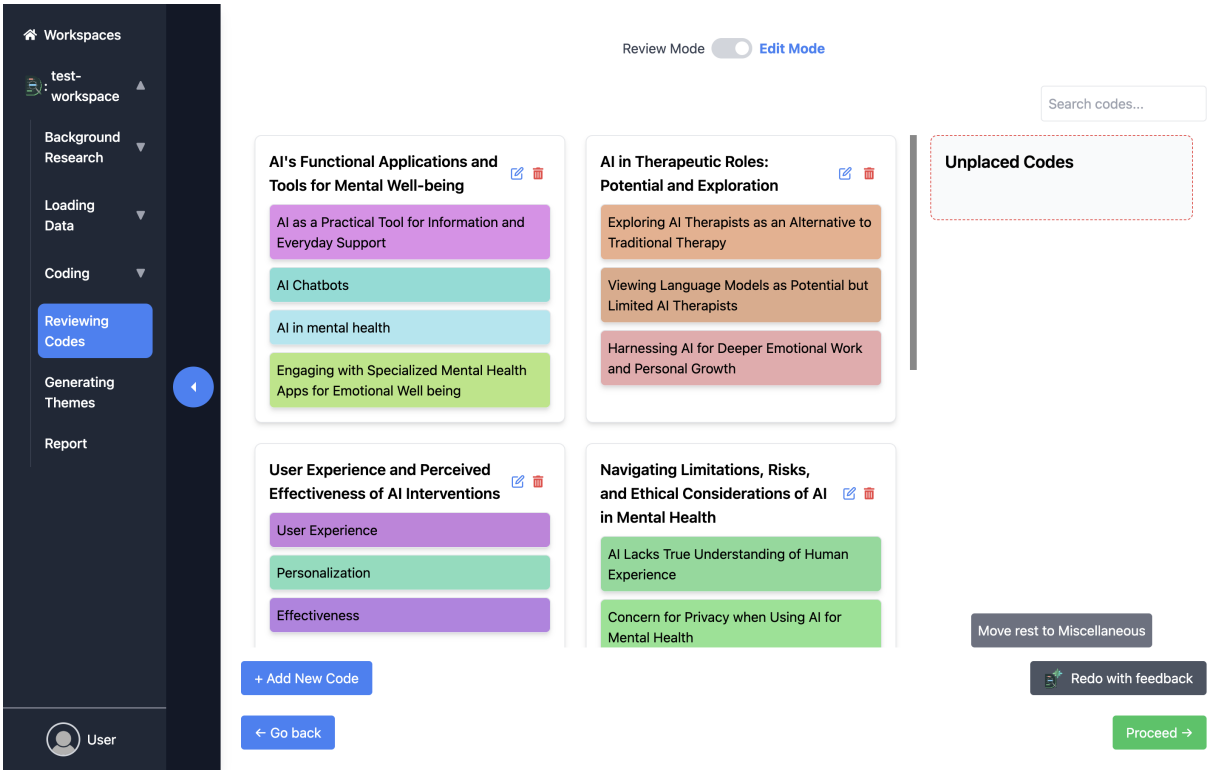


Figure 3.5: During the *Reviewing Codes* (Phase 4), DeTAILS' *Edit View* presents the LLM-generated clusters that merge similar codes into higher-order *reviewed codes*. Each 'bucket' is headed by the reviewed-code label and contains the coloured blocks representing the underlying codes imported from the previous coding step. Researchers can curate these clusters by dragging blocks between buckets, renaming buckets, deleting them, or creating new ones. Codes that have not yet been placed accumulate in an *Unplaced Codes* bucket on the right. A *Redo with feedback* button launches a feedback modal, enabling users to describe desired changes and have the LLM accordingly regroup the codes. The toggle at the top switches back to *Review Mode*, where each reviewed code and code is shown in a table format with its supporting quotes and explanations, letting the researcher audit and iterate until satisfied before clicking *Proceed*.

codes, clearly illustrating their interrelationships (Figure 3.5). Researchers can edit the visualized clusters using manual organization tools (such as drag-and-drop functionality to reposition code blocks or create, rename, and delete buckets) and further fine-tune the grouping by triggering an LLM refinement through a 'Redo with feedback' function based on user feedback. Once the reviewed code groupings are finalized, an alternate view displays these codes alongside their associated original quotes and explanations, maintaining a clear link to the source data.

LLM-driven clustering surfaces latent semantic relationships among codes, enriching the depth of analysis by revealing patterns that manual grouping alone might overlook.

3.5 Generating Themes

The final analytical step involves generating overarching themes from the reviewed codes. DeTAILS prepares a final codebook memory snapshot including all reviewed codes, their definitions, and representative examples (quotes) and provides this, along with the list of unique reviewed codes (this is sent twice — once as part of the table and once as a separate list — to avoid LLM hallucinations), to the LLM. The model then performs a clustering task to group these reviewed codes into potential themes capturing major patterns in the data.

Similar to the Reviewing Codes phase, these suggested themes are presented visually using an interactive 'bucket' interface, where each theme bucket contains the reviewed codes encompassed within it (Figure 3.6). The researcher retains full control over this final structure through direct editing — dragging reviewed codes between themes, creating new themes, merging or renaming them — ensuring the final themes accurately reflect their interpretation. An LLM refinement option, using a 'Redo with feedback' function, is also available to assist in exploring alternative thematic structures.

AI-assisted theme generation proposes high-level patterns for refinement, combining LLM breadth (scale and consistency) with human insights (theoretical framing), a synergy unavailable in fully manual mapping.

3.6 Report

Once themes are finalized, DeTAILS assists in compiling and exporting the findings. The report interface provides flexibility through toggles for 'Detailed' versus 'Summary' views

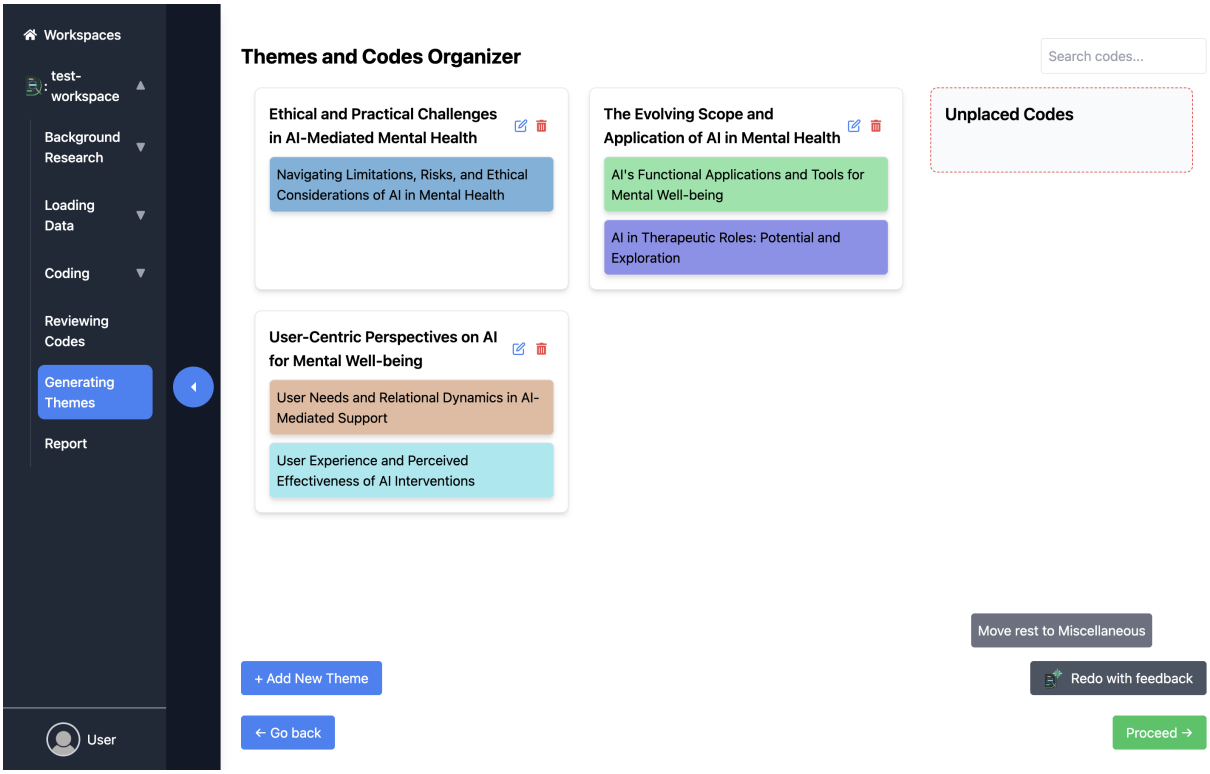


Figure 3.6: In the *Generating Themes* (Phase 5), DeTAILS elevates the analysis from codes to overarching patterns (themes). The interface shows 'buckets' headed by their proposed theme title; within it, coloured blocks list the *reviewed codes* that the LLM has clustered together. Researchers can reshape the thematic structure by dragging blocks between theme buckets, creating new themes or deleting/editing them. Any codes that have not yet been assigned accumulate in an *Unplaced Codes* tray on the right. Selecting *'Redo with feedback'* opens a feedback modal so the LLM can suggest a new grouping based on the researchers' guidance. Once satisfied, clicking *Proceed* locks the themes and passes them to the final report stage. (Note: The initial number of themes generated increases with the number and diversity of the reviewed codes, but researchers can specify their desired number of themes through feedback).

and options to organize the report, for instance, by 'Theme & Code' or 'Post by Post'. Unlike most current tools that limit data downloads to their proprietary formats, DeTAILS enables exports of the final, structured report in the widely compatible CSV format, suitable for sharing, further analysis in other tools, or direct incorporation into publications.

Although the final write-up of the analysis still remain a manual process, DeTAILS generates a structured report — segmenting each result into Code-Quote-Explanation sections — so that researchers can quickly review key insights and focus their efforts on in-depth analysis.

3.7 Technical Implementation

DeTAILS is implemented as a self-contained, cross-platform desktop application that tightly integrates a front-end interface with a back-end analysis engine. The frontend is built using Electron [39] (with a React [41] UI), which allows the toolkit to run on Windows, macOS, and Linux without additional setup. The Electron container bundles the Python backend, vector database (ChromaDB [118]), and local language model runtime (Ollama [93]) with the goal of enabling researchers to install and run DeTAILS easily as a single application.

This architecture provides a clear separation of concerns: frontend handles user interactions and visualization, while the back-end performs data handling, LLM prompt generation, and computation.

The backend uses FastAPI [98], a high-performance asynchronous web framework in Python. We deploy the FastAPI server with multiple Uvicorn [28] worker processes (three for handling HTTP requests and one dedicated to WebSocket communication) to fully utilize multi-core processors and handle concurrent operations without blocking the user interface. FastAPI was chosen for its native async/await support and robust WebSocket integration, allowing us to serve both RESTful API calls and real-time event streams within one cohesive server. For example, as the LLM processes data or indexes are built, the back-end streams progress messages to the front-end via a WebSocket connection. All source code is available at DeTAILS GitHub repository [108].

3.7.1 Services & Data Stores

All data in DeTAILS is stored locally to maintain privacy and support offline use. We employ SQLite [60] in write-ahead logging mode as our embedded relational store for analysis artifacts — Reddit posts, comments, codes, codebooks, themes, and application

state — because its file-based design imposes no external dependencies and avoids the packaging conflicts of client–server engines under PyInstaller [97]. Write-ahead logging decouples readers and writers, reducing lock contention when multiple back-end workers and the UI access the database concurrently.

3.7.2 LLM Providers

Integrating large language models into the toolkit required careful design to support multiple model providers and to ensure consistent, interpretable outputs. DeTAILS addresses this with a modular LLM orchestration layer built on LangChain [22], which allows us to interface with various LLM APIs (e.g., OpenAI's models, Google's models via Vertex AI [52]) as well as local models via Ollama. We implemented a provider-factory pattern so that each LLM backend is encapsulated in a provider class with a standardized interface. This abstraction means the core logic of DeTAILS can remain unchanged regardless of which model is in use — adding a new model (for example, an Anthropic model or any Azure-hosted model) only requires writing a small provider subclass that knows how to call that API. All providers register with the factory, and the user's choice of model (e.g., "OpenAI-GPT4" or "Ollama-Llama3.2") is resolved to the appropriate provider at runtime.

In our current implementation, four providers are supported: OpenAI, Google's Vertex AI, the local Ollama server (which can run models such as LLaMA or Gemma models on the users' hardware), and Google's AI Studio.

3.7.3 Prompts

Each phase of the analysis workflow is powered by carefully engineered prompt structures that guide the LLM's behavior and output format (Table 3.1). We created distinct prompt scripts for every major task: related concepts, concept outline, initial coding, initial codebook, final coding, reviewing codes, and theme generation. These prompts provide the model with role-specific instructions and a consistent structure for responses. For example, in the Background Research phase, the system prompt reminds the model of the main topic and research questions, and asks for a JSON list of exactly 10 related concepts (each 1–3 words long). In the Coding phase, the prompt template includes the full text of a Reddit post (with all comments) and instructs the model to extract meaningful quotes and label them with a short code and an explanation, all formatted as a JSON array of objects. We enforce a strict JSON schema in the prompts by explicitly showing an example JSON

structure in the user prompt and ending the instruction with a reminder like "Return only a JSON object with the following fields... no extra text." The model's output is then parsed by our back-end using a regex that captures the JSON block.

Each prompt unfolds as an ordered five-stage scaffold chain-of-thought, beginning with Role and Expertise, which primes the model to adopt an appropriate persona — be it "Expert qualitative researcher" or "Specialist in Braun & Clarke's six-phase thematic analysis" — to maintain the right depth, tone, and methodological rigour from the outset. It then moves to Scope and Task Objective, where the exact analytic goal is named (for example, "Extract ten 1–3-word concepts" or "Group lower-level codes into higher-order code clusters"), sharply focusing the model's attention and preventing drift into tangential or overly creative responses. Next comes Context / Input Data, wherein the relevant materials — research questions, transcripts, literature summaries, prior code lists — are injected so that all output remains firmly grounded in the users' data rather than generic knowledge. For more intricate activities, a Step-by-Step Process Guidance section follows, laying out the methodological workflow (scan for salient ideas, extract verbatim quotes, label codes, define clusters) to guide the model through a logical, human-inspired progression rather than asking it to invent its own path.

Throughout the iterative 'redo' cycles, DeTAILS uses specialized prompt variants to incorporate user feedback into the LLMs' next response. When a researcher isn't satisfied with the LLMs' suggestion — say the grouping of codes into a theme — they can provide written feedback or select items to keep/discard. The toolkit then formulates a follow-up prompt: it includes the users' feedback (e.g., "the last grouping merged two concepts that should stay separate") and often the previous output, and asks the model to regenerate a revised proposal. We maintain separate prompt templates for these refinement steps. This iterative approach lets the model know why the user was dissatisfied and what adjustments are expected, essentially creating a dialogue where the model acts on critique. By chaining prompts in this manner, the system can zero in on results that meet the researcher's needs without the user having to do all adjustments manually. It's an example of human-LLM collaboration: the human gives high-level direction, and the LLM does the heavy lifting of reorganizing or rephrasing content accordingly. All prompts used in DeTAILS can be found in the Appendix A.

3.7.4 Verification & Consistency

Finally, we incorporated several measures to ensure the reliability and correctness of the LLM-generated outputs. DeTAILS verifies every quote and code that the model produces, to avoid hallucinations. After an LLM call (for coding or theme suggestions), the back-end

cross-checks each quoted text segment against the original Reddit post to confirm that the segment appears verbatim. Any quote that cannot be found in the source text is discarded as a likely hallucination.

Section	Context	Coding	Codebook	Clustering
Role & Expertise	Expert qualitative researcher	Expert qualitative researcher, specializing in Braun–Clarke's six-phase thematic analysis	Expert qualitative researcher	Expert qualitative researcher, specializing in Braun–Clarke's six-phase thematic analysis
Task Objective	Extract N 1–3-word concepts	Label transcript passages with evidence-based codes	Define each unique code	Group lower-level codes into higher-order clusters or themes
Context & Input Data	• Main topic • Research questions • User-uploaded background literature	• Transcript • Main topic • Research questions • Concept outline	• Codes and their summarized explanations	• List of codes • (Optional) Table of codes and summaries
Process Guidance	1. Map concepts to your research questions. 2. Identify N 1–3-word concepts. 3. Check for relevance, diversity and uniqueness.	1. Read transcript for context 2. Extract exact quotes 3. Assign concise labels	1. Identify patterns across explanations 2. Generate definitions	1. Examine each code's meaning 2. Merge semantically related codes 3. Assign descriptive cluster names/themes
Output Format	{ "concepts": ["c1", "c2", "c3"] }	{ "codes": [{"quote": "...", "explanation": "...", "code": "...", "source": {...} }] } or {"codes": []}	{ "code1": "def1", "code2": "def2" }	{ "higher_level_codes": [{"name": "...", "codes": [...] }] } or { "themes": [{"name": "...", "codes": [...] }] }

Table 3.1: Prompt Structure Across Multiple Phases of DeTAILS: Context, Coding, Codebook and Clustering with each prompt containing Role & Expertise, Task Objective, Context & Input Data, Process Guidance, and Output Format

Chapter 4

Evaluation: Usability Study

To understand how DeTAILS can support qualitative researchers in a realistic thematic analysis scenario, we conducted a participant study with qualitative analysts performing reflexive thematic analysis on their own data. This evaluation focused on how participants interacted with the tool, the workflow's usability, and the tool's support for their research tasks.

All participants provided informed consent, and the study has undergone a thorough review process and received ethics clearance from the University of Waterloo Research Ethics Board (REB #47221).

4.1 Participants and Recruitment

We recruited 18 qualitative researchers with varying levels of experience and disciplinary backgrounds in reflexive thematic analysis (see Table 4.1). To capture a range of perspectives, we used purposive sampling with inclusion criteria of prior involvement in thematic analysis and qualitative coding. Participants were categorized by TA experience as novice (less than 1 year experience, n=6), proficient (1–4 years experience, n=6), or expert (more than 4 years experience, n=6). This stratification ensured feedback from both newcomers and seasoned practitioners in TA.

Participants were recruited via direct email invitations using academic and professional networks. Prior to their session, each participant was asked to send their research questions, 2–5 representative research papers related to the questions, and the name of a subreddit they were familiar with and interested in analyzing.

4.2 Materials and Apparatus

We focused on Reddit data given its heterogeneity and scale [96], which has led to its use in numerous HCI studies (e.g., [100, 21, 63, 79]). Reddit hosts millions of topical sub-communities (subreddits) encompassing diverse domains and vast amounts of participant-generated text. Working with subreddit data allowed us to evaluate DeTAILS on large-scale, noisy real-world text, while also enabling each participant to explore content relevant to their interests. For each participant's chosen subreddit, we downloaded the archived Reddit data using DeTAILS. This approach ensured consistency in data format and took advantage of an existing open dataset for social media research.

In preparation for each session, we pre-loaded the participant's chosen subreddit data into DeTAILS. For each participant, we downloaded all posts and comments from the last full month of their chosen subreddit [125] and processed them through the DeTAILS pipeline before the session, so we could present a full-scale analysis to them at the end of their evaluation. Participants then started the analysis during their session in a new workspace.

Sessions were conducted using a MacBook Air with an Apple M1 chip (8-core CPU, 7-core GPU) and 16 GB RAM. This machine ran the DeTAILS application with the gemini-2.5-pro-preview-03-25 LLM via the Google Vertex AI API. Participants who attended in person used this laptop directly. For remote participants, we utilized Microsoft Teams' screen-sharing and remote control functionality: the author shared their screen running DeTAILS, and the participant was granted control to operate the interface virtually. This setup allowed online participants to fully interact with the toolkit (e.g., clicking buttons, highlighting text, editing code labels) as if they were seated at the machine. Audio and video were streamed via Teams for communication. Two participants opted for in-person sessions in our lab, while the remaining 16 participated remotely via Teams.

4.3 Study Procedure

We conducted each evaluation session as a single 2 to 2.25 hour engagement (with an optional 10 minute break in between), structured into several phases: an initial interview, a tutorial demonstration, an interactive analysis using DeTAILS (the main usability test), a follow-up exploration of large-scale results, and post-study surveys. Throughout all phases, we encouraged participants to think aloud, i.e. to verbalize their thoughts, expectations, and reasoning as they worked. We also employed a semi-structured interview approach during and after each task to probe specific aspects of their experience. The researcher

was present to prompt participants with questions from an interview guide and to answer any questions the participants had. While participants conducted their thematic analysis using DeTAILS, the interviewer remained neutral—refraining from altering their judgments or asking leading questions—only stepping in to clarify the tool's functionality or study questions when necessary

- **Introduction & Background of Participant (10 min):** At the start of the session, we collected background information through a brief interview (Appendix B.2). This initial conversation established each participant's context and helped us understand their baseline practices and needs.

- **Tutorial Demonstration (10 min):** Next, we provided a guided walkthrough of the DeTAILS interface and workflow. The interviewer demonstrated each step of the toolkit using an example dataset. By the end of this walkthrough, they had a basic understanding of how to operate DeTAILS and what to expect at each step. This training phase was important to mitigate any learning curve effects once the actual analysis began.

- **Interactive Thematic Analysis Task (90 min):** The core of the study was having participants use DeTAILS to perform reflexive thematic analysis on 30 discussion post transcripts from their chosen subreddit (with each transcript consisting of an original post and its comments). This subset size of 30 was chosen to keep interactive analysis feasible within the session time while providing enough data to generate meaningful codes and themes. We asked participants to approach this task as if it were part of their own research. The task was divided into the same sequence of steps as the toolkit's workflow, with the participant in control and the researcher observing and prompting discussion (Appendix B.2). Participants were also free to ask any clarifying questions about the tool during the study.

  By the end of the analysis task, each participant had experienced the full lifecycle of using DeTAILS on a dataset relevant to their own work. This provided a grounded basis for them to evaluate the tool's usability and usefulness.

- **Large-Scale Dataset Review (10 min):** After completing the interactive analysis on the 30-post subset, we transitioned to showing the participant the pre-computed large-scale results for the full month of subreddit data (December 2024) that we had prepared earlier. We loaded the test workspace in DeTAILS and allowed the participant to freely browse and inspect each stage's output for the larger dataset. We also gathered any additional feedback at this stage: participants could comment

on the overall workflow now that they had seen it operate on two scales, and suggest improvements or features.

- **Post-Study Surveys (5 min):** Finally, participants completed three standardized questionnaires to quantify their experience with DeTAILS. First, NASA Task Load Index (NASA-TLX) [55] to assess the perceived cognitive workload of using the toolkit. NASA-TLX is a widely used multidimensional scale that captures a participant's mental, physical, and temporal demands, performance effort, and frustration while performing a task. Second, participants filled out the AttrakDiff [56] questionnaire, which measures their experience along pragmatic and hedonic dimensions. AttrakDiff uses semantic differentials (pairs of opposite adjectives) to rate qualities such as the tool's efficiency, clarity, stimulation, and novelty. This allowed us to gauge both the usability (pragmatic quality) of DeTAILS and its participant engagement (hedonic quality) in a systematic way. Third, we gave a Perceived Usefulness Scale [32]. Perceived usefulness reflects the degree to which a person believes that using a system will enhance their job performance. This was aimed at understanding whether participants felt the tool would be beneficial and worth adopting in their own research workflow or not. Together, these three surveys provided a quantitative profile of workload, participant experience, and utility for DeTAILS from the participants' perspective.

Throughout the study procedure, we strove to maintain a balance between observing DeTAILS usage (to uncover usability issues or effective features) and gathering participants' expert judgments on the tool's support for qualitative analysis.

4.4 Data Collection and Analysis

To evaluate DeTAILS, we collected and analyzed both quantitative and qualitative data. Quantitative data included metrics derived from system logs that captured participant interactions on analysis tasks, as well as scores from three post-study surveys. Qualitative data consisted of recordings from think-aloud protocols and semi-structured interviews conducted throughout the study sessions.

We analyzed two data sources: logs of user actions in the database and subjective surveys administered after the study. The system logs contained a record of every user interaction, such as time taken at each phase, additions, edits, deletions, and the acceptance or rejection of system suggestions. These logs enabled the quantification of metrics

for interactions performed in DeTAILS. The subjective questionnaires were the NASA-TLX, AttrakDiff, and Perceived Usefulness. In addition to these quantitative sources, we also collected qualitative data through interview transcripts obtained by transcribing audio recordings of each session, which included think-aloud protocols and semi-structured interviews.

For metrics across actions performed in DeTAILS (like accepting/editing/deleting concepts/codes/themes), we computed precision, recall, and F1 scores using cosine similarity for weighting. Weighted versions of precision, recall, and F1 scores were applied to Related Concepts, Concept Outline, Initial Coding, and Global Coding to value edits. For the Initial Codebook, we used cosine similarity alone to measure semantic similarities between definitions. For Reviewing Codes and Generating Themes, we used macro precision, recall, and F1 scores.

To analyze time spent across the different study phases, we conducted a one-way repeated-measures ANOVA (RM ANOVA). We assessed the assumption of sphericity using Mauchly's test; if this assumption was violated, a Greenhouse-Geisser correction was applied. The non-parametric Friedman test was used as an alternative if the assumptions for RM ANOVA were not met. Significant omnibus tests were followed by post-hoc comparisons using Wilcoxon signed-rank tests with a Bonferroni correction to identify specific differences between phases. To compare the three expertise groups (Novices, Proficients, and Experts), we performed more inferential statistical tests. Assumption checks—Shapiro-Wilk for normality and Levene's test for homogeneity of variance—guided inferential test selection. If data were normally distributed ($p > 0.05$) with equal variances ($p > 0.05$), a one-way ANOVA was used; if variances were unequal, Welch's ANOVA was applied; and if normality was violated, the Kruskal-Wallis test was used. Pairwise comparisons used t-tests, Welch's t-tests, or Mann-Whitney U tests based on the same assumptions, with a Bonferroni-adjusted $\alpha = 0.0167$ for multiple comparisons.

The NASA-TLX survey was administered on a 0–210 scale (later scaled down to 0–100), with each of its six subscales rated on a 21-point scale. The AttrakDiff questionnaire used 7-point bipolar semantic differential scales ranging from 1 to 7 across all of its 28 dimensions. Perceived Usefulness questionnaire items (6 items) were measured on a 5-point Likert scale from 1 (strongly disagree) to 5 (strongly agree).

Both the survey data and the calculated performance metrics were statistically analyzed. Specifically, the F1 scores from the Related Concepts, Concept Outline, Initial Coding, Global Coding, Reviewing Codes, and Generating Themes phases, along with all survey scores, were analyzed using means, standard deviations, and medians. To compare the three expertise groups (Novices, Proficients, and Experts), we performed inferential statistical tests. Assumption checks—Shapiro-Wilk for normality and Levene's test for ho-

mogeneity of variance—guided inferential test selection. If data were normally distributed ($p > 0.05$) with equal variances ($p > 0.05$), a one-way ANOVA was used; if variances were unequal, Welch's ANOVA was applied; and if normality was violated, the Kruskal-Wallis test was used. Pairwise comparisons used t-tests, Welch's t-tests, or Mann-Whitney U tests based on the same assumptions, with a Bonferroni-adjusted $\alpha = 0.0167$ for multiple comparisons.

For the qualitative data, session recordings were transcribed verbatim. We then performed a reflexive thematic analysis to identify key patterns in participants' reasoning, user experience, and feedback across TA phases using DeTAILS.

ID	TA Experience	Expertise Level	Field	Occupation	AI Used for TA	Subreddit	Start - End Date	Description
P1	¡1 year	Novice	XR Gaming	Master's Student	ChatGPT	r/artificial	06/07/2008 - 31/12/2024	A subreddit dedicated to news, research, and discussions on artificial intelligence and its applications.
P2	¡1 year	Novice	Explainable AI	PhD Candidate	ChatGPT	r/ChatGPTCoding	06/12/2022 - 31/12/2024	Community for ChatGPT code samples, usage tips, and creative AI-driven demos.
P3	1–2 years	Proficient	NLP Classification	Jr. Consultant	–	r/WeAreTheMusicMakers	09/09/2008 - 31/12/2024	Music production, composition techniques, and audio engineering discussions.
P4	2–3 years	Proficient	Human–AI Collaboration	PhD Candidate	Atlas.TI (AI features)	r/humanresources	29/05/2008 - 31/12/2024	HR best practices, talent management insights, and collaboration strategies.
P5	3–4 years	Proficient	Secure HCI	Master's Student	GPT-4	r/Chatbots	18/01/2009 - 31/12/2024	Chatbot development, conversational AI frameworks, and real-world use cases.
P6	5–6 years	Expert	Social Stigma	PhD Candidate	NVivo (AI features)	r/beyondthebump	29/04/2012 - 31/12/2024	Support for new parents navigating postpartum and early parenting challenges.
P7	¡1 year	Novice	NLP Classification	Post-doc Researcher	Multiple LLMs	r/ArtificialInteligence	06/05/2016 - 31/12/2024	AI research, ethics debates, and industry application discussions.
P8	¡1 year	Novice	Human–Robot Interaction	Post-doc Researcher	Multiple LLMs	r/CaregiverSupport	30/03/2013 - 31/12/2024	Advice and emotional support for family and professional caregivers.
P9	¡1 year	Novice	Youth Smoking	PhD Candidate	ChatGPT	r/Canadian_ecigarette	23/10/2013 - 31/12/2024	Vaping products, health impacts, and regulation updates in Canada.
P10	¡1 year	Novice	Healthcare Trust	Undergrad Student	–	r/AskBalkans	05/03/2019 - 31/12/2024	Q&A on Balkan culture, travel tips, and current affairs.
P11	4–5 years	Expert	Accessible Gaming	PhD Candidate	–	r/disability	12/03/2008 - 31/12/2024	News, resources, and perspectives for individuals with disabilities.
P12	4–5 years	Expert	Data Integration	PhD Candidate	–	r/bigdata	05/01/2011 - 31/12/2024	Big data storage, processing pipelines, and predictive analytics.
P13	3–4 years	Proficient	Health Systems	PhD Candidate	–	r/ImmigrationCanada	10/02/2013 - 31/12/2024	Immigration advice, visa procedures, and settlement experiences.
P14	1–2 years	Proficient	Haptic UX	Product Analyst	ChatGPT	r/userexperience	19/06/2018 - 30/12/2024	UX design methods, usability testing, and prototyping best practices.
P15	12–13 years	Expert	Public Health	PhD Candidate	–	r/blackladies	16/11/2012 - 31/12/2024	Safe space for Black women to share, support, and celebrate achievements.
P16	3–4 years	Proficient	VR Analytics	PhD Candidate	ChatGPT	r/virtualreality	28/08/2011 - 31/12/2024	VR hardware, software platforms, and immersive design discussions.
P17	10–11 years	Expert	Dementia Technology	PhD Candidate	ChatGPT	r/dementia	10/01/2010 - 31/12/2024	Resources and support for those affected by dementia and their caregivers.
P18	11–12 years	Expert	Human Factors	Assistant Professor	–	r/smartwatch	21/01/2013 - 31/12/2024	Smartwatch tech, app development, and wearable computing trends.

Table 4.1: Demographics of Study Participants: Years of Experience, TA Expertise Level, Field of Specialization, Current Occupations, AI used for TA, Subreddit details: Name, Start Date - End Date, Description

Chapter 5

Results

We now present the empirical findings from our user study evaluating DeTAILS through a mixed-methods approach involving 18 participants. Overall, the results demonstrate that DeTAILS fosters high alignment between LLM-generated outputs and user refinements, reduces cognitive demands of TA, and enhances perceived efficiency, with consistent support through distinct phases of TA.

5.1 Quantitative Results

Drawing from F1 scores, we assess the alignment between DeTAILS-generated outputs and participants' final refinements throughout the workflow phases, complemented by statistical analyses to examine differences by expertise. We also analyzed the time each participant spent in each phase using statistical tests. Additionally, we report on user workload via NASA-TLX scores, perceived usefulness scale, and hedonic/pragmatic qualities using the AttrakDiff questionnaire.

5.1.1 F1 Scores

F1 scores revealed increase in agreement within each phase (Figure 5.1). In the Background Research phase, initial Related Concepts suggestions achieved a mean weighted F1 of 0.86 (SD = 0.13), improving to 0.98 (SD = 0.03) in the Concept Outline step. This refinement extended to the Coding phase, where Initial Coding yielded a mean weighted F1 of 0.90 (SD = 0.17) across participants, rising to 0.97 (SD = 0.04) in Global Coding. Reviewing

Codes phase produced average macro F1 scores of 0.90 (SD = 0.20). The process peaked in Generating Themes phase, where all participants reached macro F1 = 1.00, reflecting full consensus on DeTAILS-proposed themes without modifications.

All omnibus tests failed to reject the null hypothesis ($p > 0.05$), indicating no overall differences in F1 scores by expertise. We assessed using the Shapiro-Wilk test for normality ($\alpha = 0.05$) and Levene's test for homoscedasticity. The Related Concepts phase satisfied both (Shapiro-Wilk $p > 0.05$; Levene's $p = 0.2440$), yielding a nonsignificant ANOVA: $F(2, 15) = 0.22$, $p = 0.8032$, $\eta^2 = 0.03$ (novice: $n = 6$, $M = 0.88$, $SD = 0.08$; proficient: $n = 6$, $M = 0.83$, $SD = 0.15$; expert: $n = 6$, $M = 0.87$, $SD = 0.17$). Other phases (Concept Outline, Initial Coding, Global Coding, Reviewing Codes) violated normality ($p < 0.05$) but met homoscedasticity (Levene's $p = 0.1519$–0.5960), with nonsignificant Kruskal-Wallis results: $H(2) = 1.95$–5.82, $p = 0.0544$–0.7664, $\eta^2 = 0.03$–0.34 (Generating Themes excluded for zero variance).

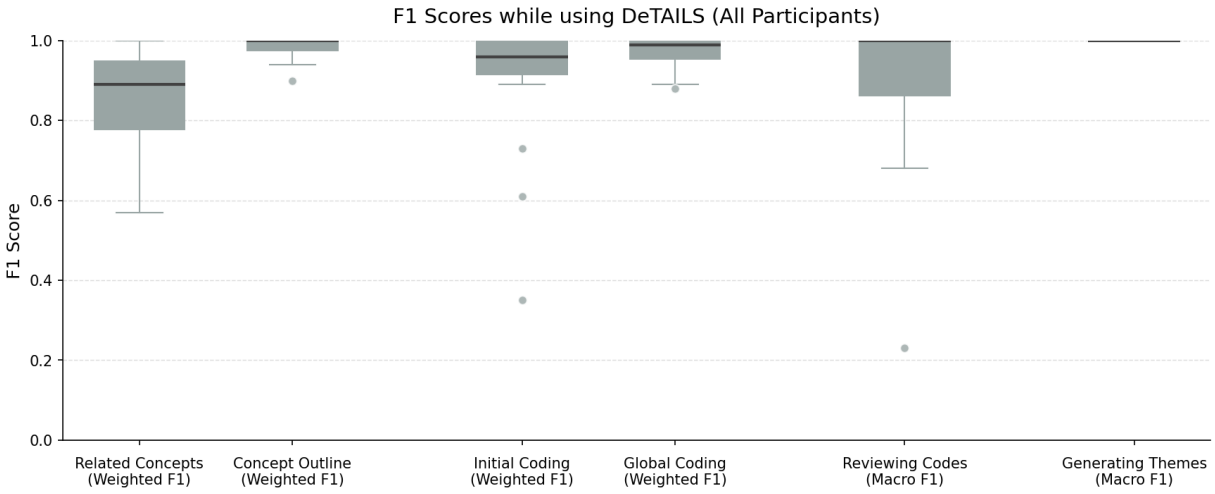


Figure 5.1: F1 Scores (All Participants) while using DeTAILS: Background Research Phase (Related Concepts, Concept Outline), Coding Phase (Initial Coding, Global Coding), Reviewing Codes Phase, Generating Themes Phase

5.1.2 Time Analysis

Our analysis revealed that the Initial Coding took significantly more time than any other steps (Figure 5.2). We assessed sphericity using Mauchly's test ($\alpha = 0.05$), which was significant, $W = 0.197$, $p = 0.003$, indicating a violation of sphericity. Accordingly, we

applied the Greenhouse–Geisser correction ($\varepsilon = 0.623$); the one-way repeated-measures ANOVA then revealed a robust effect of Phase on time spent, $F(2.49, 42.37) = 22.77$, $p < 0.001$, $\eta_p^2 = 0.57$. A nonparametric Friedman test corroborated this finding, $\chi^2(4) = 44.62$, $p < 0.001$. Post-hoc comparisons using Wilcoxon signed-rank tests with a Bonferroni correction showed that the Initial Coding step lasted significantly longer than Related Concepts ($p = 0.002$), Concept Outline ($p = 0.004$), Initial Codebook ($p < 0.001$), and Global Coding ($p = 0.004$). Global Coding itself was also significantly longer than Initial Codebook ($p = 0.004$). No other pairwise contrasts reached significance.

When comparing expertise groups, we generally found no significant differences in the time spent on each phase, with one key exception. The only statistically significant difference occurred during the Related Concepts phase, where a Welch's ANOVA (F(2, 9.26) = 6.51, p = 0.017) revealed that proficient users (M=5.37 min) spent significantly more time than novices (M = 2.49 min, t(10) = -3.74, p = 0.004). No other statistically significant differences in time spent between expertise groups were found in any of the other phases. These comparisons were made using one-way ANOVAs or their non-parametric equivalents after checking assumptions for normality (Shapiro-Wilk test) and homogeneity of variances (Levene's test).

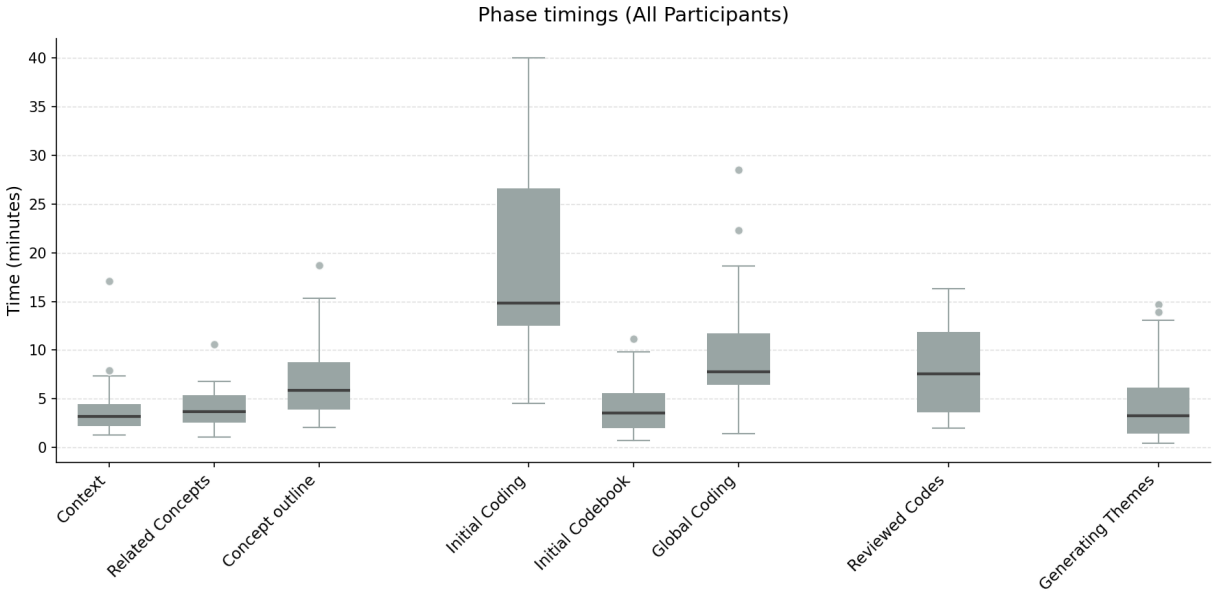


Figure 5.2: Participant timings (by phase) while using DeTAILS: Background Research Phase (Context, Related Concepts, Concept Outline), Coding Phase (Initial Coding, Global Coding), Reviewing Codes Phase, Generating Themes Phase

5.1.3 NASA-TLX

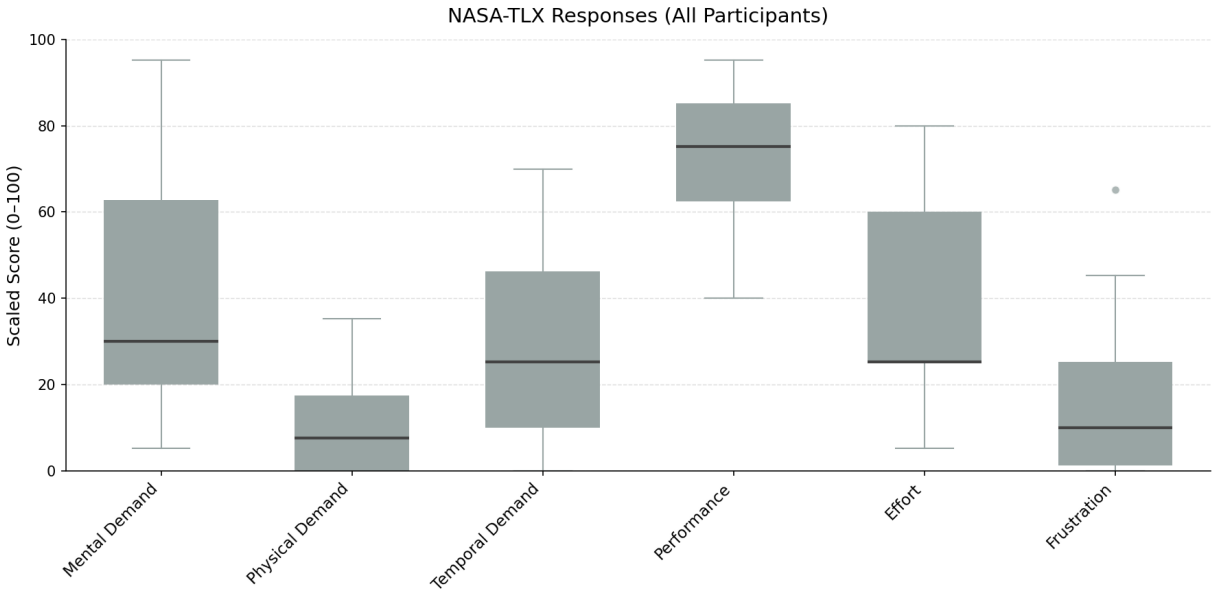


Figure 5.3: NASA-TLX Responses (All Participants) after using DeTAILS. It measures mental, physical, and temporal demands, performance, effort, and frustration through participants' self-reports.

Participants reported low to moderate cognitive load when using DeTAILS (Figure 5.3), with mental demand averaging 39/100, effort at 37/100, and temporal demand at 29/100. Frustration was notably low (median 10/100, with 89% scoring $\leq$30/100) paired with high self-assessed performance (mean 73/100, median 75/100). Overall workload averaged 26.3/100 (SD = 12.4), showing low overall perceived workload across participants.

Statistical analysis revealed no significant differences in workload across expertise levels. We applied appropriate tests based on distributional assumptions (Shapiro-Wilk test and Levene's test): one-way ANOVA for Mental Demand ($F(2, 15) = 0.94$, $p = 0.4123$, $\eta^2 = 0.11$), Temporal Demand ($F(2, 15) = 0.22$, $p = 0.8064$, $\eta^2 = 0.03$), and Effort ($F(2, 15) = 0.12$, $p = 0.8919$, $\eta^2 = 0.02$); Welch's ANOVA for Performance ($F(2, 9.32) = 0.43$, $p = 0.6652$, $\eta^2 = 0.07$); and Kruskal-Wallis tests for Physical Demand ($H(2) = 1.08$, $p = 0.5837$, $\eta^2 = 0.06$) and Frustration ($H(2) = 3.23$, $p = 0.1993$, $\eta^2 = 0.19$).

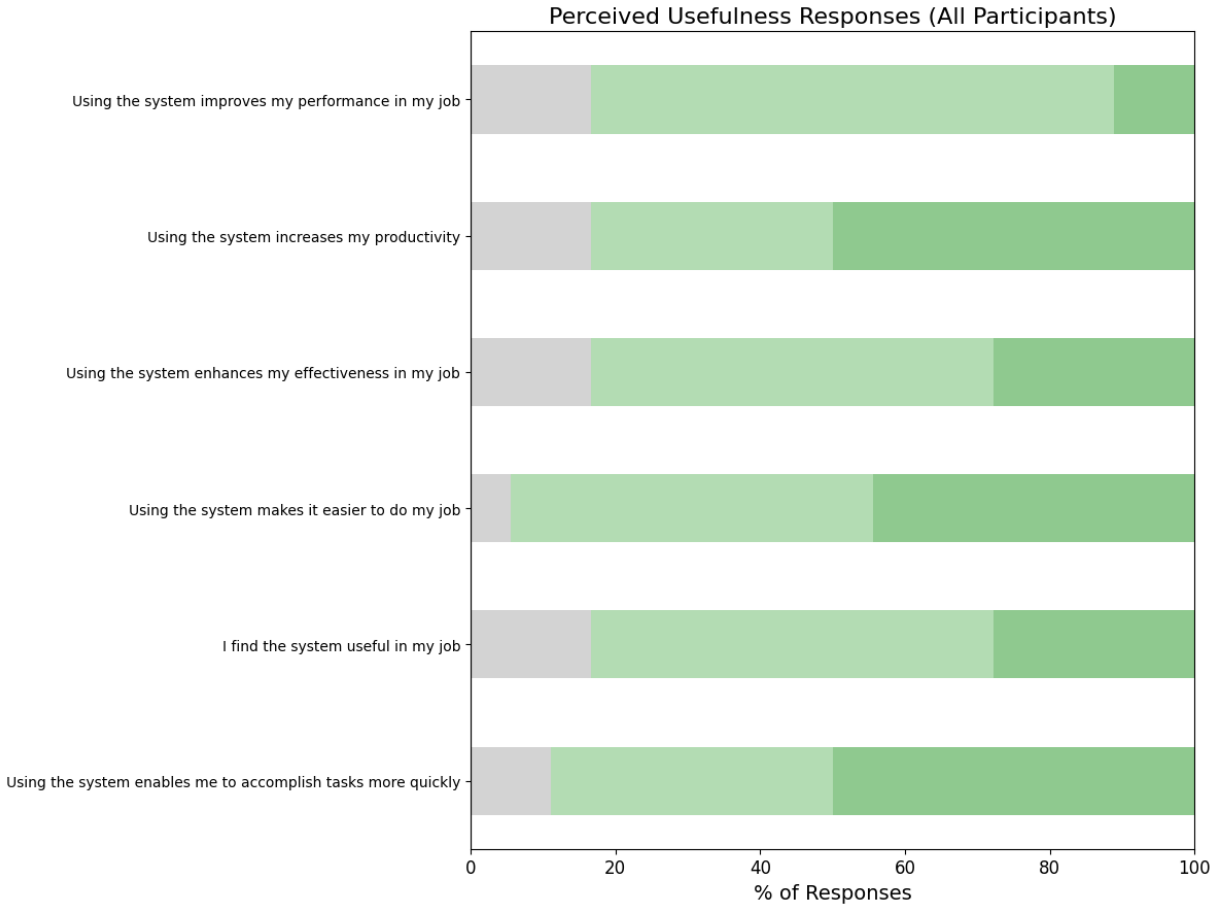


Figure 5.4: Perceived Usefulness Responses (All Participants) after using DeTAILS. It measures "Using the system improves my performance in my job," "Using the system increases my productivity," "Using the system enhances my effectiveness in my job," "Using the system makes it easier to do my job," "I find the system useful in my job," and "Using the system enables me to accomplish tasks more quickly" through participants' self-reports.

5.1.4 Perceived Usefulness Scale

Participants exhibited highly positive attitudes toward the utility of DeTAILS in their work, with the Perceived Usefulness Scale yielding an average score of 4.21/5, between "very useful" and "extremely useful" (Figure 5.4). No ratings fell below 3 ("moderately useful"), and 86% of responses were 4 or 5, indicating strong agreement. Prominent metrics included "Using the system enables me to accomplish tasks more quickly" (mean = 4.4/5) and "makes it easier to do my job" (mean = 4.4/5). 15 of 18 participants rated at least four items as 4 or 5, suggesting broad belief in enhanced productivity and effectiveness. Furthermore, all expertise groups reported medians of 4.5 for both "Increases Productivity" and "Accomplish Quickly," highlighting consistent perceived benefits in efficiency.

No significant differences were found on any of the Perceived Usefulness scales across expertise levels. To evaluate differences by expertise (novice, proficient, expert; $n = 6$ each), distributional assumptions were tested using Shapiro–Wilk for normality and Levene's for homogeneity of variance. Only "Improves Performance" met criteria (Shapiro–Wilk $p > 0.05$; Levene's $p = 0.1622$), yielding a non-significant one-way ANOVA: $F(2, 15) = 0.17$, $p = 0.8433$, $\eta^2 = 0.02$ (Novice: $M = 3.83$, $SD = 0.75$; Proficient: $M = 4.00$, $SD = 0.63$; Expert: $M = 4.00$, $SD = 0.00$). The other five items violated normality, leading to Kruskal–Wallis tests with p-values from 0.1137 to 0.8679 (all > 0.05), e.g., Enhances Effectiveness: $H(2) = 4.35$, $p = 0.1137$, $\eta^2 = 0.26$.

5.1.5 AttrakDiff

The 28-item AttrakDiff questionnaire assessed participants' subjective impressions of DeTAILS on pragmatic (task-oriented, usability-focused) and hedonic dimensions (emotional, experiential, non-task-oriented), yielding overwhelmingly positive results (Figure 5.5). On pragmatic quality, participants rated the tool as clearly structured (mean 6.3/7), straightforward (mean 5.7/7), practical (mean 2.0/7, where 1 is fully practical), and simple (mean 2.8/7, where 1 is simple). It was also viewed as professional (mean 2.1/7), presentable (mean 6.1/7), and attractive (mean 5.9/7). Hedonic quality ratings were similarly favorable, with the tool described as creative (mean 5.9/7), inventive (mean 2.3/7, where 1 is inventive), captivating (mean 5.8/7), inviting (mean 5.9/7), and integrating (mean 5.7/7). These metrics confirm high usability and enjoyment across the sample of 18 participants.

No significant differences were found on most AttrakDiff items across expertise levels. Normality and homoscedasticity were tested using Shapiro–Wilk and Levene's tests ($\alpha = 0.05$), leading to one-way ANOVA for 16 items meeting assumptions and Kruskal–Wallis for the remaining 12. Most items showed no significant group differences (p-

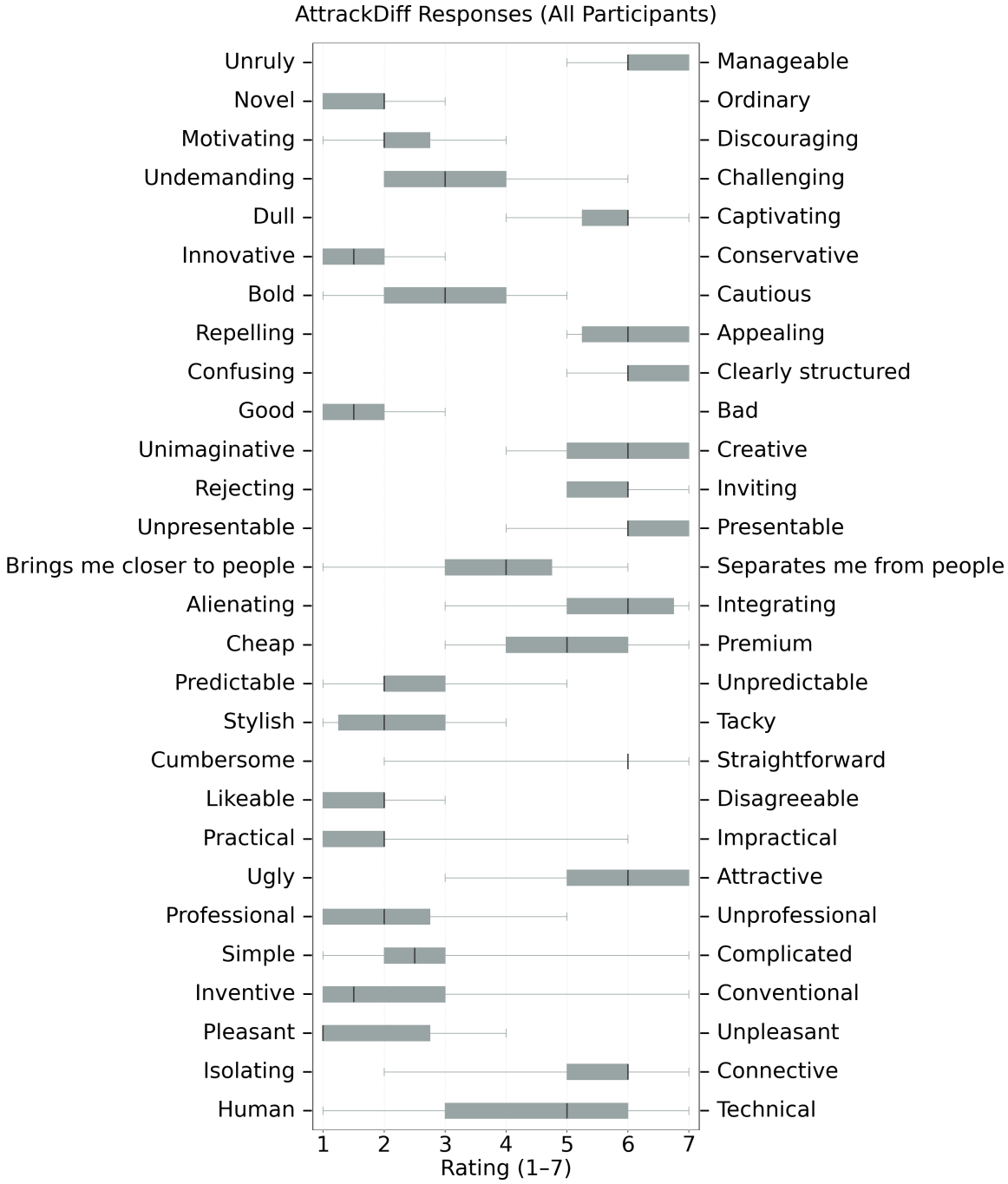


Figure 5.5: AttrakDiff Responses (All Participants) after using DeTAILS. It measures pragmatic (task-oriented, usability-focused) and hedonic dimensions (emotional, experiential, non-task-oriented) through participants' self-reports.

values 0.1196–1.0000), including pragmatic aspects like Human/Technical ($F(2, 15) = 1.06$, $p = 0.3715$, $\eta^2 = 0.12$), Professional/Unprofessional ($F(2, 15) = 0.04$, $p = 0.9642$, $\eta^2 = 0.00$), and Practical/Impractical ($H(2) = 0.05$, $p = 0.9738$, $\eta^2 = 0.00$), as well as hedonic qualities like Inventive/Conventional ($H(2) = 0.50$, $p = 0.7793$, $\eta^2 = 0.03$), Dull/Captivating ($F(2, 15) = 0.25$, $p = 0.7851$, $\eta^2 = 0.03$), and Motivating/Discouraging ($F(2, 15) = 0.07$, $p = 0.9314$, $\eta^2 = 0.01$). A trend emerged on Simple/Complicated (Kruskal–Wallis $\chi^2(2) = 5.88$, $p = 0.0529$, $\eta^2 = 0.35$; novices $Mdn = 2.00$, experts $Mdn = 3.00$), with novices perceiving greater simplicity than experts (post-hoc $t(10) = -2.87$, $p = 0.0165$, Cohen's $d = -1.66$, Bonferroni-corrected $\alpha = 0.0167$).

5.2 Qualitative Results

We structure this section by the toolkit's four main phases (these use LLMs for computation) — *Background Research*, *Coding*, *Reviewing Codes*, and *Generating Themes* — mirroring the reflexive thematic analysis workflow. In each phase, we describe overall usage patterns (noting differences by experience level when present) and compare their typical manual practice to the DeTAILS-supported approach. We then detail the impacts of specific features on their experience, using verbatim quotes to illustrate how each feature changed their analysis. Throughout, we highlight moments where the LLM assistant surfaced new insights or prompted participants to adjust their approach, as well as any limitations or cautions participants noted.

5.2.1 Phase 1: Background Research

All participants found that DeTAILS' Background Research features expedited their analysis. Rather than repeatedly rereading and annotating the same literature to define key concepts, analysts could upload documents they had already reviewed and have related concepts automatically generated—freeing them to focus on validating results. One novice (P2) remarked, *"Doing it manually would have taken an order of magnitude longer... it basically takes all of the legwork out of it."* This pace was also driven by the trust participants gained in DeTAILS's outputs—since its suggestions aligned with their expectations, they could rapidly verify results and move forward. Even experienced analysts found DeTAILS's suggestions largely aligned with their expectations, instilling confidence that important ideas were not missed. As one expert (P17) noted, background research *"pulled pretty well everything I was expecting"* and even surfaced a couple of concepts *"I wasn't necessarily looking for but knew were there... now I am really interested in it,"* referring

to how "fear of falling" is linked to "falls rehabilitation in dementia research"—a topic they had not explicitly mentioned in their research questions, but which emerged as relevant. When asked why this concept interested them, the participant explained: *"So, like fear of falling, I have another paper on fear of falling and how linked it is to fall risk but I hadn't made any research questions about it, but it's also taken the articles [uploaded research papers] so it was encouraging to see that the articles were picking up what I already knew, just didn't mention in the research question... So it's kind of encouraging like that extra verification as well that you're on the right track."* Later in the analysis, codes related to the fear of falling appeared from caregivers' comments about the lived experiences of people with dementia.

Context

Participants likened DeTAILS' Context to a snapshot of their mental map, offloading the cognitive load of recalling where they started. P11 observed, *"I am a visual person. It felt like all the information I needed for my research was present, all there on one screen."* Each participant brought a curated set of source documents to the study, enabling DeTAILS to perform what P13 described as *"basically the first literature scan"* for them. By offloading this work, participants were able to devote their efforts to reviewing and refining the list of concepts. As P11 reflected, *"Using the toolkit, I just have to look at what was already created... that felt a lot easier than doing it myself."* This initial consolidation then served as the foundation for generating Related Concepts and Concept Outline.

Experts nevertheless emphasized that the quality of context is contingent on the inputs they provide. As P18 noted, *"The more you explain what you're looking for, the better the output,"* noting that in their study of *how peer discussions shape perceptions of health technology,* they included additional information- "resources required to support technology adoption" which was reflected as a related concept—"facilitating conditions" (see Figure 5.6)—and, although not particularly prominent in the data, later appeared as a code assigned to two quotes.

Several participants (P3, P13, P14, P17) highlighted the advantage of persistent access to their research context: P17 stated, *"I like that at any point I could go back and edit Context to update the entire research focus,"* commenting on how in the future they could analyze multiple research questions separately using the same literature as context. Participant P3 said, *"I could go back to the very first step [Context] and change it... so I would automatically redefine further steps [Related Concepts],"* reflecting on their desire to avoid related concepts rooted in human emotions and focus solely on "AI's pattern recognition in songs", which they achieved by providing additional information "avoid emotion based

concepts", to the DeTAILS' Context for their research on "the impact of AI in the music industry." This made one of the related concepts change from "Emotional Expressions" to "Understanding Song Concepts".

Related Concepts

Participants across skill levels valued how the related concepts DeTAILS generated surfaced overlooked ideas, related terms, and synonyms. Related concepts acted as a collaborative *"second opinion"* (P14) or *"hidden brainstorming partner"* (P4), rather than accepting every suggestion outright, making the process engaging and efficient. Participant (P2) felt that *"It [finding key concepts] was a total... solo slog... but then here this thing jumps in like an expert partner... and suddenly analysis is actually pretty fun"* adding, *"It came up with stuff I wouldn't think of."*

Novice analysts found that DeTAILS offered a broader, "less biased" perspective as mentioned by P10. Initially focused on trust and bribery in their research on Romanian-Balkan healthcare, P10 was intrigued by the concept of *"low salaries."* This factor was significant in the literature on the region, but they said *"it is not something I want to [re]search in this work"* and did not end up selecting this related concept. They noted, *"At least the tool is a little bit less biased... The fact that it counted low salaries as one of the concepts is very interesting, it made me aware of things I maybe wouldn't be aware of,"* highlighting research aspects that are overlooked. To remove some of the suggestions produced because of the broader perspective of Related Concepts, P10 noted that, by providing feedback on redo—"Medical bribery can be linked to the 'gift-giving' culture found in Eastern Europe and Balkan countries"—the concept's focus shifted to cultural norms, ethics, and the doctor–patient encounter. As they remarked, *"Once I added additional information... literally everything else was pretty spot on to what I'm most interested in. It became even more accurate – it switched the narrative almost."* These early adjustments yielded four final themes—"Challenges in Healthcare Delivery: Quality, Infrastructure, and Patient Access", "Historical Legacies Shaping Contemporary Healthcare", "Systemic Deficiencies: Governance, Trust, and Ethics in Healthcare", and "The Medical Workforce: Pressures, Expertise, and Systemic Impacts"—all of which more focused on cultural and ethical dimensions rather than low salaries.

Experts used it to identify potential oversights. In their research on *"assistive technology use"*, P11 noticed an emphasis on the concept of *"independence,"* which, according to them, *"it's not like offensive or anything, but it's something that I would change in the future based on current critical disability theory directions. But I will leave it for now because I can understand why the language model came to that specific conclusion there based on*

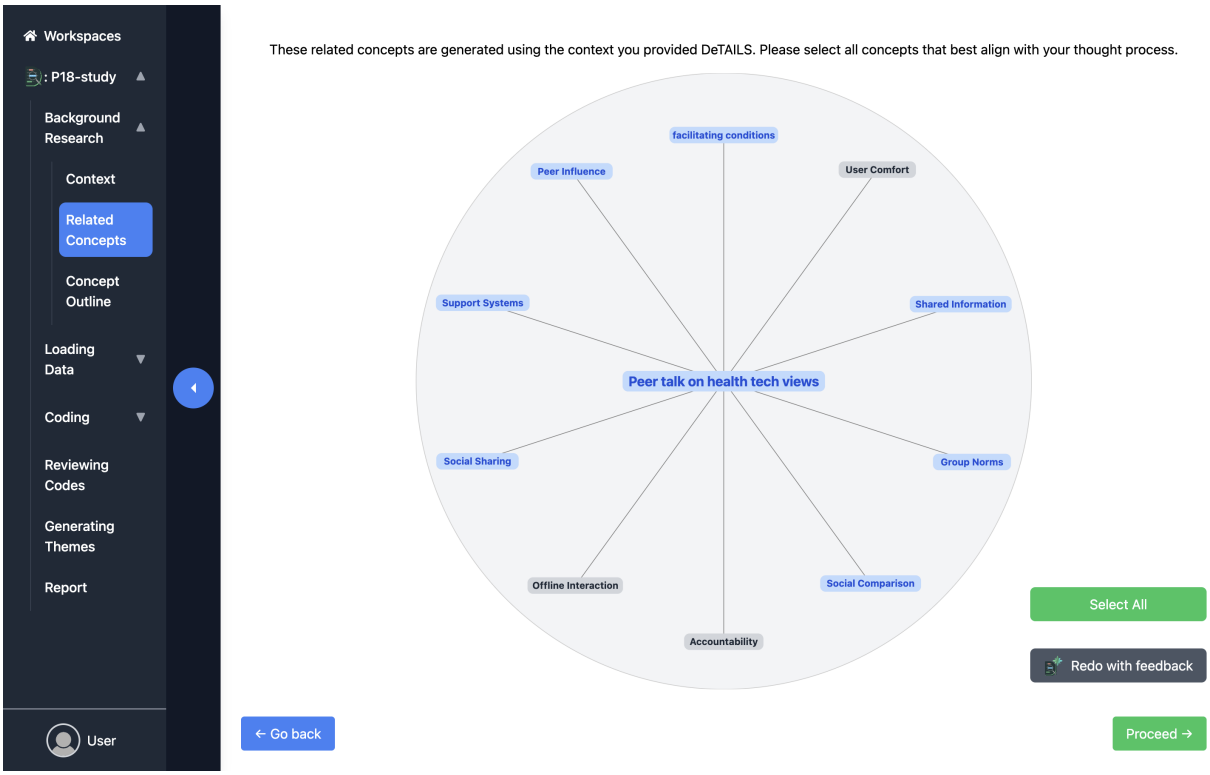


Figure 5.6: Related Concepts generated by DeTAILS for the research on how peer discussions shape perceptions of health technology and edited by P18

my papers. Everything else looks good." They saw this as a source-material bias and were prompted to seek additional literature on independence in their future research. Similarly, expert P18 valued it for revealing an underlying idea, stating, *"related concept found that from my research paper, nice!"* referring to how "Social Comparison" was highlighted as a related concept for their research on "how peer discussions shape perceptions of health technology." Later on in the study this also emerged a code, clustered under the reviewed code "Social Context and Normative Influence in Peer Health Tech Discussions".

Concept Outline

Participants also appreciated that the concept outline provided a brief description for each related concept drawn from their source literature. These read like scholarly explanations, helping them understand how each idea was described in context. As one novice analyst (P9) put it, *"It seems like what I'd read in a background section of a paper explaining these concepts."* While a few participants felt some descriptions were a bit wordy, those who spoke about its quality emphasized their accuracy and professionalism. For example, P11—an expert in disability research—noted that *"the descriptions were correct. It wasn't offensive... that was a plus,"* and P10 (novice analyst) similarly remarked that *"there wasn't any hallucinations really; it was all pretty accurate"*

5.2.2 Phase 3: Coding

Participants universally found DeTAILS' Coding much faster than manual coding. One participant remarked, *"In just seven minutes, the codes are generated... the change is so apparent"* (P16). Another explained, *"I would typically spend seven to eight hours sifting through everything... [DeTAILS' Coding] accomplished this in minutes"* (P12).

After the initial coding, participants adopted reviewer roles—verifying and refining coded entries and, on occasion, re-examining posts deemed irrelevant. As one participant noted, *"I deliberately inspected a couple of posts marked as non-relevant to ensure they truly had nothing of value"* (P16). Participants P4 and P17 likened the workflow to *"having a junior coder perform the first pass,"* followed by their own verification and high-level adjustments. Nonetheless, they cautioned that DeTAILS may miss subtle contextual cues or tonal subtleties—such as sarcasm conveyed through an ironic emoji- which was also illustrated by P4's reference to a code *"Preference for Human-Generated Resumes Over AI Alternatives,"* to the quote (reddit comment), *"...we promise no AI-generated resumes. ¡upside-down face emoji¿."* This highlighted the indispensable role of human expertise

in interpreting such subtleties, even as DeTAILS alleviates the drudgery of preliminary coding.

Many were surprised at how closely the codes generated by DeTAILS matched those they would have created themselves. *"It's usually what my coding looks like,"* said P17 after comparing DeTAILS-generated codes—such as "Physical Health Comorbidities and Medication Related Fall Risk"—to text excerpts like: *"Along with dementia, she has several other health issues. Severe arthritis, the need for some new joints, loss in sight and hearing, some new foot pain (so she can barely walk at the moment), among others."*. This comparison closely aligned with P17's expectations, as they agreed with the explanation provided by DeTAILS that *"these conditions are significant contributors to an elevated fall risk...,"* and observed that *"tool[DeTAILS] accurately inferred fall risk... from just the health issues mentioned [in the quote]."*

A concern emerged that while code labels and themes were often relevant, the specific quotations chosen to represent the code sometimes missed the mark. For some participants, the pattern of highlighted quote excerpts differed from their own style. P5 made changes to the quotes initially picked by DeTAILS and said, *"they[quotes] are correct but very short... I also add a few lines before and after [in the quote] to capture more detail of what is being said."* P15 noted they *"edited one [quote] a little bit,"* removing a few of the last words from a highlighted, lengthy quote as the *"...rest [of the quotes] were on target."*

Several participants admitted they were initially cautious about trusting the AI's coding until they examined the evidence. For example, after editing the codes P14 said *"having the codes from the beginning is new for me and I feel like I don't initially trust it... but the more I read, it's like ok – these are general themes around these quotes I was looking for."* In other words, once they read the explanations and the actual quote excerpts under each code, they recognized that the categories were valid. This pattern of initial skepticism turning into agreement was common, especially for those unfamiliar with letting an AI take the first pass.

5.2.3 Phase 4: Reviewing Codes

In the Reviewing Codes phase, participants saw that DeTAILS automatically grouped similar codes into clusters (higher-level codes). Overall, they found these groupings sensible and useful for streamlining the codebook. For example, P10 (novice) said *"I wouldn't change any of these groupings... they're grouped pretty well,"* and only changed one cluster name. Experts like P11 and P18 merged more codes into the same cluster to consolidate them even further, as P11 said, *"I could focus on thinking about connections...instead of*

worrying did I code every single thing." Many were pleased that they didn't have to manually discover redundant codes — the system had put similar codes in the same bucket and made the clustering process easier for them. One novice (P1) noted that the toolkit *"showed me if anything was basically the same, which definitely saves time."* P12 appreciated the boost, saying, *"after a quick pass organizing codes it wasn't necessary to invest the usual many hours into this step,"* allowing them to shift focus from mechanical cleanup to higher-order interpretation.

Reviewing Codes also surfaced connections participants hadn't initially considered. For P6, whose research focused on *'parents' sentiments and concerns regarding vaccines,"* DeTAILS produced four reviewed codes: Perceived Vaccine Risks and Experienced Impacts, Navigating the Vaccine Information Ecosystem and Decision-Making, Parental Vaccination Stances, Choices, and Their Social Ramifications, and Engagement with and Responses to Vaccine Misinformation. While this structure was not what P6 would have initially created, they recognized its value. *"I think I would have probably done a lot of this very different, but I like the way it [DeTAILS] did it too,"* P6 said. They attributed this to their own focused research question — *"What are the current sentiments and concerns expressed by parents of young children on Reddit regarding vaccines?"* — noting, *"I think I just would have come in with a different like theoretical framework... so it wouldn't have ever done it that way, but I actually do like it this way."* This process provided P6 with an alternative thematic framework to report in their findings, which they recognized as a valid—though different—way to interpret their data and chose to adopt.

An interesting strategy for clustering codes came from P13, who didn't want to lose unique but important points: they created a *"Miscellaneous"* bucket to hold one-off codes that had only a single example but were still noteworthy. *"Even if it's just one person saying it, but it's a really important comment, I will keep it"* they explained, reasoning that those insights should be retained separately rather than discarded. There was one instance where the code clusters did not resonate with the participant's thinking process. Participant P7 had 25 codes, which were initially grouped into 5 reviewed codes: *"Perspectives on AI: Critiques, Capabilities, and Future"; "AI for Enhanced Information Access and Comprehension"; "Advancing AI Capabilities and Developmental Methodologies"; "AI for Work Augmentation and Scientific Acceleration";* and *"AI as a Catalyst for Learning and Critical Engagement."* After analyzing each reviewed code and its associated codes, P7 decided to consolidate them into three reviewed codes: *"Perspectives on AI: Critiques, Capabilities, and Future"; "AI for Work Augmentation and Scientific Acceleration";* and *"Miscellaneous."* They explained, *"I want all the reviewed codes, but I don't want my focus to move away from the most important ones... I need the others, but I create an 'Others' group and use them only to fill gaps."* This outcome demonstrated that even during the Re-

viewing Codes phase, human intervention was necessary to fully capture the participant's research goals.

5.2.4 Phase 5: Generating Themes

Participants found the DeTAILS-generated themes logical and aligned with their own interpretation of the data. For example, P15 (expert), analyzing discussions on *r/blackladies* about Black womens' experiences with multimorbidity, found the themes to be *"very, very, very relevant."* They were particularly impressed with the *"sophisticated language and conceptual grouping in themes"* like *"Systemic Constraints and Enduring Legacies: Navigating Adversity in Health and Society,"* and *"The Interwoven Tapestry of Lived Experience: Identity, Support, and Agency in Multimorbidity"* given the complexity of their topic. P15 noted, *"for it to be able to generate themes and I can tell like, yeah, this is the language we would utilize and these are the main concepts that I would be looking for or would generally see, I think is pretty accurate"* (see Figure 5.7).

In general, participants reported that each theme corresponded to a major aspect of their research question. This accuracy was frequently highlighted, with one expert (P6) observing, *"I would organize a paper this way too,"* and a proficient analyst (P3) giving a succinct endorsement: *"This is exactly what I would name them."* P9 noted that DeTAILS' Generating Themes allowed them to skip the *"messy middle"* of theme development, estimating the tool had *"done 90% of the work."* One participant (P13, proficient) noted DeTAILS's theme titles were surprisingly well-phrased, *"almost like a manuscript title,"* which meant they could spend less time wordsmithing.

On the other hand, P14 liked the themes though they commented the wording was a bit "magical" referring to the theme *"Championing Clarity, Accessibility, and Inclusivity in Educational Content and Platform Design."* DeTAILS also surfaced participant-familiar, non-obvious concept connections in the generated themes. For instance, P5 was intrigued when DeTAILS clustered codes about *"AI Chatbot Adoption Drivers and Processes"* and *"User Trust and Reliability Assessment of AI Chatbots"* into a cohesive theme titled *"Dynamics of User Adoption and Trust in AI Chatbots."* This way of generating themes articulated a direct link between the two concepts, prompting P5 to reflect: *"Adoption is directly related to trust and reliability... I was not expecting it would get that, and I like how it did that".* P13 reflected on their journey from skepticism to trust: after reviewing the final themes, they admitted, *"I actually really like the themes. I like them more than the codes, so maybe I just have to trust the process that we'll get there."*

At the same time, participants recognized the importance of their own role in interpreting and finalizing the analysis. As P17 put it, *"even with an AI partner, ultimately the*

researcher is the instrument. The human decides what story to tell with the themes. You can add things along the way based on your understanding of the data, and themes are just an overview of the pattern in the data. At the end of the day, I will go through the whole process multiple times just to know how to report the actual findings from those patterns." They saw DeTAILS as an assistant that accelerates the analytic steps while keeping the researcher in control of meaning-making.

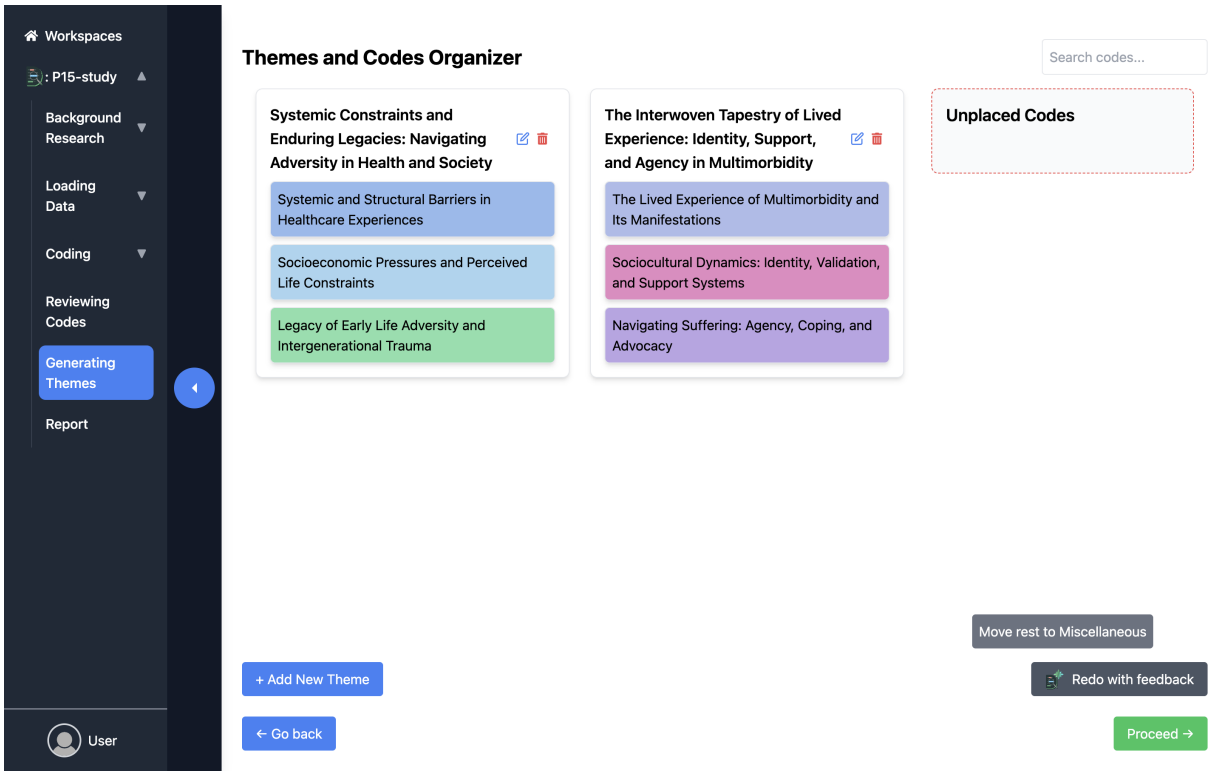


Figure 5.7: Themes generated by DeTAILS for the research on Black womens' experiences with multimorbidity by P15

Chapter 6

Discussion

DeTAILS was developed in response to key challenges of traditional thematic analysis — notably, its labour intensity and difficulty scaling to large datasets — and researchers' hesitancy to trust algorithms with qualitative coding. By integrating an LLM into a stepwise, interactive workflow, we aimed to dramatically accelerate coding while preserving researcher agency and trust. Our study confirmed that DeTAILS can indeed accelerate analysis and generate coherent, valuable codes and themes that closely align with human interpretations.

We now reflect on how DeTAILS addresses concerns raised in prior work like reliability, interpretability, and trust in AI, and outline the broader implications for LLM-assisted qualitative research. These insights illustrate a path toward significantly faster yet still reflexive thematic analysis.

6.1 LLM-assisted Coding Accelerates Analysis

Participants reported that DeTAILS, through LLM-assisted coding, accelerated thematic analysis. P16 noted, *"In just seven minutes, the codes are generated... the change is so apparent,"* and P12 recalled that manual coding typically took seven to eight hours for a similar analysis, whereas while using DeTAILS it was completed in minutes. Measured timings (Figure 5.2) confirmed this speed: participants completed coding their datasets in an average of 33 minutes. Participants also perceived DeTAILS as efficient, with 16/18 agreeing that "Using the system enables me to accomplish tasks quickly" (Figure 5.4), and easy to use, with low NASA-TLX scores (Figure 5.3). Throughout the coding process

accuracy remained high, with F1 raising from 0.90 in initial coding to 1.0 when generating themes (Figure 5.1), indicating that the LLM-generated codes and themes matched participants' expectations.

Prior work on LLM-assisted coding often forces researchers into a trade-off between speed and accuracy. Some systems prioritize group consensus over individual speed. For instance, CollabCoder [47] enables multiple analysts to code inductively with optional LLM suggestions, but analysis speed hinges on analysts' discussion time for agreement [47]. Other approaches use simple prompting styles. Though initially fast, they yield superficial outputs needing manual refinement, negating time savings [57, 101, 102]. Even commercial tools like NVivo provide partial automation when using LLMs. They suggest granular codes but leave theme development to users, failing to speed up the complete workflow [82]. The key challenge, therefore, has been to create an AI assistant integrated system that provides end-to-end acceleration for a single analyst.

DeTAILS uniquely accelerates the entire coding workflow by integrating inductive discovery with deductive application. The process begins inductively, as the LLM generates emergent codes from an initial data subset, which the researcher then refines to create a data-grounded codebook. This initial codebook is then applied deductively across the full dataset, yet allowing for new codes to emerge. This hybrid approach resolves a central and time consuming bottleneck of qualitative analysis: coding data. The inductive phase solves the "cold start" problem of creating a relevant codebook, and the high-quality output of that phase makes the subsequent deductive application both fast and accurate.

6.2 LLMs Help Transform Topics into Themes

Participants praised DeTAILS for converting codes into coherent, relevant themes. P15, studying health discussions, was *"very, very, very impressed"* with its *"sophisticated language and grouping"*, which captured all main concepts. P3 said generated theme titles were *"exactly what I would name them."* Themes aligned with participants' research goals, translating codes into broader patterns. Titles were well-phrased, *"almost like a manuscript title"* (P13), easing the task of wording them correctly. DeTAILS managed the labor-intensive *"messy middle"* of condensing codes, doing *"90% of the work"* (P9). All 18 participants accepted AI-generated themes unmodified, as measured by a macro F1 score of 1.00 (Figure 5.1), indicating that the LLM-generated themes matched participants' expectations exactly. 15/18 participants agreed that "Using the system enhances my effectiveness in my job"(Figure 5.4).

Existing thematic analysis tools struggle to incorporate researchers' domain knowledge,

and tend to produce purely descriptive topics that lack depth and require extensive manual work to form meaningful themes. Tools like the CTA Toolkit [48, 71], which use topic models like LDA or BERTopic, cluster data via word co-occurrence or semantic similarity. The topics they produce require further interpretation to uncover their intended meaning and avoid surface-level insights [25, 49, 85]. Systems like LLooM [75] use LLMs to generate coherent, diverse, and high-level concepts, but often fail to take into account the researcher's insights, resulting in more comprehensive yet generic topics that require tedious interpretation post computation.

DeTAILS generated coherent themes based on a researcher's questions by identifying links between the codes they created. These links then revealed hidden patterns in the data. For instance, P5 noted that it made surprisingly good connections, like *"AI chatbot adoption"* and *"user trust"* under a user-AI trust dynamics theme. Participants remarked that the patterns then served as reflective prompts to explore code connections for deeper insights. DeTAILS also connects themes back to supporting codes, quotes, and the LLMs' earlier explanations of context for each code assignment. To move beyond generic topics, future systems should therefore ground theme generation in researchers' domain knowledge and reveal code connections in a transparent, traceable manner that prompts deeper analytical reflection.

6.3 Integrating Qualitative Research and Chain-of-Thought

DeTAILS's design externalizes the analyst's thought process into a structured, visible chain of tasks. The system decomposes the complex, often internalized, process of thematic analysis into a six-phase workflow (Figure 3.1). This structure requires the analyst to engage deliberately at each stage, from defining the initial research context to refining the final themes. This step-by-step engagement was found to be useful in the user study; for instance, P6 said, *"I like how it looks... the sidebar with all the steps,"* P11 felt that *"all the information I needed for my research was present, all there,"* while P3 appreciated the ability to *"go to any step... see what I was thinking"* at any time. This chained process, where the output of one step becomes the explicit input for the next, builds user confidence, as exemplified by P13's journey to *"trust the process,"* with 15/18 agreeing with "I find the system useful in my job"(Figure 5.4).

Existing QDA-LLM integrations struggle to support the analyst's iterative sense-making. Frameworks such as LLM-in-the-loop [31] use code examples for coding the data without

sharing their logic with the LLM, this approach often fails to incorporate what the researcher is actually looking for in the data [31, 34]. Systems such as MindCoder [46] automate an end-to-end qualitative pipeline yet make human intervention optional — analysts can provide their insights when they want but are not required to review each stage — potentially missing critical corrections early on. Tools like LLooM [75] use multi-step prompting but emphasize automated outputs, often bypassing the analyst's ongoing reflection for efficiency. Moreover, these approaches lock in only the researcher's initial objectives, following a predetermined sequence of tasks and overlooking evolving insights that arise throughout the analysis.

DeTAILS transforms qualitative analysis by turning LLM into a collaborative partner that embeds researchers' judgment as a core part of its transparent, step-by-step workflow. This "Chain-of-Tasks" offers clear goals and suggestions, reducing busywork so analysts can focus on insights, while allowing them to input domain knowledge that dynamically refines future steps and maintains a smooth, iterative flow while guiding the AI. By fostering deeper engagement through traceable AI logic (like code-quote-explanations) and human interpretation, DeTAILS ensures that researchers' reflexive thinking drives meaningful discoveries, setting a new standard for AI tools that prioritize human insight over automation.

6.4 Agency and Trust

Participants reported a strong sense of agency and trust in DeTAILS as a supportive tool. They described themselves as *"in the driver's seat,"* (P4) and the AI as a *"junior coder"* (P17) who will do the analysis as a first pass and then they can verify the work. This sense of agency was reflected in high AttrakDiff ratings for manageability (mean = 5.97) (Figure 5.5), as P17 asserted: *"ultimately the researcher is the instrument... AI is just like an extra set of eyes."* Trust emerged iteratively through interaction, with participants shifting from initial wariness to confidence as they used the tool. For instance, P13 admitted to not trusting the initial codes but came to *"trust the process"* upon review. The ultimate testament to trust in DeTAILS was when participants asked, *"I would love to explore it more for my research"* (P8), *"so when's this available? When can I use this?"* (P17).

Building researcher trust in AI is an ongoing challenge. "Black box" designs, like single-step AI coding, fail to build trust by hiding reasoning [74, 138]. When many analysts agree on an AI suggestion, people trust it more but this stops individual analysts from exploring more on their own [47]. While the literature has long recommended general trust-enhancing practices like human-in-the-loop control and step-by-step verification as potential solutions to these challenges [81], a key gap remains: existing systems often fall short in seamlessly

integrating these principles of trust and agency into a reflexive workflow for thematic analysis.

Trustworthy human–AI collaboration in qualitative research requires AI tools that are transparent, editable, and keep the researcher in control. Participants often reported increased trust in DeTAILS associated with the ability to inspect and modify its computational steps and with AI outputs they judged to be accurate. DeTAILS offers an interface for cross-checking AI-proposed codes and themes against raw data, including code explanations to enhance transparency. Participants perceived AI outputs as accurate, boosting their trust in the system. This design surpasses opaque, one-shot outputs, enabling interactive partnerships where analysts guide and correct AI, retaining decision authority. Such processes build earned trust and preserve interpretive agency.

Chapter 7

Limitations and Future Work

There are several limitations that suggest directions for future work.

First, our evaluation was short-term and controlled; we lack evidence of how DeTAILS would fare in long-term, real-world use. A longitudinal deployment study is needed to observe whether researchers continue to adopt the toolkit over extended projects, how it impacts their workflow and rigour over time, and what new use cases or challenges emerge in naturalistic settings. Understanding sustained adoption and effects on qualitative practice will be crucial before broader adoption.

Second, we did not directly compare DeTAILS with established qualitative analysis software like NVivo or Atlas.ti. Such tools are widely used for coding and theming, and a head-to-head evaluation (especially on large or complex datasets) would clarify DeTAILS's relative strengths and weaknesses. Future work should involve comparative studies to benchmark the toolkit's efficiency, theme quality, and user experience against these existing platforms. This comparison could help identify scenarios where an LLM-supported approach provides the most value, and also highlight any features from traditional software that DeTAILS might integrate.

Third, our current implementation explores only a narrow range of prompting techniques and model types. We used a general-purpose LLM with a chain-of-thought prompting workflow, without running parallel studies with other prompting strategies (e.g., ReAct prompting [134]) or specialized models. It remains an open question how different LLM configurations might influence the outcomes. Future research should experiment with diverse prompting approaches and model variants — for instance, comparing reasoning-intensive prompts versus direct answers, or using domain-specific fine-tuned models in place of a general model. These explorations could improve the quality of AI-generated codes and

themes, and inform guidelines on pairing the right prompting technique or model with particular qualitative analysis tasks.

Chapter 8

Conclusion

We introduced DeTAILS as a Human-LLM collaboration toolkit that integrates LLM assistance into each phase of reflexive thematic analysis that the human guides. Key features—like the persistent "memory snapshot" to inject context into every prompt and the "redo-with-feedback" loop or interactive manual intervention for iterative refinement—ensure continuity from coding through theme generation while keeping the researcher's influence at the forefront. The interface makes every LLM-generated code and theme inspectable and editable, promoting transparency and enabling analysts to verify and adjust outputs against the raw data.

Our study with 18 analysts confirmed DeTAILS's effectiveness and ease of use. Participants used the toolkit to perform thematic analysis on a heterogeneous text dataset of their choice, reporting that the iterative LLM support accelerated their process of generating codes and themes accurately. All 18 participants agreed with the DeTAILS-generated themes, requiring no modifications, indicating that the workflow preserved interpretive depth and earned their trust. Participants described feeling "in the driver's seat" of the analysis and viewed the LLM as a diligent assistant rather than an autonomous agent, highlighting how the tool fostered agency and trust.

In conclusion, this work demonstrates that LLM assistance can be scaled up without undermining qualitative depth. DeTAILS's memory-grounded prompt chaining and interactive feedback loops illustrate how LLMs can amplify analysts' capabilities while preserving their control and reflexivity. Ultimately, our approach shows that LLMs can accelerate and scale thematic analysis, yet still uphold the reflexive depth of the researcher's interpretive process, as evidenced by high alignment scores, low workload, and positive trust feedback.

References

[1] Jos Akkermans and Sophie Kubasch. "The Research Audit Trail: Methodological Guidance for Application in Practice". In: *Electronic Journal of Business Research Methods* 19.1 (2021), pp. 1–12. URL: https://academic-publishing.org/index.php/ejbrm/article/view/2033.

[2] Ian Arawjo et al. "ChainForge: An Open-Source Visual Programming Environment for Prompt Engineering". In: *UIST '23 Adjunct*. New York, NY, USA: ACM, 2023, pp. 1–3. DOI: 10.1145/3586182.3616660. URL: https://doi.org/10.1145/3586182.3616660.

[3] Jennifer Attride-Stirling. "Thematic networks: an analytic tool for qualitative research". In: *Qualitative Research* 1.3 (2001), pp. 385–405. DOI: 10.1177/146879410100100307. eprint: https://doi.org/10.1177/146879410100100307. URL: https://doi.org/10.1177/146879410100100307.

[4] B. Baker, J. Hoffmann, et al. "Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation". In: *arXiv preprint arXiv:2503.11926* (2025). URL: https://arxiv.org/abs/2503.11926.

[5] Muneera Bano, Didar Zowghi, and Jon Whittle. "Exploring Qualitative Research Using LLMs". In: *arXiv preprint arXiv:2306.13298* (2023). URL: https://arxiv.org/abs/2306.13298.

[6] Tehmina N. Basit. "Manual or electronic? The role of coding in qualitative data analysis". In: *Educational Research* 45.2 (2003), pp. 143–154. DOI: 10.1080/0013188032000133548.

[7] Jason Baumgartner et al. *The Pushshift Reddit Dataset*. 2020. arXiv: 2001.08435 [cs.SI]. URL: https://arxiv.org/abs/2001.08435.

[8] Melissa Beresford et al. "Coding Qualitative Data at Scale: Guidance for Large Coder Teams Based on 18 Studies". In: *International Journal of Qualitative Methods* 21 (2022). DOI: 10.1177/16094069221075860.

[9] Maciej Besta et al. “Graph of Thoughts: Robot Scientist Explorations via Multimodal Chain of Thought”. In: *arXiv preprint arXiv:2403.04690* (2024). URL: https://arxiv.org/abs/2403.04690.

[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. “Latent Dirichlet Allocation”. In: *Journal of Machine Learning Research* 3.4–5 (Jan. 2003), pp. 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993. URL: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

[11] Rishi Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2022. arXiv: 2108.07258 [cs.LG]. URL: https://arxiv.org/abs/2108.07258.

[12] Richard Boyatzis. “Transforming Qualitative Information: Thematic Analysis and Code Development”. In: (Jan. 1998).

[13] Jordan Boyd-Graber, Yuening Hu, and David Mimno. “Applications of Topic Models”. In: *Found. Trends Inf. Retr.* 11.2–3 (July 2017), pp. 143–296. ISSN: 1554-0669. DOI: 10.1561/1500000030. URL: https://doi.org/10.1561/1500000030.

[14] Virginia Braun and Victoria Clarke. “Using thematic analysis in psychology”. In: *Qualitative Research in Psychology* 3 (Jan. 2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa.

[15] Virginia Braun and Victoria Clarke. “Reflecting on reflexive thematic analysis”. In: *Qualitative Research in Sport, Exercise and Health* 11.4 (2019), pp. 589–597. DOI: 10.1080/2159676X.2019.1628806. eprint: https://doi.org/10.1080/2159676X.2019.1628806. URL: https://doi.org/10.1080/2159676X.2019.1628806.

[16] Virginia Braun and Victoria Clarke. “One size fits all? What counts as quality practice in (reflexive) thematic analysis?” In: *Qualitative Research in Psychology* 18.3 (2021), pp. 328–352. ISSN: 1478-0895, 1478-0887. DOI: 10.1080/14780887.2020.1769238.

[17] Virginia Braun and Victoria Clarke. “One size fits all? What counts as quality practice in (reflexive) thematic analysis?” In: *Qualitative Research in Psychology* 18.3 (2021), pp. 328–352. DOI: 10.1080/14780887.2020.1769238. eprint: https://doi.org/10.1080/14780887.2020.1769238. URL: https://doi.org/10.1080/14780887.2020.1769238.

[18] Jenna Burrell. "How the machine 'thinks': Understanding opacity in machine learning algorithms". In: *Big Data & Society* 3.1 (2016), p. 2053951715622512. DOI: 10.1177/2053951715622512. eprint: https://doi.org/10.1177/2053951715622512. URL: https://doi.org/10.1177/2053951715622512.

[19] Ashley Castleberry and Amanda Nolen. "Thematic analysis of qualitative research data: Is it as easy as it sounds?" In: *Currents in Pharmacy Teaching and Learning* 10.6 (2018), pp. 807–815. ISSN: 1877-1297. DOI: https://doi.org/10.1016/j.cptl.2018.03.019. URL: https://www.sciencedirect.com/science/article/pii/S1877129717300606.

[20] Abinaya Chandrasekar et al. "Making the most of big qualitative datasets: a living systematic review of analysis methods". In: *Frontiers in Big Data* Volume 7 - 2024 (2024). ISSN: 2624-909X. DOI: 10.3389/fdata.2024.1455399. URL: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2024.1455399.

[21] Eshwar Chandrasekharan et al. "The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–25. DOI: 10.1145/3274301. URL: https://doi.org/10.1145/3274301.

[22] Harrison Chase. *LangChain*. https://github.com/langchain-ai/langchain. Version 0.3.27.

[23] B. Chen, R. Patel, A. Lee, et al. "Large Language Models in Qualitative Research: Can We Do the Data Justice?" In: *arXiv preprint arXiv:2410.07362* (2024). URL: https://arxiv.org/abs/2410.07362.

[24] Y. Chen, X. Li, J. Kim, et al. "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying About Prompt Formatting". In: *arXiv preprint arXiv:2310.11324* (2023). URL: https://arxiv.org/abs/2310.11324.

[25] Yingying Chen et al. "What We Can Do and Cannot Do with Topic Modeling: A Systematic Review". In: *Communication Methods and Measures* 17 (Jan. 2023), pp. 1–20. DOI: 10.1080/19312458.2023.2167965.

[26] R. Chew, P. Singh, et al. "Artificial Intelligence and Qualitative Research: The Promise and Perils of Large Language Model (LLM) 'Assistance'". In: *Poetics* 103 (2024), p. 101892.

[27] Robert Chew et al. "LLM-assisted Content Analysis: Using Large Language Models to Support Deductive Coding". In: *arXiv preprint arXiv:2306.14924* (2023). URL: https://arxiv.org/abs/2306.14924.

[28] Tom Christie. *Uvicorn*. https://www.uvicorn.org/. Accessed: 2025-07-07.

[29] Joseph Paul Cohen and Henry Z Lo. "Academic Torrents: A Community-Maintained Distributed Repository". In: *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*. ACM. 2014, pp. 1–2. DOI: 10.1145/2616498.2616528.

[30] Yann Collet and Murray Kucherawy. *Zstandard Compression and the application/zstd Media Type*. RFC 8478. Internet Engineering Task Force, 2018. DOI: 10.17487/RFC8478.

[31] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. "LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis". In: (2023). arXiv: 2310.15100 [cs.CL]. URL: https://arxiv.org/abs/2310.15100.

[32] Fred D. Davis. "Perceived usefulness, perceived ease of use, and user acceptance of information technology". In: *MIS Quarterly* 13.3 (1989), pp. 319–340. DOI: 10.2307/249008.

[33] S. De Paoli. "Further Explorations on the Use of Large Language Models for Thematic Analysis. Open-Ended Prompts, Better Terminologies and Thematic Maps". In: *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 25.2 (2024).

[34] S. De Paoli. "Performing an Inductive Thematic Analysis of Semi-Structured Interviews With a Large Language Model: An Exploration and Provocation on the Limits of AI in Qualitative Research". In: *Social Science Computer Review* 42.3 (2024), pp. 678–695.

[35] Stefano De Paoli and Walter S. Mathis. "Reflections on Inductive Thematic Saturation as a Potential Metric for Measuring the Validity of an Inductive Thematic Analysis with LLMs". In: *Quality & Quantity* 59.1 (2025), pp. 683–709.

[36] M. S. Deiner, Y. Zhang, J. Williams, et al. "Large Language Models Can Enable Inductive Thematic Analysis of a Social Media Corpus in a Single Prompt: Human Validation Study". In: *JMIR Infodemiology* 4.1 (2024), e59641.

[37] J. Drápal, P. Novák, and P. Švec. *Using Large Language Models to Support Thematic Analysis in Empirical Legal Studies*. SSRN Electronic Journal. 2023.

[38] Zackary Okun Dunivin. "Scalable Qualitative Coding with LLMs: Chain-of-Thought Reasoning Matches Human Performance in Some Hermeneutic Tasks". In: *arXiv preprint arXiv:2401.15170* (2024). URL: https://arxiv.org/abs/2401.15170.

[39] Electron. *Electron*. https://www.electronjs.org/docs/latest. Accessed: 2025-07-07.

[40] T. Eloundou et al. "Large Language Models: A Survey". In: *arXiv preprint arXiv: 2402.06196* (2024). URL: https://arxiv.org/abs/2402.06196.

[41] Facebook. *React*. https://react.dev/. Accessed: 2025-07-07.

[42] Jessica L. Feuston and Jed R. Brubaker. "Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis". In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021). DOI: 10.1145/3479856. URL: https://doi.org/10.1145/3479856.

[43] Linda Finlay. "Negotiating the swamp: the opportunity and challenge of reflexivity in research practice". In: *Qualitative Research* 2.2 (2002), pp. 209–230. DOI: 10.1177/146879410200200205. eprint: https://doi.org/10.1177/146879410200200205. URL: https://doi.org/10.1177/146879410200200205.

[44] Andrew Gallant. *ripgrep: a line-oriented search tool*. https://github.com/BurntSushi/ripgrep.

[45] Yasir Gamieldien, Jennifer M. Case, and Andrew Katz. "Advancing Qualitative Analysis: An Exploration of the Potential of Generative AI and NLP in Thematic Coding". In: *SSRN Electronic Journal* (2023). URL: https://doi.org/10.2139/ssrn.4487768.

[46] Jie Gao, Zhiyao Shu, and Shun Yi Yeo. *MindCoder: Automated and Controllable Reasoning Chain in Qualitative Analysis*. 2025. arXiv: 2501.00775 [cs.HC]. URL: https://arxiv.org/abs/2501.00775.

[47] Jie Gao et al. *CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models*. 2024. arXiv: 2304.07366 [cs.HC]. URL: https://arxiv.org/abs/2304.07366.

[48] Robert Gauthier and James Wallace. "The Computational Thematic Analysis Toolkit". In: *Proceedings of the ACM on Human-Computer Interaction* 6 (Jan. 2022), pp. 1–15. DOI: 10.1145/3492844.

[49] Omar El-Gayar et al. *A Comparative Analysis of the Interpretability of LDA and LLM for Topic Modeling: The Case of Healthcare Apps*. Tech. rep. 423. Research & Publications, 2024. URL: https://scholar.dsu.edu/bispapers/423.

[50] Simret Araya Gebreegziabher et al. "PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581352. URL: https://doi.org/10.1145/3544548.3581352.

[51] M. F. P. Gillies et al. "Theme and Topic: How Qualitative Research and Topic Modeling Can Be Brought Together". In: *ArXiv* abs/2210.00707 (2022). URL: https://api.semanticscholar.org/CorpusID:252683229.

[52] Google. *Vertex AI Documentation*. https://cloud.google.com/vertex-ai/docs.

[53] Maarten Grootendorst. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. 2022. arXiv: 2203.05794 [cs.CL]. URL: https://arxiv.org/abs/2203.05794.

[54] Greg Guest, Kathleen M. MacQueen, and Emily E. Namey. *Applied Thematic Analysis*. Thousand Oaks, California: SAGE Publications, Inc., 2012. DOI: 10.4135/9781483384436. URL: https://methods.sagepub.com/book/mono/applied-thematic-analysis/toc.

[55] Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research". In: *Human Mental Workload*. Ed. by Peter A. Hancock and Najmedin Meshkati. Amsterdam: North-Holland, 1988, pp. 139–183. DOI: 10.1016/S0166-4115(08)62386-9.

[56] Marc Hassenzahl, Michael Burmester, and Franz Koller. "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität". In: *Mensch & Computer 2003. Interaktion in Bewegung*. Ed. by Jürgen Ziegler and Gerd Szwillus. Stuttgart: B.G. Teubner, 2003, pp. 187–196.

[57] Adam S. Hayes. ""Conversing" With Qualitative Data: Enhancing Qualitative Research Through Large Language Models (LLMs)". In: *International Journal of Qualitative Methods* 24 (2025), p. 16094069251322346. DOI: 10.1177/16094069251322346. eprint: https://doi.org/10.1177/16094069251322346. URL: https://doi.org/10.1177/16094069251322346.

[58] Zeyu He et al. "If in a Crowdsourced Data Annotation Pipeline, a GPT-4". In: *CHI '24*. New York, NY, USA: ACM, 2024, pp. 1–25. DOI: 10.1145/3613904.3642834. URL: https://doi.org/10.1145/3613904.3642834.

[59] Zeyu He et al. “If in a Crowdsourced Data Annotation Pipeline, a GPT-4”. In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI ’24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642834. URL: https://doi.org/10.1145/3613904.3642834.

[60] Richard D Hipp. *SQLite*. https://sqlite.org/. Accessed: 2025-07-07.

[61] Chenyu Hou et al. “Prompt-based and Fine-tuned GPT Models for Context- Dependent and -Independent Deductive Coding in Social Annotation”. In: *LAK ’24*. 2024, pp. 518–528.

[62] Michelle K. Jamieson, Gisela H. Govaart, and Madeleine Pownall. “Making the most of big qualitative datasets: a living systematic review of analysis methods”. In: *Frontiers in Big Data* 7 (2024), p. 1455399. DOI: 10.3389/fdata.2024.1455399.

[63] Shagun Jhaver et al. “”Did You Suspect the Post Would be Removed?”: Understanding User Reactions to Content Removals on Reddit”. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–33. DOI: 10.1145/3359294. URL: https://doi.org/10.1145/3359294.

[64] Jialun Aaron Jiang et al. “Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449168. URL: https://doi.org/10.1145/3449168.

[65] Jialun Aaron Jiang et al. “Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1 (Apr. 2021). DOI: 10.1145/3449168. URL: https://doi.org/10.1145/3449168.

[66] Jim Wallace. *PASS*. https://git.uwaterloo.ca/jrwallace/PASS/. Accessed: 2025-07-07.

[67] Jessica L. Johnson, Donna Adkins, and Sheila Chauvin. “A Review of the Quality Indicators of Rigor in Qualitative Research”. In: *American Journal of Pharmaceutical Education* 84.1 (2020), p. 7120. ISSN: 0002-9459. DOI: https://doi.org/10.5688/ajpe7120. URL: https://www.sciencedirect.com/science/article/pii/S0002945923016066.

[68] Adam Johs et al. *Qualitative Investigation in Explainable Artificial Intelligence: A Bit More Insight from Social Science*. 2020. arXiv: 2011.07130 [cs.HC].

[69] Shivani Kapania et al. "'I'm Categorizing LLM as a Productivity Tool': Examining Ethics of LLM Use in HCI Research Practices". In: *Proc. ACM Hum.-Comput. Interact.* 9.2 (May 2025). DOI: 10.1145/3711000. URL: https://doi.org/10.1145/3711000.

[70] Andrew Katz, Gabriella Coloyan Fleming, and Joyce Main. "Thematic Analysis with Open-Source Generative AI and Machine Learning: A New Method for Inductive Qualitative Codebook Development". In: *arXiv preprint arXiv:2410.03721* (2024). URL: https://arxiv.org/abs/2410.03721.

[71] Amandeep Kaur and James Wallace. *Moving Beyond LDA: A Comparison of Unsupervised Topic Modelling Techniques for Qualitative Data Analysis of Online Communities*. Dec. 2024. DOI: 10.48550/arXiv.2412.14486.

[72] Elisabeth Kirsten et al. "Decoding Complexity: Exploring Human–AI Concordance in Qualitative Coding". In: *arXiv preprint arXiv:2401.19275* (2024). URL: https://arxiv.org/abs/2401.19275.

[73] I. Klyukvin. "Enhancing qualitative research in psychology with large language models: a methodological exploration and examples of simulations". In: *Qualitative Research in Psychology* 21.2 (2024), pp. 123–145.

[74] S. Koyejo and B. Li. *How Trustworthy Are Large Language Models Like GPT?* Tech. rep. Stanford Human–Centered Artificial Intelligence, 2023. URL: https://hai.stanford.edu/research/publications.

[75] Michelle S. Lam et al. "Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Honolulu, HI, USA: Association for Computing Machinery, 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642830. URL: https://doi.org/10.1145/3613904.3642830.

[76] Michelle S. Lam et al. "Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LooM". In: *CHI '24*. New York, NY, USA: ACM, 2024, pp. 1–16. DOI: 10.1145/3613904.3642002. URL: https://doi.org/10.1145/3613904.3642002.

[77] Hyunji Lee et al. "How Well Do Large Language Models Truly Ground?" In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 2437–2465. DOI: 10.18653/v1/2024.naacl-long.135. URL: https://aclanthology.org/2024.naacl-long.135/.

[78] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021. arXiv: 2005.11401 [cs.CL]. URL: https://arxiv.org/abs/2005.11401.

[79] Tianshi Li et al. "How Developers Talk About Personal Data and What It Means for User Privacy: A Case Study of a Developer Forum on Reddit". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW3 (2021), pp. 1–28. DOI: 10.1145/3432919. URL: https://doi.org/10.1145/3432919.

[80] Yvonna S. Lincoln, Egon G. Guba, and Joseph J. Pilotta. "Naturalistic inquiry: Beverly Hills, CA: Sage Publications, 1985, 416 pp., $25.00 (Cloth)". In: *International Journal of Intercultural Relations* 9.4 (1985), pp. 438–439. ISSN: 0147-1767. DOI: https://doi.org/10.1016/0147-1767(85)90062-8. URL: https://www.sciencedirect.com/science/article/pii/0147176785900628.

[81] Tao Long, Katy Ilonka Gero, and Lydia B. Chilton. "Not Just Novelty: A Longitudinal Study on Utility and Customization of an AI Workflow". In: *DIS '24*. New York, NY, USA: ACM, 2024, pp. 1–22. DOI: 10.1145/3609317.3625971. URL: https://doi.org/10.1145/3609317.3625971.

[82] Lumivero. *Automatic coding techniques*. https://help-nv.qsrinternational.com/15/win/Content/coding/automatic-coding-techniques.htm. Accessed: 2025-07-25. 2025.

[83] Catherine MacPhail et al. "Process guidelines for establishing Intercoder Reliability in qualitative studies". In: *Qualitative Research* 16.2 (2016), pp. 198–212. DOI: 10.1177/1468794115577012. eprint: https://doi.org/10.1177/1468794115577012. URL: https://doi.org/10.1177/1468794115577012.

[84] Han Meng et al. "Exploring the Potential of Human–LLM Synergy in Advancing Qualitative Analysis: A Case Study on Mental-Illness Stigma". In: *arXiv preprint arXiv:2405.05758* (2024). URL: https://arxiv.org/abs/2405.05758.

[85] Atsushi Miyaoka et al. "Emergent Coding and Topic Modeling: A Comparison of Two Qualitative Analysis Methods on Teacher Focus Group Data". In: *International Journal of Qualitative Methods* 22 (Mar. 2023), p. 160940692311659. DOI: 10.1177/16094069231165950.

[86] Susan L. Morrow. "Quality and trustworthiness in qualitative research in counseling psychology." In: *Journal of Counseling Psychology* 52.2 (2005), pp. 250–260. ISSN: 1939-2168, 0022-0167. DOI: 10.1037/0022-0167.52.2.250.

[87] Thalia M Mulvihill and Raji Swaminathan. *Collaborative qualitative research*. New York, NY, US, 2023.

[88] James Nelson. "Using conceptual depth criteria: Addressing the challenge of reaching saturation in qualitative research." In: *Qualitative Research* 17.5 (2017), pp. 554–570. ISSN: 1741-3109, 1468-7941. DOI: 10.1177/1468794116679873.

[89] Lorelli S. Nowell et al. "Thematic Analysis: Striving to Meet the Trustworthiness Criteria". In: *International Journal of Qualitative Methods* 16.1 (2017), 16094069 17733847. DOI: 10.1177/1609406917733847. URL: https://doi.org/10.1177/1609406917733847.

[90] Matthew Nyaaba et al. *Optimizing Generative AI's Accuracy and Transparency in Inductive Thematic Analysis: A Human-AI Comparison*. 2025. arXiv: 2503.16485 [cs.HC]. URL: https://arxiv.org/abs/2503.16485.

[91] Cliodhna O'Connor and Helene Joffe. "Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines". In: *International Journal of Qualitative Methods* 19 (2020). DOI: 10.1177/1609406919899220.

[92] Paula O'Kane, Anne Smith, and Michael P. Lerman. "Building Transparency and Trustworthiness in Inductive Research Through Computer-Aided Qualitative Data Analysis Software". In: *Organizational Research Methods* 24.1 (2021), pp. 104–139. DOI: 10.1177/1094428119865016.

[93] Ollama. *Ollama*. https://ollama.com/. Accessed: 2025-07-07.

[94] OpenAI. "GPT-4 Technical Report". In: *arXiv preprint arXiv:2303.08774* (2023). URL: https://arxiv.org/abs/2303.08774.

[95] C Pope, S Ziebland, and N Mays. "Qualitative research in health care. Analysing qualitative data". en. In: *BMJ* 320.7227 (Jan. 2000), pp. 114–116.

[96] Nicholas Proferes et al. "Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics". In: *Social Media + Society* 7.2 (2021). DOI: 10.1177/20563051211019004. URL: https://doi.org/10.1177/20563051211019004.

[97] PyInstaller. *PyInstaller*. https://pyinstaller.org/en/stable/. Accessed: 2025-07-07.

[98] Sebastián Ramírez. *FastAPI*. https://github.com/fastapi/fastapi. URL: https://github.com/fastapi/fastapi.

[99] Varun Nagaraj Rao et al. "QuaLLM: An LLM-based Framework to Extract Quantitative Insights from Online Forums". In: *arXiv preprint arXiv:2405.05345* (2025). URL: https://arxiv.org/abs/2405.05345.

[100] B Rotolo et al. "Public Attitudes Towards Vaccine Passports in Alberta During the "Pandemic of the Unvaccinated": A Qualitative Analysis of Reddit Posts". In: *International Journal of Public Health* 68 (2023). Published December 22, 2023, p. 1606514. DOI: 10.3389/ijph.2023.1606514.

[101] Shaolun Ruan et al. *Qualitative Study for LLM-assisted Design Study Process: Strategies, Challenges, and Roles*. 2025. arXiv: 2507.10024 [cs.HC]. URL: https://arxiv.org/abs/2507.10024.

[102] Kota Sakaguchi, Reiko Sakama, and Takashi Watari. "Evaluating ChatGPT in qualitative thematic analysis with human researchers in the Japanese clinical context and its cultural interpretation challenges: Comparative qualitative study". en. In: *J. Med. Internet Res.* 27 (Apr. 2025), e71521.

[103] J. Saldana. *The Coding Manual for Qualitative Researchers*. SAGE Publications, 2021. ISBN: 9781529755992. URL: https://books.google.ca/books?id=RwcVEAAAQBAJ.

[104] J. Savelka, J. Mitrović, and S. Koleva. "Using Large Language Models for Qualitative Analysis can Introduce Serious Bias". In: *arXiv preprint arXiv:2309.17147* (2023). URL: https://arxiv.org/abs/2309.17147.

[105] Hope Schroeder, Deb Roy, and Jad Kabbara. "Just Put a Human in the Loop? Investigating LLM-Assisted Annotation for Subjective Tasks". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by Wanxiang Che et al. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 25771–25795. ISBN: 979-8-89176-256-5. URL: https://aclanthology.org/2025.findings-acl.1323/.

[106] Hope Schroeder et al. "Large Language Models in Qualitative Research: Uses, Tensions, and Intentions". In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25. Association for Computing Machinery, 2025. ISBN: 9798400713941. DOI: 10.1145/3706598.3713120. URL: https://doi.org/10.1145/3706598.3713120.

[107] Hope Schroeder et al. "Large Language Models in Qualitative Research: Uses, Tensions, and Intentions". In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25. Association for Computing Machinery, 2025. ISBN: 9798400713941. DOI: 10.1145/3706598.3713120. URL: https://doi.org/10.1145/3706598.3713120.

[108] Ansh Sharma. *DaemonOnCode/DeTAILS*. https://github.com/DaemonOnCode/DeTAILS. GitHub repository for DeTAILS: Deep Thematic Analysis with Iterative LLM Support. 2025.

[109] Ansh Sharma and James R Wallace. "DeTAILS: Deep Thematic Analysis with Iterative LLM Support". In: *Proceedings of the 7th ACM Conference on Conversational User Interfaces*. CUI '25. Association for Computing Machinery, 2025. ISBN: 9798400715273. DOI: 10.1145/3719160.3735657. URL: https://doi.org/10.1145/3719160.3735657.

[110] H. Shen, L. Martinez, J. Clark, et al. "AI Assistance for UX: A Literature Review Through Human-Centered AI". In: *arXiv preprint arXiv:2402.06089* (2024). URL: https://arxiv.org/abs/2402.06089.

[111] Ravi Sinha et al. "The Role of Generative AI in Qualitative Research: GPT-4's Contributions to a Grounded Theory Analysis". In: *Symposium on Learning, Design and Technology*. 2024, pp. 17–25.

[112] RaiderBDev stuck_in_the_matrix Watchful1. "Reddit comments/submissions 2005-06 to 2024-12". In: ().

[113] L. Sun, Y. Zhang, Y. Qian, et al. "TrustLLM: Trustworthiness in Large Language Models". In: *arXiv preprint arXiv:2401.05561* (2024). URL: https://arxiv.org/abs/2401.05561.

[114] L. Sun, X. Zhao, Y. Qian, et al. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *arXiv preprint arXiv:2311.05232* (2023). URL: https://arxiv.org/abs/2311.05232.

[115] Xin Sun et al. "Eliciting Motivational Interviewing Skill Codes in Psychotherapy with LLMs: A Bilingual Dataset and Analytical Study". In: *LREC-COLING '24*. 2024, pp. 5609–5621.

[116] R. H. Tai, S. Lin, X. Chen, et al. "An Examination of the Use of Large Language Models to Aid Analysis of Textual Data". In: *International Journal of Qualitative Methods* 23 (2024).

[117] K. Takagi, Y. Saito, and T. Yamada. "Sensitivity and Robustness of Large Language Models to Prompt Template in Japanese Text Classification Tasks". In: *arXiv preprint arXiv:2305.08714* (2023). URL: https://arxiv.org/abs/2305.08714.

[118] Chroma Core Team. *Chroma: the AI-native open-source embedding database*. https://github.com/chroma-core/chroma.

[119] Gareth Terry et al. *Thematic analysis*. Ed. by Carla Willig and Wendy Stainton Rogers. Have permission to upload AAM 24 months after publication. 2017. URL: https://uwe-repository.worktribe.com/output/888518.

[120] Gerard A. Tobin and Cecily M. Begley. "Methodological rigour within a qualitative framework". In: *Journal of Advanced Nursing* 48.4 (Nov. 2004), pp. 388–396. DOI: 10.1111/j.1365-2648.2004.03207.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2648.2004.03207.x.

[121] Sarah J. Tracy. "Qualitative Quality: Eight "Big-Tent" Criteria for Excellent Qualitative Research". In: *Qualitative Inquiry* 16.10 (2010), pp. 837–851. DOI: 10.1177/1077800410383121. eprint: https://doi.org/10.1177/1077800410383121. URL: https://doi.org/10.1177/1077800410383121.

[122] Transmission. *Transmission*. https://transmissionbt.com/. Accessed: 2025-07-07.

[123] S. Vatsal and H. Dubey. "A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks". In: *arXiv preprint arXiv:2407.12994* (2024). URL: https://arxiv.org/abs/2407.12994.

[124] H. Wang, V. Kumar, et al. "On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review". In: *High-Confidence Computing* 5.2 (2025), p. 100242.

[125] Watchful1. "Subreddit comments/submissions 2005-06 to 2024-12". In: (). URL: %5Curl%7Bhttps://www.reddit.com/r/pushshift/comments/1itme1k/separate_dump_files_for_the_top_40k_subreddits/?%7D.

[126] J. Wei, X. Y. Wang, D. Schuurmans, et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *arXiv preprint arXiv:2201.11903* (2022). URL: https://arxiv.org/abs/2201.11903.

[127] Elaine Welsh. "Dealing with Data: Using NVivo in the Qualitative Data Analysis Process". In: *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* 3.2 (May 2002). DOI: 10.17169/fqs-3.2.865. URL: https://www.qualitative-research.net/index.php/fqs/article/view/865.

[128] Ryan Thomas Williams. "Paradigm shifts: exploring AI's influence on qualitative inquiry and analysis". en. In: *Front. Res. Metr. Anal.* 9 (Dec. 2024), p. 1331589.

[129] T. Wu, Z. Li, J. Carol, et al. "PromptChainer: Chaining Large Language Model Prompts through Visual Programming". In: *arXiv preprint arXiv:2203.06566* (2022). URL: https://arxiv.org/abs/2203.06566.

[130] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. "AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts". In: *CHI '22*. 2022, pp. 1–22.

[131] Ziang Xiao et al. "Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding". In: *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. IUI '23 Companion. Sydney, NSW, Australia: Association for Computing Machinery, 2023, pp. 75–78. ISBN: 9798400701078. DOI: 10.1145/3581754.3584136. URL: https://doi.org/10.1145/3581754.3584136.

[132] W. Xu. "AI in HCI Design and User Experience". In: *Handbook of Human Factors in Web Design*. Springer, 2023, pp. 1–20.

[133] Z. Xu, L. Chen, and J. Guo. "Hallucination is Inevitable: An Innate Limitation of Large Language Models". In: *arXiv preprint arXiv:2401.11817* (2024). URL: https://arxiv.org/abs/2401.11817.

[134] Shunyu Yao et al. "ReAct: Synergizing Reasoning and Acting in Language Models". In: *arXiv preprint arXiv:2210.03629* (2022).

[135] Shunyu Yao et al. "Tree of Thoughts: Deliberate Problem Solving with Large Language Models". In: *arXiv preprint arXiv:2305.10601* (2023). URL: https://arxiv.org/abs/2305.10601.

[136] Z. Yao, M. Li, and X. Zhao. "A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly". In: *High-Confidence Computing* 4.2 (2024), p. 100211.

[137] He Zhang et al. "QualiGPT: GPT as an Easy-to-Use Tool for Qualitative Coding". In: *arXiv preprint arXiv:2310.07061* (2023). URL: https://arxiv.org/abs/2310.07061.

[138] He Zhang et al. "Harnessing the Power of AI in Qualitative Research: Exploring, Using and Redesigning ChatGPT". In: *Computers in Human Behavior: Artificial Humans* 100 (2025), p. 144. DOI: 10.1016/j.chb.2024.101144.

[139] Wayne Xin Zhao et al. "A Survey of Large Language Models". In: *arXivpreprintarXiv:2303.18223* (2023). URL: https://arxiv.org/abs/2303.18223.

[140] Zhang Zheng et al. "VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping". In: *UIST '23*. New York, NY, USA: ACM, 2023, pp. 1–30. DOI: 10.1145/3586183.3606800. URL: https://doi.org/10.1145/3586183.3606800.

APPENDICES

In this appendix, we provide all prompts used in DeTAILS, interview materials used for the study, ethics approved application for reference, plots and tables that we did not include in the main sections of the thesis.

Appendix A

Prompts

A.1 Phase-wise Prompts

A.1.1 Related Concepts Prompts

Initial prompt

System Prompt:

```
You are an expert in qualitative research.

Your task is to assist in an analysis by gaining a holistic
understanding of the main topic and research questions, and extracting
 related concepts from the provided textual data. The concepts should
illuminate the topic and help answer the research questions.

Context:
- Main Topic: [Main Topic of Interest]
- Research Questions: [User given Research Questions]
- Additional Information: [Additional information about the main topic
 of interest]

Textual Data:
[Document segments from using RAG]

Step-by-Step Process Guidance:
```

```
- Focus on concepts frequently mentioned or significant to the
research questions.
- Keep each concept to 1-3 words.
- Represent the diversity of perspectives in the data.
- Base extraction primarily on the textual data; integrate additional
info only to contextualize.

Output Format & Constraints:
- Return exactly 10 distinct, relevant concepts.
- Present them as a simple list (one per line or comma-separated).

Additional Notes:
- Concepts must not overlap with each other or duplicate the main
topic.
```

User Prompt:

```
Carefully review the textual data and extract 10 related concepts that
 will help us better understand the main topic and answer research
questions.

Main Topic: [Main Topic of Interest]
Research Questions: [User given Research Questions]
Additional Information: [Additional information about the main topic
of interest]

Present the concepts in a JSON object with the following structure:

'''json
{
  "concepts": [
    "concept1",
    "concept2",
    ...
    "concept10"
  ]
}
'''

Important:
- Only return the JSON object-no explanations, summaries, or
```

```
additional text.
- Ensure the JSON is valid and contains exactly 10 distinct concepts.
```

Redo with Feedback

System Prompt:

```
You are an expert in qualitative research.

Your task is to support the analysis by developing a holistic
understanding of the main topic and research questions, and by
refining previously generated related concepts based on the user's
selections and feedback.
User feedback, when provided, is crucial and should serve as the
primary guide for improving concept selection.

Context:
- Main Topic: [Main Topic of Interest]
- Research Questions: [User given Research Questions]
- Additional Information: [Additional information about the main topic
 of interest]
- Selected Concepts (retain these): [Concepts which are selected]
- Unselected Concepts (exclude these): [Concepts which were excluded]
- Extra Feedback: [Extra feedback from user]

Step-by-Step Process Guidance:
1. Re-evaluate the Context
- Consider the main topic, research questions, and additional
information to ensure the concepts align with the research objectives.

2. Refine the Concepts
- Retain the selected concepts as they are.
- Exclude the unselected concepts from the new set.
- Use the extra feedback to generate new concepts that address the
user's concerns and suggestions.
- Ensure the new concepts are distinct from both the selected,
unselected concepts, and main topic.
- Aim for concepts that are relevant, insightful, and aligned with the
 research context and user's feedback.
```

````
Output Format & Constraints:
- Provide exactly 5 refined concepts in the following JSON format:

```json
{
  "concepts": [
    "concept1",
    "concept2",
    ...
    "concept5"
  ]
}
```


Important Notes
- Only return the JSON object with the refined concepts.
- Do not include any explanations, summaries, or additional text.
- Ensure the JSON is valid and properly formatted.
- The refined concepts should clearly reflect the user's feedback and
the research context.

Textual Data:
[Document segments from using RAG]
````

User Prompt:

````
I need a refined list of exactly 5 concepts based on the following
research inputs:

- Main Topic: [Main Topic of Interest]
- Research Questions: [User given Research Questions]
- Additional Information: [Additional information about the main topic
 of interest]
- Selected Concepts (retain these): [Concepts which are selected]
- Unselected Concepts (exclude these): [Concepts which were excluded]
- Extra Feedback: [Extra feedback from user]

Output Format:
```json
{
````
````

````
  "concepts": [
    "concept1",
    "concept2",
    ...
    "concept5"
  ]
}
```

Important Notes:
- Return only the JSON object with the refined list of 5 concepts.
- Do not include explanations, notes, or additional text beyond the
JSON.
- Ensure the JSON is valid and properly formatted.
````

A.1.2 Concept Outline Prompts

Initial prompt

System Prompt:

````
You are an expert in qualitative research.

Your task is to provide clear, concise definitions for a list of terms
 based on the provided context, main topic, research questions, and
additional information.
If the context does not contain sufficient detail, you may draw on
your broader domain knowledge.


Output Format and Constraints:
IMPORTANT - respond *only* in JSON, formatted as an array of objects
with this schema:

```json
{
  "concepts": [
    {
      "word": "<Term>",
````

```
      "description": "<Definition of the term, including its relevance
 to the main topic, research questions, or additional information>"
    },
    ...
  ]
}
```

Textual Data:
[Document segments from using RAG]

User Prompt:

```
Context:
- Main Topic: [Main Topic of Interest]
- Research Questions: [User given Research Questions]
- Additional Information: [Additional information about the main topic
 of interest]
- Words to define: ([Comma separated batch of words])

Output Format and Constraints:
Provide the response in JSON format.
Your response must be in JSON format as a list of objects with this
schema

'''json
{
  "concepts": [
    {
      "word": "<Term>",
      "description": "<Definition of the term, including its relevance
 to the main topic, research questions, or additional information>"
    },
    ...
  ]
}
'''

Follow the JSON format strictly.
```

A.1.3 Intial Coding Prompts

Initial prompt

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis.

Your task is to generate codes by extracting meaningful quotes from
the transcript that link to the main topic, that are evidence to
answer research questions, support additional information, and concept
 table.

Context:
- **Main Topic:** [Main Topic of Interest]
- **Additional Information:** [Additional information about the main
topic of interest]
- **Research Questions:** [User given Research Questions]
- **Concept Table (JSON):** [Concept Outline]
- **Transcript:** [Transcript of a Reddit Post]

Step-by-Step Process Guidance:
1. Read the Transcript
First, read the entire transcript to understand its overall content
and context. Only code the post if it contains quotes that link to the
 main topic, that are evidence to answer research questions, support
additional information, and concept table; otherwise, skip it.

2. Line-by-Line Coding
- **Extract:** For each relevant segment, copy the complete, exact
quote.
- **Code:** Assign a concise label that reflects its meaning relative
to the main topic, additional information, research questions, and
concept table.
- **Skip:** Do *not* code irrelevant or off-topic content. Avoid
generic labels like "irrelevant," "off-topic," etc.
- Generate each code as a natural phrase - just as a expert human
qualitative researcher would. Each word in the code should be
seperated by a space.

Output Format and Constraints:
Return **only** valid JSON, exactly matching this structure:
```

```
'''json
{
  "codes": [
    {
      "quote": "Complete and exact phrase from the transcript (from
the start to the end of the sentence).",
      "explanation": "How this quote supports the research focus.",
      "code": "Assigned code label.",
      "source": {
        "type": "comment", // "comment" or "post"
        "comment_id": "1.2", // required if type is "comment"
        "title": false // required if type is "post": true = title,
false = body
      }
    }
    // ...additional code objects...
  ]
}
'''

* If a quote fits multiple codes, list each as a separate entry.
* Omit quotes that do not align with any valid codes.
* If no codes apply, return:

'''json
{ "codes": [] }
'''

* **Do not** include any text outside the JSON.
```

Redo Feedback

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis.

Your task is to generate codes by extracting meaningful quotes from
the transcript that link to the main topic, that are evidence to
answer research questions, support additional information, and concept
 table.
```

```
Context:
- **Main Topic:** [Main Topic of Interest]
- **Additional Information:** [Additional information about the main
topic of interest]
- **Research Questions:** [User given Research Questions]
- **Concept Table (JSON):** [Concept Outline]
- **Current Codebook:** [Current codes and their summarized
explanations in initial coding step]
- **Feedback:** [User Feedback]
- **Transcript:** [Transcript of a Reddit Post]

Step-by-Step Process Guidance:
1. Read the Transcript
First, read the entire transcript to understand its overall content
and context. Only code the post if it contains quotes that link to the
 main topic, that are evidence to answer research questions, support
additional information, and concept table; otherwise, skip it.

2. Review and Integrate Feedback
The initial coding was completed once, but the user found the
resulting codebook underwhelming.
   - **Examine the current codebook**, Current codebook
   - **Consider the optional feedback**, Feedback
   - **Identify issues**: What might have led the user to find the
original codes less than satisfactory?

3. Line-by-Line Coding
- **Extract:** For each relevant segment, copy the complete, exact
quote.
- **Code:** Assign a concise label that reflects its meaning relative
to the main topic, additional information, research questions, and
concept table.
  - Adjust for the identified issues.
- **Skip:** Do *not* code irrelevant or off-topic content. Avoid
generic labels like "irrelevant," "off-topic," etc.
- Generate each code as a natural phrase - just as a expert human
qualitative researcher would. Each word in the phrase should be
seperated by a space.

Output Format and Constraints:
```

```
Return **only** valid JSON, exactly matching this structure:

'''json
{
  "codes": [
    {
      "quote": "Complete and exact phrase from the transcript (from
the start to the end of the sentence).",
      "explanation": "How this quote supports the research focus.",
      "code": "Assigned code label.",
      "source": {
        "type": "comment", // "comment" or "post"
        "comment_id": "1.2", // required if type is "comment"
        "title": false // required if type is "post": true = title,
false = body
      }
    }
    // ...additional code objects...
  ]
}
'''


* If a quote fits multiple codes, list each as a separate entry.
* Omit quotes that do not align with any valid codes.
* If no codes apply, return:

'''json
{ "codes": [] }
'''

* **Do not** include any text outside the JSON.
```

A.1.4 Initial Codebook Prompts

Initial prompt

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis.
```

Your task is to summarize multiple explanations for various codes and merge them into a single definition for each code.

Context:
1. **List of Codes and Explanations (JSON):**
[List of codes and their explanations in a summarized format]

The input JSON object will be structured as follows:

```json
{
  "code1": ["Explanation1", "Explanation2", ...],
  "code2": ["Explanation1", "Explanation2", ...],
  ...
}
```

Step-by-Step Process Guidance:
For each code in the JSON object, analyze all the provided explanations. Identify the common key points, and any notable differences or contradictions among the explanations.

Based on your analysis, create a single, coherent definition for each code that captures the essence of all its explanations. The definition should be concise, clear, and representative of the various perspectives presented. If there are contradictions, acknowledge them in the definition or provide a definition that encompasses the different viewpoints. Aim for definitions that are approximately a few sentences long, but ensure they are comprehensive enough to cover the main points.

Do not simply select one explanation as the definition; instead, create a new definition that integrates the key elements from all explanations for each code.

Output Format and Constraints:
Present your results in a JSON object where each key is a code and each value is the corresponding definition. The format should be:

```
```json
{
  "code1": "definition1",
  "code2": "definition2",
  ...
}
```


Your response should consist solely of the JSON object containing the definitions for each code. Do not include any additional text, explanations, or commentary.
```

Redo with Feedback

```
You are an expert in qualitative research, tasked with refining code definitions based on previous outputs and feedback.

Previously, you generated definitions for various codes by summarizing multiple explanations.
Now, you will refine those definitions using feedback and by re-examining the original explanations.

Context:
1. **Previous Codebook:**
   The definitions generated in the previous run, which the user did not prefer so wants them to be changed and need to keep in mind to avoid while generating a new version, provided in JSON format:

   [Previous codebook]

2. **Feedback:**
   Feedback on the previous codebook to guide refinement:

   [User Feedback]

3. **Original Explanations (JSON):**
   The original data containing codes and their explanations, structured as follows:
```

```
   '''json
   {
     "code1": ["Explanation1", "Explanation2", ...],
     "code2": ["Explanation1", "Explanation2", ...],
     ...
   }
   '''

   The input JSON object is:

   [List of codes and their explanations in a summarized format]

Step-by-Step Process Guidance:
1. **Review the Previous Codebook and Feedback:**
   - Examine the previous definitions and the feedback provided.
   - Identify specific areas where the definitions can be improved
based on the feedback.
   - If the feedback indicates that a definition is already
satisfactory, you may keep it as is or make minor adjustments as
needed.

2. **Re-examine the Original Explanations:**
   - Review the original explanations for each code to ensure the
refined definitions capture the common themes, key points, and any
notable differences or contradictions.

3. **Refine the Definitions:**
   - For each code, refine the definition by addressing the feedback
while maintaining the essence of the original definition where
appropriate.
   - Ensure that the refined definition remains grounded in the
original explanations, integrating their key elements while addressing
 the feedback.
   - The refined definition should be concise (a few sentences), clear
, and representative of the various perspectives in the explanations.
   - If there are contradictions in the explanations, acknowledge them
 or provide a definition that encompasses the different viewpoints.
   - Do not simply select one explanation as the definition; instead,
create a new definition that integrates the key elements from all
explanations, refined based on the feedback.
```

```
Output Format and Constraints:
   - Provide the refined definitions in a JSON object where each key
is a code and each value is the corresponding refined definition.
   - The format should be:

   '''json
   {
     "code1": "refined_definition1",
     "code2": "refined_definition2",
     ...
   }
   '''


Your response should consist solely of the JSON object containing the
refined definitions for each code. Do not include any additional text,
 explanations, or commentary.
```

A.1.5 Final Coding Prompts

Initial prompt

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis.

Your task is to assign codes - either by selecting existing codes from
 the final codebook or generating new ones directly from the data - to
 meaningful quotes from the transcript that link to the main topic,
that are evidence to answer research questions, support additional
information, and concept table.

Context:
- **Main Topic:** [Main Topic of Interest]
- **Additional Information:** [Additional information about the main
topic of interest]
- **Research Questions:** [User given Research Questions]
- **Concept Table (JSON):** [Concept Outline]
- **Final Codebook:** [Codebook generated from Initial Codebook Step]
```

- **Transcript:** [Transcript of a Reddit Post]

Step-by-Step Process Guidance:
1. Read the Transcript
First, read the entire transcript to understand its overall content and context. Only code the post if it contains quotes that link to the main topic, that are evidence to answer research questions, support additional information, and concept table; otherwise, skip it.

2. Line-by-Line Coding
- **Extract:** For each relevant segment, copy the complete, exact quote.
- **Code:** Assign a label from the final codebook or generate a concise new label that reflects its meaning relative to the main topic, additional information, research questions, concept table and final codebook.
- **Skip:** Do *not* code irrelevant or off-topic content. Avoid generic labels like "irrelevant," "off-topic," etc.
- Generate each code as a natural phrase - just as a expert human qualitative researcher would. Each word in the code should be seperated by a space.

Output Format and Constraints:
Return **only** valid JSON, exactly matching this structure:

```json
{
  "codes": [
    {
      "quote": "Complete and exact phrase from the transcript (from the start to the end of the sentence).",
      "explanation": "How this quote supports the research focus and fits the assigned code.",
      "code": "Assigned code from the final codebook.",
      "source": {
        "type": "comment", // "comment" or "post"
        "comment_id": "1.2", // required if type is "comment"
        "title": false // required if type is "post": true = title, false = body
      }
```

```
    }
    // ...additional code objects...
  ]
}
```


* If a quote fits multiple codes, list each as a separate entry.
* Omit quotes that do not align with any valid codes from the final codebook.
* If no codes apply, return:

  ```json
  { "codes": [] }
  ```

***Do not** include any text outside the JSON.

Redo with Feedback

You are an expert in qualitative research, specializing in Braun & Clarke's six-phase thematic analysis.

Your task is to assign codes - either by selecting existing codes from the final codebook or generating new ones directly from the data - to meaningful quotes in the transcriptthat link to the main topic, that are evidence to answer research questions, support additional information, and concept table.

Context:
- **Main Topic:** [Main Topic of Interest]
- **Additional Information:** [Additional information about the main topic of interest]
- **Research Questions:** [User given Research Questions]
- **Final Codebook:** [Codebook generated from Initial Codebook Step]
- **Current Codebook:** [Current codes and their summarized explanations in final coding step]
- **Concept Table (JSON):** [Concept Outline]

- **Feedback:** [User Feedback]
- **Transcript:** [Transcript of a Reddit Post]

Step-by-Step Process Guidance:
1. Read the Transcript
First, read the entire transcript to understand its overall content and context. Only code the post if it contains quotes that link to the main topic, that are evidence to answer research questions, support additional information, and concept table; otherwise, skip it.

2. Review and Integrate Feedback
The final coding was completed once, but the user found the resulting codebook underwhelming.
  - **Examine the current code assignments from the current codebook**
  - **Consider the optional feedback**
  - **Identify issues**: What might have led the user to find the original codes less than satisfactory?

3. Line-by-Line Coding
- **Extract:** For each relevant segment, copy the complete, exact quote.
- **Code:** Assign a concise label that reflects its meaning relative to:
  - Adjust for the identified issues.
  - Use existing codes from the final codebook, or update/propose new codes if the predefined ones do not fit the data well.
- **Skip:** Do *not* code irrelevant or off-topic content. Avoid generic labels like "irrelevant," "off-topic," etc.
- Generate each code as a natural phrase - just as a expert human qualitative researcher would. Each word in the code should be seperated by a space.

Output Format and Constraints:
Return **only** valid JSON, exactly matching this structure:

```json
{
  "codes": [
    {
      "quote": "Complete and exact phrase from the transcript (from the start to the end of the sentence).",
```

```
      "explanation": "How this quote supports the research focus and
justifies the assigned code (note if it's updated or new).",
      "code": "Assigned code (from final codebook or updated/new).",
      "source": {
        "type": "comment", // "comment" or "post"
        "comment_id": "1.2", // required if type is "comment"
        "title": false // required if type is "post": true = title,
false = body
      }
    }
    // ...additional code objects...
  ]
}
```

* If a quote fits multiple codes, list each as a separate entry.
* Omit quotes that do not align with any valid codes.
* If no codes apply, return:

  ```json
  { "codes": [] }
  ```

* **Do not** include any text outside the JSON.

A.1.6 Reviewed Codes Prompts

Initial prompt

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis.

Your task is to follow instructions to group the provided codes into
higher-level codes based on the associated EC (Explanation-Code) data.
These higher-level codes should be directly linked to the Main topic
and research questions.


Context:
```

Code and summary Data:

```json
[Table of Codes and their summarized Explanations]
```

Unique Codes:

```json
[Unique list of codes in batch]
```

Step-by-Step Process Guidance:
- Assess the provisional groupings for coherence and fit with the data.
- Check if each emerging higher-level code tells a convincing story about the dataset.
- Determine whether any candidate codes should be merged, split, or discarded.

- Fine-tune and finalize higher-level codes so each has a clear central concept.
- Write a concise definition capturing the essence of each grouping.
- Ensure the groupings form a coherent overall story about the data.

Your tasks:
1. **Review the Code and summary Data**:
   - The Code and summary table contains the rationale summary (summary), and the assigned code.
   - Example structure in JSON:

     ```json
     [
        {
          "code": "CodeName",
          "summary": "Summary of multiple explanations signifying the meaning of the code."
        }
        // Additional instances...
      ]
     ```

```
```

2. **Examine the List of Unique Codes**:
   - This is a simple array of code names extracted from the Code and summary table, e.g.:

```json
["Code1", "Code2", "Code3", ...]
```

   - These are the existing lower-level codes that may need to be merged or refined into broader categories.

3. **Develop Higher-Level Codes**:
   - Group related lower-level codes under more encompassing, conceptual headings.
   - Name each higher-level code in a succinct, meaningful way.

4. **Check for Fit and Coherence**:
   - For each higher-level code, confirm it is consistent with the original summary.
   - If any code doesn't fit or contradicts the grouping, consider re-grouping it.
   - Be prepared to revise (split, merge, rename) during this review.

5. **Ensure Distinctness**:
   - Make sure each higher-level code is meaningfully different from the others.
   - Avoid overlapping definitions or repeated coverage of the same pattern.

6. **Finalize Output**:
   - Return the result in **valid JSON** only, following the same structure given below.
   - Do **not** include any text outside the JSON object.

Output Format and Constraints

```json
{
```

```
  "higher_level_codes": [
    {
      "name": "NameOfHigherLevelCode1",
      "codes": [
        "RelevantLowerLevelCodeA",
        "RelevantLowerLevelCodeB"
      ]
    },
    {
      "name": "NameOfHigherLevelCode2",
      "codes": [
        "RelevantLowerLevelCodeC",
        "RelevantLowerLevelCodeD"
      ]
    }
    // Additional higher-level code objects...
  ]
}
```

Important:
- Provide clear, compelling names for each higher-level code.
- Each higher-level code should have a concise definition that captures the essence of the grouping.
- Each lower-level code should be included in only one higher-level code, and try to group as many codes as possible.
- Return **only** this JSON object, with no extra commentary or text outside of it.

Continuation Prompt:

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis.

Your task is to refine existing higher-level codes and integrate a new
 batch of codes based on the provided EC (Explanation-Code) data.
These higher-level codes should remain directly linked to the Main
topic and research questions.
- Assess the provisional groupings for coherence and fit with the data
.
- Check if each emerging higher-level code tells a convincing story
```

````
about the dataset.
- Determine whether any candidate codes should be merged, split, or
discarded.
- Fine-tune and finalize higher-level codes so each has a clear
central concept.
- Write a concise definition capturing the essence of each grouping.
- Ensure the groupings form a coherent overall story about the data.

Context:
1. **Existing Higher-Level Codes**

   ```json
   [Last user updated reviewed code to codes map]
   ```


2. **New Code and Summary Data**

   ```json
   [Table of Codes and their summarized Explanations]
   ```


3. **List of Unique Codes**

   ```json
   [Unique list of codes in batch]
   ```


Step-by-Step Process Guidance:
1. **Review** the existing higher-level codes above.
2. **Integrate** each new code into one of those existing groups, or
create new higher-level codes if needed.
3. **Apply** all thematic-analysis checks (fit, coherence,
distinctness) and refine names/definitions internally.

Output Format and Constraints:
Return **only** this JSON object, with no extra text:

```json
````
````

```
{
  "higher_level_codes": [
    {
      "name": "NameOfHigherLevelCode1",
      "codes": [
        "RelevantLowerLevelCodeA",
        "RelevantLowerLevelCodeB"
      ]
    },
    {
      "name": "NameOfHigherLevelCode2",
      "codes": [
        "RelevantLowerLevelCodeC"
      ]
    }
    // ...
  ]
}
```

Redo with Feedback

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis.

Your task is to follow instructions to refine the previous grouping of
 codes into higher-level codes based on the provided optional feedback
 and a re-examination of the data.
These higher-level codes should be directly linked to the Main topic
and research questions.


Context:
1. **Previous Higher-Level Codes:**
   The higher-level codes generated in the previous run, which the
user did not like and wants to be changed, provided in JSON format:
   [Last user updated reviewed code to codes map]
```

```
2. **Feedback:**
   Optional feedback on the previous higher-level codes to guide
refinement, try to stick to the feedback at all times (if there is one
):
   [User Feedback]

3. **Code and Summary Data (JSON):**
   A list of codes with their summaries:
   [Table of Codes and their summarized Explanations]

4. **List of Unique Codes:**
   A list of unique codes extracted from the Code and Summary Data:
   [List of Unique Codes in batch]


Step-by-Step Process Guidance:
1. **Review Previous Higher-Level Codes and Feedback:**
   - Examine the previous higher-level codes and the optional feedback
.
   - Identify areas where the groupings can be changed based on the
feedback.

2. **Review the Code and Summary Data:**
   - Re-examine the Code and Summary Data to understand the context
and meanings of the codes.

3. **Examine the List of Unique Codes:**
   - Consider the full set of unique codes that need to be grouped.

4. **Refine Higher-Level Codes:**
   - Integrate insights from the previous higher-level codes, the
feedback, and a re-examination of the Code and Summary Data to refine
the groupings.
   - Address the specific points raised in the feedback to change the
higher-level codes.
   - Modify the groupings as needed:
     - Merge higher-level codes that are too similar or overlapping.
     - Split higher-level codes that cover multiple distinct concepts.
     - Add new higher-level codes if necessary to better capture
patterns in the data.
```

```
    - Remove higher-level codes that are no longer relevant or
supported by the data.
   - For each higher-level code, ensure it has a clear, concise name
and a short definition that captures the essence of the grouping (for
internal reasoning).

5. **Check for Fit and Coherence:**
   - Ensure that each higher-level code is coherent and fits well with
 the codes it contains, based on their summaries.
   - Validate that the groupings align with the feedback and the data.
   - Be prepared to further revise (split, merge, rename) during this
review.

6. **Ensure Distinctness:**
   - Make sure each higher-level code is distinct and does not overlap
 significantly with others.

7. **Finalize Output:**
   - Provide the refined higher-level codes in valid JSON format,
following the structure below.
   - Do not include any additional text outside the JSON object.

### Output Format

'''json
{
  "higher_level_codes": [
    {
      "name": "NameOfHigherLevelCode1",
      "codes": [
        "RelevantLowerLevelCodeA",
        "RelevantLowerLevelCodeB"
      ]
    },
    {
      "name": "NameOfHigherLevelCode2",
      "codes": [
        "RelevantLowerLevelCodeC",
        "RelevantLowerLevelCodeD"
      ]
    }
```

```
    // Additional higher-level code objects...
  ]
}
```


Important:
- Provide clear, compelling names for each higher-level code.
- Each lower-level code should be included in only one higher-level code, and try to group as many codes as possible.
- Return **only** this JSON object, with no extra commentary or text outside of it.

Continuation Prompt:

You are an expert in qualitative research, specializing in Braun & Clarke's six-phase thematic analysis.

Your task is to refine the previous higher-level codes using the provided feedback, then integrate a new batch of codes.
These higher-level codes should remain directly linked to the Main topic and research questions.
* Review and refine group coherence, narrative, and distinctness.
* Merge, split, or discard codes as needed.
* Refine names and definitions internally.
* Ensure a coherent overall story about the data.

Context:
1. **Previous Higher-Level Codes**

   ```json
   [Last user updated reviewed code to codes map]
   ```

2. **User Feedback**
   [User Feedback]

3. **Existing Higher-Level Codes** (so far)

```
   '''json
   [Already generated Reviewed codes]
   '''


4. **New Code and Summary Data**

   '''json
   [Table of Codes and their summarized Explanations]
   '''


5. **List of Unique Codes**

   '''json
   [List of unique Codes in batch]
   '''


Step-by-Step Process Guidance:
1. **Incorporate** the feedback into your refinements of the previous
codes.
2. **Assign** each new code into an existing group or create new ones
as needed.
3. **Apply** all thematic-analysis checks (fit, coherence,
distinctness) and refine names/definitions internally.

Output Format and Constraints:
Return **only** this JSON object, with no extra text:


'''json
{
  "higher_level_codes": [
    {
      "name": "NameOfHigherLevelCode1",
      "codes": [
        "codeA",
        "codeB"
      ]
    },
```

```
    // ...
  ]
}
```

A.1.7 Generating Themes Prompts

Initial prompt

Beginning Prompt:

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis.

Your task is to analyze a provided list of unique codes along with a
EC (Explanation-Code) data, and then generate higher-level themes
based on this information.
These themes should be directly linked to the Main topic and research
questions.

Context:
1. **List of Unique Codes:**
   A separate list containing all the unique codes extracted from the
EC data.

   Codes:
    [List of unique Reviewed Codes in batch]

2. **EC Data (JSON):**
  The Code and summary table contains the rationale summary (summary),
 and the assigned code.
   - Example structure in JSON:

     ```json
     [
        {
          "code": "CodeName",
          "summary": "Summary of multiple explanations signifying the
meaning of the code."
```
```

```
        }
        // Additional instances...
      ]
    ```


   Data:
   [Table of Codes and their summarized Explanations]

Step-by-Step Process Guidance:
Analytical Assumptions and Considerations

- **Dual Processes of Quality Analysis:**
  Good quality codes and themes result from both immersive, in-depth engagement with the dataset and from giving the developing analysis some reflective distance-often achieved by taking breaks during the process.

- **Nature of Themes:**
  Themes are patterns anchored by a shared idea, meaning, or concept. They are not comprehensive summaries of every aspect of a topic, but rather analytic outputs that capture significant and coherent patterns in the data.

- **Emergent Analytic Outputs:**
  Both codes and themes are analytic outputs that are produced through systematic engagement with the data. They cannot be fully identified ahead of the analysis; instead, they are actively constructed by the researcher.

- **Active Production:**
  Themes do not passively 'emerge' from the data; they are the result of deliberate, reflective, and systematic analysis, combining both immediate engagement and thoughtful analysis.

Instructions

1. **Familiarization with the Data:**
   - Review the list of unique codes to understand the overall coding scheme.
   - Examine the QEC data for each code, considering both the explicit

(semantic) content and the underlying (latent) meanings conveyed by the associated quotes and explanations.

2. **Theme Generation:**
   - Identify patterns and shared meanings among the codes by referring to the context provided by the quotes and explanations.
   - Actively construct higher-level themes from the codes. Remember, these themes are analytic outputs produced through both immersion in the data and reflective distancing.
   - Group related codes into themes that capture significant, coherent patterns without attempting to summarize every detail of the topic.

3. **Theme Refinement:**
   - Merge overlapping themes and separate those that conflate multiple distinct ideas.
   - Validate the coherence of each theme against the detailed context provided by the associated quotes and explanations.

4. **Theme Naming:**
   - Assign concise, evocative names to each theme that accurately reflect the central ideas of the grouped codes.

Output Format and Constraints:
   - Provide your final output strictly in valid JSON format without any additional commentary.
   - The JSON structure should be as follows:

   ```json
   {
     "themes": [
       {
         "theme": "Theme Name",
         "codes": ["Code1", "Code2", "Code3"]
       }
       // Additional theme objects...
     ]
   }
   ```

   No additional text outside the JSON.

Continuation Prompt:

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis. You have already generated the
following themes from earlier batches:

'''json
[Already generated theme names]
'''


Now integrate this new batch of data:

1. **List of Unique Codes:**

'''json
[List of unique Reviewed Codes in batch]
'''

2. **EC Data (JSON):**

'''json
[Table of Codes and their summarized Explanations]
'''

Follow the same analytic process:

* Identify patterns and shared meanings among codes.
* Integrate each new code into an existing theme or create new themes
as needed.
* Merge overlapping themes or split those conflating distinct ideas.
* Validate coherence of each theme against the QEC data.
* Assign concise, evocative names to any new themes.

**Output Format**
Return **only** this JSON object, with no extra commentary:

'''json
{
  "themes": [
```

```
    {
      "theme": "Theme Name 1",
      "codes": ["CodeA", "CodeB"]
    },
    {
      "theme": "Theme Name 2",
      "codes": ["CodeC"]
    }
    // ...
  ]
}
```

Redo with Feedback

Beginning Prompt:

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis. Previously, you generated themes
 based on a provided list of unique codes and a EC (Explanation-Code)
data. Now, you are tasked with refining those themes by incorporating
optional feedback and re-analyzing the original data to produce
improved higher-level themes. These themes should be directly linked
to the Main topic and research questions.


Context:
1. **Previous Themes:**
   The themes generated in the previous run, which the user did not
like and wants to be changed, provided in JSON format:

   [Last user updated themes to reviewed codes map]

2. **Feedback:**
   Optional feedback on the previous themes to guide refinement:

   [User Feedback]

3. **List of Unique Codes:**
```

A separate list containing all the unique codes extracted from the EC data.

Codes:
[List of unique Reviewed Codes in batch]

4. **EC Data (JSON):**
The Code and summary table contains the rationale summary (summary), and the assigned code.
- Example structure in JSON:

```json
[
   {
     "code": "CodeName",
     "summary": "Summary of multiple explanations signifying the meaning of the code."
   }
   // Additional instances...
 ]
```

Data:
[Table of Codes and their summarized Explanations]

Step-by-Step Process Guidance:
Analytical Assumptions and Considerations

- **Dual Processes of Quality Analysis:**
Good quality codes and themes result from both immersive, in-depth engagement with the dataset and from giving the developing analysis some reflective distance-often achieved by taking breaks during the process.

- **Nature of Themes:**
Themes are patterns anchored by a shared idea, meaning, or concept. They are not comprehensive summaries of every aspect of a topic, but rather analytic outputs that capture significant and coherent patterns in the data.

- **Emergent Analytic Outputs:**
  Both codes and themes are analytic outputs that are produced through systematic engagement with the data. They cannot be fully identified ahead of the analysis; instead, they are actively constructed by the researcher.

- **Active Production:**
  Themes do not passively 'emerge' from the data; they are the result of deliberate, reflective, and systematic analysis, combining both immediate engagement and thoughtful distance.

Instructions

1. **Review Previous Themes and Feedback:**
   - Examine the previous themes and the feedback provided.
   - Assess the strengths and weaknesses of the previous themes, identifying areas for improvement based on the feedback (e.g., missed patterns, unclear groupings, or misaligned interpretations).

2. **Familiarization with the Data:**
   - Re-examine the list of unique codes and the QEC data to refresh your understanding of the dataset.
   - Consider both the explicit (semantic) content and the underlying (latent) meanings conveyed by the quotes and explanations.

3. **Theme Refinement and Generation:**
   - Actively construct new or refined themes by identifying patterns and shared meanings among the codes, informed by the previous themes and feedback.
   - Modify existing themes, merge overlapping ones, separate themes that conflate distinct ideas, or create entirely new themes as needed to better capture significant and coherent patterns.
   - Validate each theme's coherence and relevance against the quotes and explanations in the QEC data.

4. **Theme Naming:**
   - Assign concise, evocative names to each theme that accurately reflect the central ideas of the grouped codes.

Output Format and Constraints:
   - Provide your final output strictly in valid JSON format without

````
any additional commentary.
   - The JSON structure should be as follows:

   ```json
   {
     "themes": [
       {
         "theme": "Theme Name",
         "codes": ["Code1", "Code2", "Code3"]
       }
       // Additional theme objects...
     ]
   }
   ```

No additional text outside the JSON.
````

Continuation Prompt:

````
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis. Previously you generated these
themes:

```json
[Last user updated themes to reviewed codes map]
```

User feedback:
[User Feedback]

So far, after integrating earlier batches, you have:

```json
[List of already generated themes]
```

Now refine and extend using this full dataset:

1. **List of Unique Codes:**

```json
````
````

```
[List of unique Reviewed Codes in batch]
```

2. **EC Data (JSON):**

```json
[Table of Codes and their summarized Explanations]
```

Your tasks:

* Incorporate the feedback into your existing themes.
* Integrate any codes not yet assigned, adding or modifying themes as needed.
* Merge, split, rename, or discard themes to improve coherence and distinctness.
* Validate each theme against the QEC data.

Output Format
Return **only** this JSON object, with no extra commentary:

```json
{
  "themes": [
    {
      "theme": "Refined Theme 1",
      "codes": ["CodeX", "CodeY"]
    },
    // ...
  ]
}
```

A.2 Refine Code Prompt

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis. Your task is to evaluate a
```

previously generated code and its corresponding quote in light of a user's comment from the chat history. You must determine whether you agree or disagree with the user's comment regarding the provided code and quote, provide a clear explanation for your stance, and output a command indicating the appropriate action. The available commands are:
- REMOVE_QUOTE: if the code/quote is deemed inappropriate or not representative.
- ACCEPT_QUOTE: if the code/quote is deemed appropriate and well-supported.
- EDIT_QUOTE: if you believe that the code/quote needs revision. In this case, provide a list of alternative code suggestions along with your explanation.
- REVERT_TO_INITIAL: if you believe that the initial code (before any modifications) is more appropriate than the current one, especially after considering the user's comment and chat history.

Input Information
- Transcript: [Transcript of the Reddit Post]
- Code: [Code]
- Quote: [Quote]
- Chat History: [Full History of interaction with user]
- User Comment: [Most recent User Comment]

Your Task
1. Analyze the provided transcript, code, quote, and chat history.
2. Determine whether you agree or disagree with the user's comment. Defend your position or, if the user's comment seems valid, proceed accordingly.
   Since codes can be subjective, think a lot before answering if your answer would make more sense or switching to the user's comment would be better.
3. Provide a concise explanation of your assessment and indicate if the code/quote requires revision.
4. Select the appropriate command:
   - Use REMOVE_QUOTE if the code/quote is inappropriate.
   - Use ACCEPT_QUOTE if the code/quote is appropriate.
   - Use EDIT_QUOTE if you believe the code/quote should be modified. In this case, include a list of alternative code suggestions.
   - Use REVERT_TO_INITIAL if you believe that the initial code is more suitable than the current one or the user specifically requests it to be reverted.

```
### Output Requirements
Return your output strictly in valid JSON format:

'''json
{
  "agreement": "AGREE" or "DISAGREE",
  "explanation": "Your explanation text here",
  "command": "REMOVE_QUOTE" or "ACCEPT_QUOTE" or "EDIT_QUOTE" or "
REVERT_TO_INITIAL",
  "alternate_codes": [ "alternative code suggestion 1", "alternative
code suggestion 2", ... ] // This field should contain a list of
revised code suggestions if command is EDIT_QUOTE; otherwise, it can
be an empty list.
}
'''

No additional commentary outside the JSON object.
```

A.3 Common Prompts

A.3.1 Summarization Prompts

```
Provide a concise summary (1-2 lines).
Respond *only* with valid JSON, for example:

'''json
{\"summary\": \"...\"}
'''

Texts:
[Texts to be summarized]
```

A.3.2 Topic Clustering Prompts

Beginning Prompt:

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis.

Your task is to review the list of initial codes and merge any that
share the same meaning into a single, more abstract code. Aim for a
balanced set of clusters-neither too many highly granular codes nor
too few overly broad ones. Use concise, descriptive names that capture
 the essence of each group.

Context:
Codes to cluster (JSON array of strings):
[Array of words]

Step-by-Step Process Guidance:
Generate each cluster code (reviewed code) as a natural phrase - just
as a expert human qualitative researcher would. Each word in the code
should be seperated by a space.

Output Format and Constraints:
Return **only** valid JSON (no extra text), wrapped in a markdown code
 block, in this format:

'''json
{
  "ClusterName1": ["codeA", "codeB", ...],
  "ClusterName2": ["codeC", "codeD", ...],
  ...
}
'''
```

Continuation Prompt:

```
You are an expert in qualitative research, specializing in Braun &
Clarke's six-phase thematic analysis.


Your task is to review the list of **more initial codes** and either
assign each one to an **existing cluster** if it fits that cluster's
meaning or create a new cluster.
```

- **Only** create a new cluster when a code does **not** clearly belong under any existing cluster name.
- Each new code must appear **exactly once** in your output.
- Generate each cluster code (reviewed code) as a natural phrase - just as a expert human qualitative researcher would. Each word in the code should be seperated by a space.

Context:
- **More initial codes (JSON array of strings):**
[Array of words]

- **Existing Clusters (JSON array of strings):**
[List of all Cluster heads currently generated]

Output Format and Constraints:
Return **only** valid JSON (no extra text), wrapped in a markdown code block, in this format:

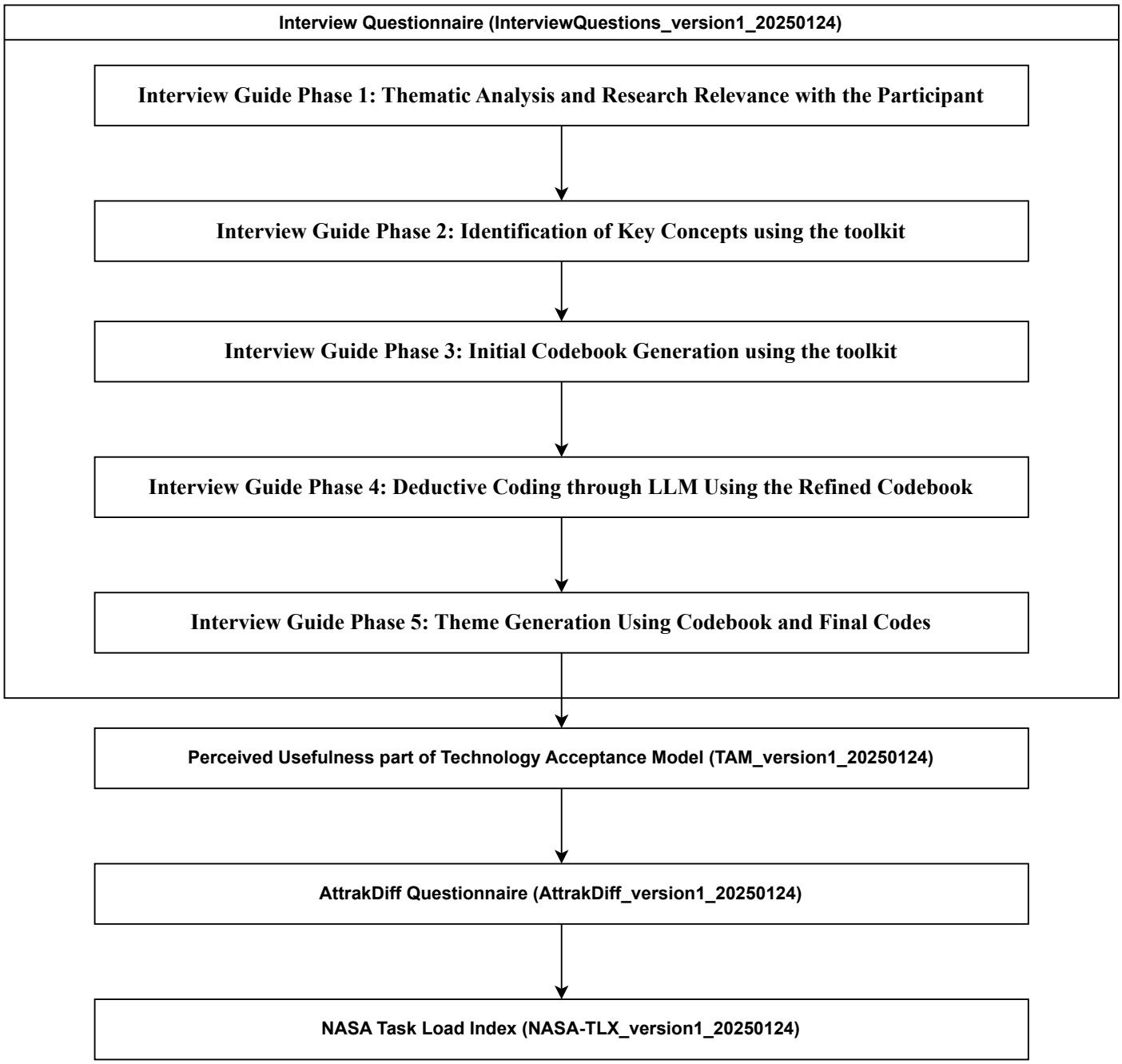
```json
{
  "ClusterName1": ["newCodeA", "newCodeB", ...],
  "ClusterName2": ["newCodeC", ...],
  "NewClusterName3": ["newCodeD", ...]
}
```

Appendix B

Interview Materials

Study Materials appendix contains

- Interview questions and survey flowchart shown to participants before the study.
- Interview questions used during the study.
- AttrakDiff questionnaire given to participants after the study.
- NASA-TLX questionnaire given to participants after the study.
- Perceived Usefulness Scale questionnaire given to participants after the study.
- Feedback letter shared with the participants after the study.


Interview Questionnaire (InterviewQuestions_version1_20250124)
Interview Guide Phase 1: Thematic Analysis and Research Relevance with the Participant
Interview Guide Phase 2: Identification of Key Concepts using the toolkit
Interview Guide Phase 3: Initial Codebook Generation using the toolkit
Interview Guide Phase 4: Deductive Coding through LLM Using the Refined Codebook
Interview Guide Phase 5: Theme Generation Using Codebook and Final Codes
Perceived Usefulness part of Technology Acceptance Model (TAM_version1_20250124)
AttrakDiff Questionnaire (AttrakDiff_version1_20250124)
NASA Task Load Index (NASA-TLX_version1_20250124)

Interview Guide Phase 1: Thematic Analysis and Research Relevance with the Participant

(This phase is conducted at the beginning of the study)

1. Introduction and Background:
    a. What is your main area of research and expertise?

    b. How long have you been involved in your field of research?

2. Relevance to Research:
    a. Where and how do you use Thematic Analysis in your research?

    b. Can you share an example of such a use case?

3. Dataset used for Thematic Analysis:
    a. What types of data do you typically work with when doing Thematic Analysis?

    b. What are the sources of these data – public/private/institutional?

4. Existing Methods used for Thematic Analysis:
    a. What methods or approaches do you currently use for Thematic Analysis?

    b. Could you share your experiences with different types of Thematic Analysis, specifically Qualitative Coding (e.g., inductive, deductive)?

    c. Do you use any specific software, tools, or libraries for Qualitative Coding?
        i. (If yes) what are they, and why do you prefer them?

5. Challenges faced in the Existing Tools used:
    a. What are the most significant challenges you face when using existing tools for Qualitative Coding?

    b. What features or improvements would you like to see in the tools or methods you use?

6. AI in Thematic Analysis:
    a. How do you envision technologies like AI, LLMs, and Machine Learning improving your workflow in Thematic Analysis?

Interview Guide Phase 2: Identification of Key Concepts using the toolkit.

(This phase is conducted after the user has used the toolkit to identify key concepts)

1. Experience of Identifying key concepts:
    a. What was your experience with the process of identifying and refining key concepts based on the literature and research questions you provided?

    b. Were there any parts of the process that required more attention?

2. Relevance of Key Concepts with User Expectations:
    a. How well did the LLM-identified key concepts align with your expectations and the research context? Were there any aspects you found particularly useful or unexpected?

3. Efficiency and Effort:
    a. What was your experience with the time and effort required to identify key concepts using this tool compared to your usual methods?

4. Confidence in the Approach:
    a. How do you feel about the key concepts identified with LLM assistance? In what ways did it support or challenge your existing approach?

5. Overall Impressions and Future Use of Key Concepts:
    a. What aspects of the LLM-assisted approach stood out to you, and how do you see it fitting into your future research work?

Interview Guide Phase 3: Initial Codebook Generation using the toolkit.

(This phase is conducted after the user has used the toolkit to develop an initial codebook)

1. Clarity and Interpretability of Initial Codes:
    a. How would you describe the clarity and ease of understanding of the initial codes? Were there any elements that stood out to you?

2. Initial Codes Relevance to Key Concepts and Data:
    a. How did the initial generated codes relate to the identified key concepts and ideas observed in the data? Were there any observations you made while reviewing the codes?

3. Refinement Process and Decision-Making in Initial Codebook Generation:
    a. How did you go about making decisions while working with the codes? Were there any parts of the process that required more of your attention?

    b. How would you describe your experience using the transcript view to provide feedback to the LLM?

    c. What are your impressions of the updated summary view, particularly in terms of understanding changes and feedback?

4. Codebook Alignment with User Expectations:
    a. How well did the final set of codes in the codebook align with your expectations? Were there any differences you observed from your initial thoughts?

5. Efficiency and Time Investment:
    a. How did the time required to refine and finalize the codes compare to what you expected? How manageable was the process for you?

6. Unexpected Findings or Observations:
    a. Were there any aspects of the process that you found surprising or different from what you anticipated? Can you describe your observations?

7. Overall Impressions and Future Use of Codebook Generation:
    a. How did you find this approach of refining codes through LLM by providing feedback? How do you see it fitting into your future research work?

Interview Guide Phase 4: Deductive Coding through LLM Using the Refined Codebook.

(This phase is conducted after the user has used the toolkit to review the final codes)

1. Relevance and Accuracy of Final Codes:
    a. How would you compare the LLM-coded data with your own coded data in terms of relevance to the research questions, key concepts, and the data?

    b. Can you describe any observations you made regarding the accuracy of the LLM-coded data?

2. Strengths and Weaknesses of the Deductive Coding Approach:
    a. What aspects of the deductive coding process using the toolkit stood out to you?

    b. How would you describe your experience with the LLM-assisted approach to reach final codes, including any aspects that required more attention?

3. Satisfaction and Improvements in the Deductive Coding Approach:
    a. What are your thoughts on the overall experience of using LLMs for deductive coding?

4. General Reflections on the Deductive Coding Approach:
    a. In what ways, if any, did the deductive coding approach influence your understanding of the data?

Interview Guide Phase 5: Theme Generation Using Codebook and Final Codes.

(This phase is conducted after the user has used the toolkit to review the final themes)

1. Clarity of Generated Themes:
    a. How would you describe your experience with understanding and interpreting the initial themes?

2. Relevance of Generated Themes:
    a. How do you perceive the grouping of codes within the initial themes?

    b. What are your thoughts on the relevance of the initial themes in relation to your topic of interest, the generated codes, and the data?

3. Evaluation and Future Use of Generated Themes:
    a. Can you share any observations about the initial themes? Were there any that stood out to you in any way?

    b. How do you see this approach fitting into your work?

UNIVERSITY OF WATERLOO

Block 1

Participant Identification

Default Question Block

Please provide your impressions of the product you have tested by check marking your impression on the scale between the terms offered in each line.

	1	2	3	4	5	6	7	
human	○	○	○	○	○	○	○	technical

isolating	○ ○ ○ ○ ○ ○ ○	connective
pleasant	○ ○ ○ ○ ○ ○ ○	unpleasant
inventive	○ ○ ○ ○ ○ ○ ○	conventional
simple	○ ○ ○ ○ ○ ○ ○	complicated
professional	○ ○ ○ ○ ○ ○ ○	unprofessional
ugly	○ ○ ○ ○ ○ ○ ○	attractive
practical	○ ○ ○ ○ ○ ○ ○	impractical
likeable	○ ○ ○ ○ ○ ○ ○	disagreeable
cumbersome	○ ○ ○ ○ ○ ○ ○	straightforward
stylish	○ ○ ○ ○ ○ ○ ○	tacky
predictable	○ ○ ○ ○ ○ ○ ○	unpredictable
cheap	○ ○ ○ ○ ○ ○ ○	premium
alienating	○ ○ ○ ○ ○ ○ ○	integrating
brings me closer to people	○ ○ ○ ○ ○ ○ ○	separates me from people
unpresentable	○ ○ ○ ○ ○ ○ ○	presentable
rejecting	○ ○ ○ ○ ○ ○ ○	inviting
unimaginative	○ ○ ○ ○ ○ ○ ○	creative

good	○ ○ ○ ○ ○ ○ ○	bad
confusing	○ ○ ○ ○ ○ ○ ○	clearly structured
repelling	○ ○ ○ ○ ○ ○ ○	appealing
bold	○ ○ ○ ○ ○ ○ ○	cautious
innovative	○ ○ ○ ○ ○ ○ ○	conservative
dull	○ ○ ○ ○ ○ ○ ○	captivating
undemanding	○ ○ ○ ○ ○ ○ ○	challenging
motivating	○ ○ ○ ○ ○ ○ ○	discouraging
novel	○ ○ ○ ○ ○ ○ ○	ordinary
unruly	○ ○ ○ ○ ○ ○ ○	manageable

Powered by Qualtrics

UNIVERSITY OF WATERLOO

Block 1

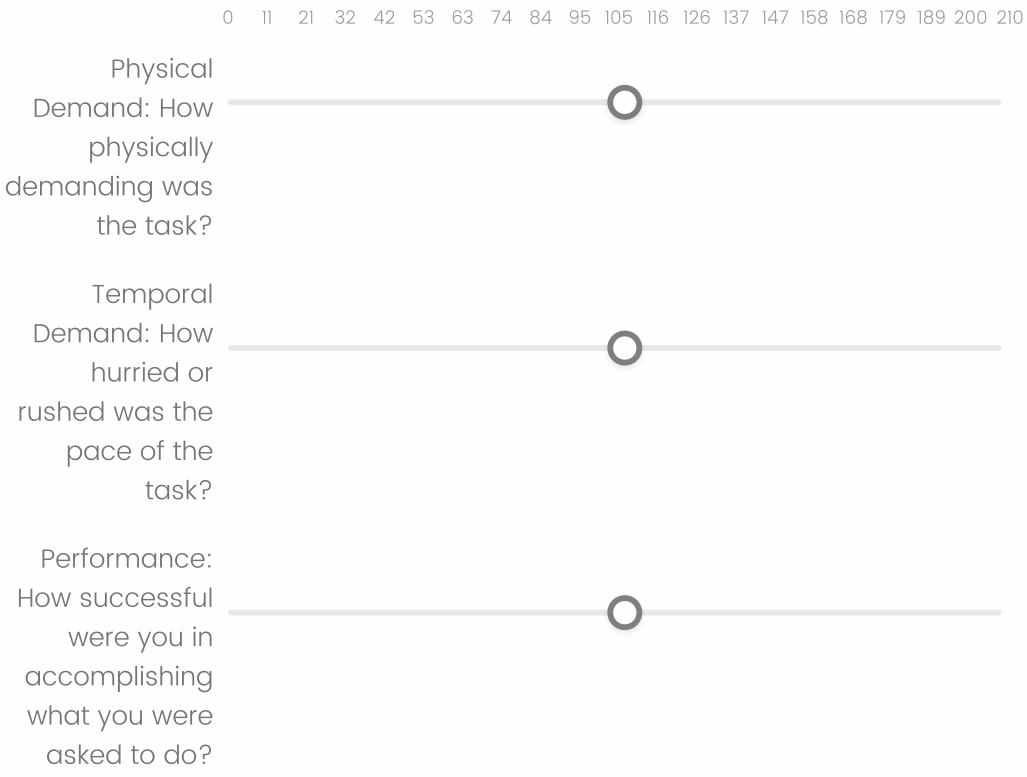
Participant Identification

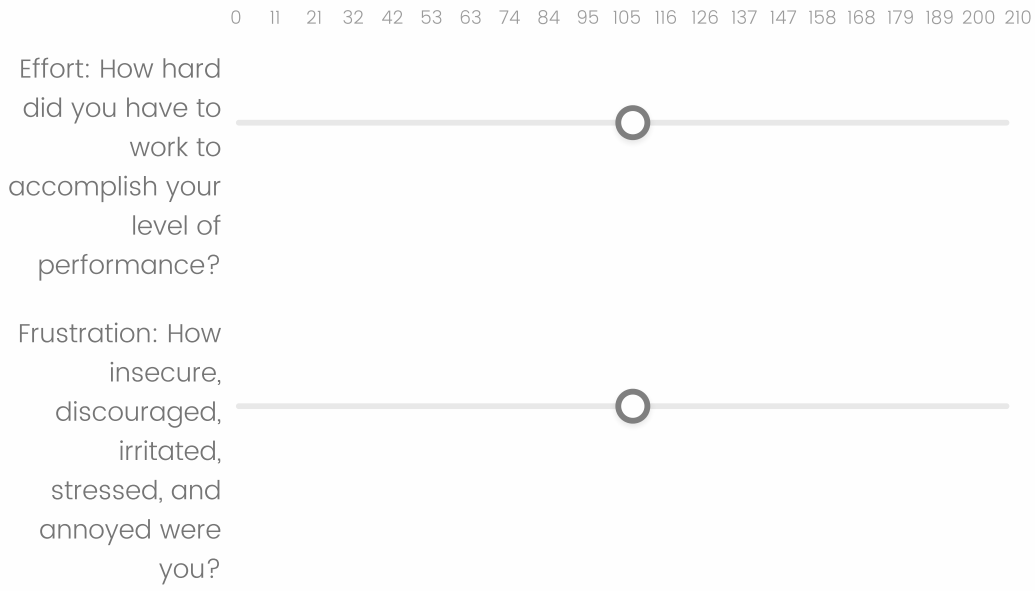
Default Question Block

Please use the sliders to answer each question below

0 11 21 32 42 53 63 74 84 95 105 116 126 137 147 158 168 179 189 200 210

Mental Demand: How mentally demanding was the task?


0 11 21 32 42 53 63 74 84 95 105 116 126 137 147 158 168 179 189 200 210
Physical Demand: How physically demanding was the task?
Temporal Demand: How hurried or rushed was the pace of the task?
Performance: How successful were you in accomplishing what you were asked to do?



Powered by Qualtrics

UNIVERSITY OF WATERLOO

Participant Identification

Perceived Usefulness (PU)

	Not at all useful	Slightly useful	Moderately useful	Very useful	Extremely useful
Using the system improves my performance in my job.	○	○	○	○	○
Using the system increases my productivity.	○	○	○	○	○

	Not at all useful	Slightly useful	Moderately useful	Very useful	Extremely useful
Using the system enhances my effectiveness in my job.	○	○	○	○	○
Using the system makes it easier to do my job.	○	○	○	○	○
I find the system useful in my job.	○	○	○	○	○
Using the system enables me to accomplish tasks more quickly.	○	○	○	○	○

Powered by Qualtrics

Ethics Material (Feedback Letter)

Dear ***(Insert Name of Participant)***,

We would like to thank you for your participation in this study. As a reminder, the purpose of this study is to investigate Human-LLM Collaboration for Deductive Thematic Analysis through Computer-Assisted Qualitative Data Analysis Software (CAQDAS).

The results of the study will provide a better understanding of the dynamics of Large Language Models (LLMs) in deductive thematic analysis. This includes, but is not limited to, evaluating the effectiveness of the proposed approach and toolkit for deductive thematic analysis (identification of key concepts, initial codebook generation, and final code and theme generation), as well as identifying any related challenges.

Please remember that your identity will be kept confidential. The data collected will not include any personal identifiers. If you are interested in receiving more information regarding the results of this study, or if you have any questions or concerns, please contact us at either the phone number or email address listed at the bottom of the page.

Should you choose to contact the researcher for any questions, please email Ansh Sharma, at a646shar@uwaterloo.ca

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB #47221). If you have questions for the Board, contact the Office of Research Ethics, toll-free at 1-833-643-2379 (Canada and USA), 1-519-888-4440, or reb@uwaterloo.ca. If you have any questions pertaining to the study specifically, please contact the researchers.

Thank you,

Jim Wallace
james.wallace@uwaterloo.ca
Associate Professor
School of Public Health Sciences
University of Waterloo, Canada

Ansh Sharma
a646shar@uwaterloo.ca
Student Investigator
David R. Cheriton School of Computer science
University of Waterloo, Canada

Appendix C

Human Ethics Protocol

Human Ethics Protocol appendix contains

- ORE approval from university of Waterloo's ethics board, File #: 47221.
- Protocols Ethics submission file.
- Information letter sent for participant recruitment.
- Poster used for participant recruitment.
- Consent letter form signed by participants before the study.

UNIVERSITY OF WATERLOO

Notification of Ethics Clearance to Conduct Research with Human Participants

---

Principal Investigator: James Wallace (School of Public Health Sciences)

Student investigator: Ansh Sharma (School of Computer Science)

File #: 47221

Title: Investigating Human-LLM Collaboration for Deductive Thematic Analysis through Computer-Assisted Qualitative Data Analysis Software (CAQDAS)

The Human Research Ethics Board is pleased to inform you this study has been reviewed and given ethics clearance.

Initial Approval Date: 02/28/25 (m/d/y)

University of Waterloo Research Ethics Boards are composed in accordance with, and carry out their functions and operate in a manner consistent with, the institution's guidelines for research with human participants, the Tri-Council Policy Statement for the Ethical Conduct for Research Involving Humans (TCPS2 2022), the Ontario Personal Health Information Protection Act (PHIPA), and all laws and regulations of the province of Ontario (as applicable). Additionally, CREB operates in a manner consistent with the International Conference for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) Guidance E6(R2): Good Clinical Practice , the International Organization for Standardization of Good Clinical Practices (GCP) as set out by ISO 14155 - Clinical investigation of medical devices for human subjects, Part C, Division 5 of the Food and Drug Regulations, Part 4 of the Natural Health Products Regulations, Part 3 of the Medical Devices Regulations. Both Boards are registered with the U.S. Department of Health and Human Services under the Federal Wide Assurance, FWA00021410, and IRB registration number IRB00002419 (HREB) and IRB00007409 (CREB).

Expiry Date: 03/01/26 (m/d/y)

Multi-year research must be renewed at least once every 12 months unless a more frequent review has otherwise been specified. Studies will only be renewed if the renewal report is received and approved before the expiry date. Failure to submit renewal reports will result in the investigators being notified ethics clearance has been suspended and Research Finance being notified the ethics clearance is no longer valid.

Level of review: Delegated Review

Signed on behalf of the Human Research Ethics Board



Heather Root, Senior Manager, Ethics, heather.root@uwaterloo.ca, 519-888-4567, ext. 41469

This above-named study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Documents reviewed and received ethics clearance for use in the study and/or received for information:

file: InformationLetter_version1_20250228.pdf

file: ParticipantConsentForm_version1_20250124.pdf

file: FeedbackLetter_version1_20250124.pdf

file: AttrakDiff_version1_20250124.pdf

file: NASA-TLX_version1_20250124.pdf

file: InterviewQuestionnaireAndSurveyFlowchart_version1_20250124.pdf

file: InterviewQuestions_version1_20250124.pdf

file: TAM_version1_20250124.pdf

Approved Protocol Version 2 in Research Ethics System

This is an official document. Retain for your files.

You are responsible for obtaining any additional institutional approvals that might be required to complete this study.

PROTOCOLS

#47221 - Investigating Human-LLM Collaboration for Deductive Thematic Analysis through Computer-Assisted Qualitative Data Analysis Software (CAQDAS)

Protocol Information

Review Type	Status	Approval Date	Renewal Date
Expedited	**Approved**	**Apr 01, 2025**	**Feb 06, 2026**

Expiration Date	Initial Approval Date	Initial Review Type
Mar 01, 2026	**Feb 28, 2025**	**Expedited**

Feedback

Approval Comment

The amendment request has been reviewed and received full ethics clearance on 04/01/2025. Please direct any further questions/concerns to Erin Van Der Meulen at ervandermeulen@uwaterloo.ca

Protocol Amendment Form

Amendment

Describe the proposed changes and why these changes are being made. Refer to the guideline for modifications to find review tips for preparing an amendment request. Review the full application and ensure that you:
1) Revise/edit all sections that relate to the amendment.
2) Upload all revised or new study materials to the appropriate section. Use track changes or highlight within the document to show where changes have been made.

Added a conference poster to recruitment.

General Information

Only the Principal Investigator/Faculty Supervisor can submit the application. This acts as a signature indicating approval of the application.

Principal Investigator / Faculty Supervisor

James Wallace

Department

School of Public Health Sciences

Study title

Investigating Human-LLM Collaboration for Deductive Thematic Analysis through Computer-Assisted Qualitative Data Analysis Software (CAQDAS)

General Questionnaire

Indicate the type of application you would like to complete

Standard application

Please confirm:

I understand that the type of applications listed above determine the form I am about to complete. If I have chosen the incorrect form I acknowledge that I may need to complete a new application.

People

Research team members based at the University of Waterloo.

Ensure **all** information in this table is completed.

Person

James Wallace

Waterloo Department

School of Public Health Sciences

Email Address

jrwallac@uwaterloo.ca

Phone

Researcher Role

Principal Investigator

Permissions

Full Access

Mandatory Training

All University of Waterloo undergraduate and graduate students, faculty, and staff must complete the TCPS 2 2022 CORE tutorial prior to submitting an application for review. See instructions or email researchethics@uwaterloo.ca if you have questions.

REQUIRED: Upload a copy of the TCPS 2 2022 certificate or a screen shot showing module completion. Applications where the certificate/screen shot is not uploaded will be sent back to the researcher and not reviewed.

TCPS2 - Wallace - 2023.pdf

Ensure **all** information in this table is completed.

Person

Ansh Sharma

Waterloo Department

School of Computer Science

Email Address

a646shar@uwaterloo.ca

Phone

2266980121

Researcher Role

Student investigator

Permissions

Full Access

Mandatory Training

All University of Waterloo undergraduate and graduate students, faculty, and staff must complete the TCPS 2 2022 CORE tutorial prior to submitting an application for review. See instructions or email researchethics@uwaterloo.ca if you have questions.

REQUIRED: Upload a copy of the TCPS 2 2022 certificate or a screen shot showing module completion. Applications where the certificate/screen shot is not uploaded will be sent back to the researcher and not reviewed.

tcps2_core_certificate.pdf

Do you have any investigators external to the University of Waterloo?

No

General details

Is this new study related to any previous application?

No

What is the estimated start and end date for the study?

Start Date (Date you anticipate beginning the recruitment of participants)

1 February 2025

End Date (Date you expect the study will end, for example, the date when there will be no further contact with participants or when the data analysis will be completed for US funded research)

31 May 2025

Does this research require approval from a UWaterloo departmental committee?

Not a department requirement

What is the level of the research to be conducted? Choose one.

Master's thesis

Will this study involve local hospitals covered by the Waterloo Wellington Research Ethics Board (WWREB) (established by Cambridge Memorial Hospital, Grand River Hospital and St. Mary's General Hospital)?

No

Has a version of this study been disapproved or rejected by any Research Ethics Board/Committee?

No

Study description

State your research question(s).

1. How effectively can LLMs assist qualitative researchers in identifying key concepts by analyzing prior literature and research questions for deductive thematic analysis? 2. How rigorously can LLM-assisted key concepts support final code and theme generation in deductive thematic analysis? 3. How do qualitative researchers perceive the usability and efficiency of an LLM-assisted deductive thematic analysis toolkit in enhancing their research workflows?

Provide a clear, detailed description of the purpose, hypothesis, aim, and objectives of this study.

This research aims to comprehensively explore the potential of Large Language Models (LLMs) in assisting qualitative researchers with deductive thematic analysis. Specifically, the study investigates how LLM-assisted key concepts and initial codebooks can support the generation of final codes and themes in qualitative research. The study also seeks to understand the usability and efficiency of an LLM-assisted deductive thematic analysis toolkit by engaging qualitative researchers through a structured user study. Participants will provide insights into their existing thematic analysis workflows, the challenges faced with current tools, and their experiences using our proposed LLM-assisted approach. The study will follow a multi-phase approach, beginning with familiarizing participants with the toolkit, followed by key concept identification using LLMs based on participant-selected research literature and research questions. The LLM will iteratively refine these key concepts using participant feedback. Subsequently, the LLM will generate an initial codebook based on the refined key concepts. Participants will iteratively refine the codebook through interactions with the LLM, allowing them to provide feedback, edit, delete, and validate the generated codes. Once finalized, the initial codebook and the refined key concepts will be used to perform deductive coding on the remaining dataset, with a comparative analysis between participant-generated and LLM-generated codes to assess accuracy, consistency, and inter-annotator agreement. We hypothesize that our LLM-assisted toolkit will enhance the rigor and efficiency of the deductive thematic analysis process by leveraging vast textual data, offering researchers a more structured and accessible approach to deductive thematic coding. The deployment of the LLM-based toolkit as an interactive and adaptive user interface will empower researchers to explore AI-assisted methods while maintaining control over the coding process. By comparing LLM-assisted and traditional coding methods, the study aims to highlight the strengths, limitations, and potential improvements of integrating AI into qualitative research workflows.

Provide background information, a rationale, and justification for conducting this study. Describe why the research is being done and what research has already been done in this area. Be sure to explain why this research is important.

Deductive thematic analysis is an important approach in qualitative research that enables researchers to systematically apply existing key concepts to analyze textual data. However, traditional qualitative coding methods, such as manual coding and rule-based approaches, present several challenges, including scalability issues, time consumption, and subjectivity. These challenges make it difficult for researchers to efficiently analyze large and complex datasets. Large Language Models (LLMs), with their advanced natural language processing capabilities, offer significant potential to assist in this process by providing contextual understanding, automating repetitive tasks, and supporting researchers in developing and refining key concepts for deductive coding. Leveraging LLMs in deductive thematic analysis can enhance the efficiency and rigor of the process while allowing researchers to focus on interpreting meaningful insights rather than being burdened by manual coding efforts. This study aims to explore the potential of deploying LLMs for deductive thematic analysis, specifically focusing on their ability to assist researchers in identifying key concepts, generating initial codebooks, and refining final codes and themes through iterative feedback loops. Existing studies, such as "LLM-in-the-loop: Leveraging Large Language Models for Thematic Analysis" and "Combining Codebook with GPT-3 for Deductive Coding," have demonstrated the feasibility of using LLMs in qualitative analysis. However, these approaches are not generalized for varying datasets, often lack seamless integration into researchers' workflows, and require significant manual intervention. Our proposed solution addresses these limitations by developing an interactive LLM-assisted toolkit that not only automates parts of the analysis but also enables researchers to iteratively refine outputs based on their domain knowledge. This ensures a balanced collaboration between human expertise and AI capabilities, with the LLM acting as an assistant rather than a replacement for the researcher. To evaluate the effectiveness, efficiency, and usability of the proposed toolkit and underlying approach, the study will involve a structured user study with qualitative researchers. Participants will use the toolkit to identify key concepts, refine initial codebooks, and apply them to analyze textual data from Reddit to generate codes and themes. The study will assess the LLM's performance using key metrics such as accuracy, inter-annotator agreement (IAA), cosine similarity, weighted precision, and weighted recall. Additionally, participant feedback on the overall usability of the toolkit will

be gathered through semi-structured interviews and established usability evaluation methods, including the Perceived Usefulness part of Technology Acceptance Model (TAM), the AttrakDiff Questionnaire, and the NASA Task Load Index (NASA-TLX). The findings from this study will provide valuable insights into the potential of LLM-assisted deductive thematic analysis to improve qualitative research workflows. Furthermore, this research aims to contribute to the broader discourse on AI integration in qualitative research by highlighting the strengths, limitations, and potential for future improvements, ultimately making the process more accessible, structured, and efficient.

In a maximum of 250 words, provide a non-scientific lay language description that summarizes the project outlining the purpose, anticipated benefits, and basic procedures. Write this summary as if it would be read by members of the general public who are not familiar with academic terms or acronyms. Use language suitable for a media release.

This project explores how artificial intelligence (AI) can assist researchers in analyzing large volumes of text more efficiently and accurately. Traditionally, qualitative researchers manually review and organize text to identify key themes and patterns - a process that can be slow, labor-intensive, and subjective. Our study aims to introduce an AI-powered tool designed to support researchers by helping them develop a structured workflow, generate themes, and refine findings, ultimately making the analysis process faster and more reliable. The purpose of this research is to evaluate how effectively this AI tool, powered by Large Language Models, can assist researchers in analyzing text from sources such as social media posts. Study participants, who are experienced researchers with little to no technical expertise, will use the tool to analyze a set of online posts and provide feedback on its usefulness, ease of use, and overall effectiveness. They will compare AI-generated results with their own analysis to determine whether the tool can enhance their workflow without replacing their expertise. By combining human judgment with AI capabilities, this tool has the potential to save time, reduce workload, and provide more consistent results across large volumes of text in qualitative research. The study findings will aid in developing more tools that help researchers adopt AI for similar tasks without requiring technical skills. Ultimately, this project aims to create a user-friendly solution for text analysis, enabling researchers from various fields to focus more on drawing meaningful insights rather than spending excessive time on manual analysis.

What is the study design?

Mixed-method, user study

Is this a pilot study?

No

Sample Size

What is the expected sample size? Outline the number of participants anticipated to take part in the study.

12

Was a formal sample size calculation completed?
See Statistics Help Desk for support in calculating sample sizes.

No

Provide a rationale for the number of participants specified

The study is conducted with domain experts. The focus would be to qualitatively analyze the experience of a small population (qualitative researchers in various domains) while performing the task of qualitative coding. Thus, the sample size has to be limited and specific.

Study sites

Where is this study taking place?

Remote (online survey, virtual/telephone interview, etc.)
On a University of Waterloo campus

Please note that different guidelines/policies may apply when participants are recruited from certain locations.

Are there any permissions required to conduct this study on campus?

No

Funding

Is the study funded/will it be funded?

Yes

Funding
List all funding sources that are new or ongoing

Funding status

Ongoing funding

Funding source is

Tri-agency / Canadian Government sponsor

Industry partner/sponsor name

Canadian Government agency

NSERC - Natural Sciences and Engineering Research Council of Canada

Charitable foundation

Other Charitable foundation name

Internal UWaterloo funding source

US federal funds

Other US federal funds sponsor

Other funding source

Program name if applicable

Discovery

Work-order or award number, if known

2022-03268

What is the expected period of funding

What is the period of funding

Funding from

1 May 2022

Funding to

30 April 2026

Conflict of interest

Are there any potential, perceived, or actual financial or non-financial conflicts of interest of the research team in undertaking the proposed research?

No

Benefits

Are there direct benefits of the proposed research to the study participants?

No

What are the scientific and/or scholarly benefits of the proposed research?

Our research enhances the accessibility and usability of Large Language Models (LLMs) for qualitative researchers in deductive thematic analysis. By developing an interactive LLM-assisted toolkit, the study offers a structured approach to qualitative data analysis, improving efficiency and reducing manual effort. This enables researchers, regardless of their technical expertise, to streamline the identification of key concepts, codebook generation, and theme identification, simplifying the overall process of deductive thematic analysis. The study also provides insights into optimizing LLMs for qualitative workflows by incorporating participant feedback and usability evaluations. It contributes to a deeper understanding of how AI, specifically LLMs, can complement human expertise while ensuring alignment with key concepts. Furthermore, the study evaluates key performance metrics, including inter-annotator agreement, precision, recall, and usability measures such as the Perceived Usefulness part of Technology Acceptance Model (TAM), and NASA Task Load Index (NASA-TLX) to assess the usability of the proposed toolkit for researchers from varied backgrounds. These findings will help refine AI-assisted qualitative research methods, promoting AI as an accessible, efficient, and reliable tool across disciplines.

Participants

Participant general categories

University of Waterloo undergraduate and/or graduate students
University of Waterloo staff and/or faculty
Adults (age 18-64 years)

Are you conducting research in classes with students as your participants to evaluate a teaching method or object?

No

Describe the sample in detail and list any specific inclusion/exclusion criteria for the study

The primary goal of this study is to gain a deep understanding of the experiences and needs of qualitative researchers from various domains in relation to the task of deductive thematic analysis. Qualitative researchers may have little to no experience with coding languages but will have experience in conducting and analyzing qualitative data. They must speak English and have experience in conducting deductive thematic analysis. Consequently, the inclusion criteria emphasize that researchers should utilize qualitative data and be capable of providing feedback on the effectiveness of our proposed approach within the context of their work, as addressed in our interview questions.

If you are excluding people on certain characteristics provide a justification for the exclusion.

Only participants speaking English will be included, as a translator is not considered for this project.

Will a screening process be used to determine eligibility in the study based on the inclusion and/or exclusion criteria identified above?

No

Recruitment

Identify from where/what sources potential participants will be recruited. If your participants are part of an organization/group please see guidance for conducting research when participants are part of an organization or group and incorporate this information into the methods section.

University of Waterloo campus (e.g., SLC, Engineering Foyer, laboratories)
Through email/internet (e.g., social media networks)

Where on campus do you intend to recruit?

University of Waterloo Public health

Indicate what email listing, internet site or network you intend to recruit from

University of Waterloo Computer Science graduate student mail list: scs-grads@lists.uwaterloo.ca

What recruitment materials will be used? See sample recruitment materials.

Email script

Posters

Upload your recruitment materials

Upload your recruitment materials

InformationLetter_version1_20250228.pdf

Study group

All participants

Upload your recruitment materials

Poster_version1_20250327.pdf

Study group

All Participants

Will potential participants be recruited through pre-existing relationships with members of the research team (e.g., employees, students, or patients of research team, acquaintances, own children or family members, colleagues, etc.)?

Yes

Outline the relationship between the researchers and potential participants (e.g., professor-student, colleagues)

Colleagues: Some participants may be colleagues of the investigators who conduct research within the faculty. Professor-Student: Some participants may be professors of the investigators, as the research is being conducted within our own faculty.

Could this relationship compromise the potential participant's freedom to decline participation?

No

Explain

The researchers will email CS graduates to invite them to participate through the graduate mailing list (scs-grads@lists.uwaterloo.ca). Additionally, the researchers will reach out to individuals (professors/colleagues) at the University of Waterloo's School of Public Health Sciences who meet the specified inclusion criteria. These potential participants, comprising colleagues and professors, will be approached based on their proficiency in qualitative research and their interest in applying AI to qualitative coding tasks. The participants will not include any current students of the Principal Investigator. Importantly, participants retain the autonomy to decline involvement without facing any personal or professional repercussions. Furthermore, participants are free to withdraw from the study at any time without any consequences or judgments imposed by the investigators.

Methods and procedures

Which of the following will be conducted for this study?

Surveys/questionnaires
One-on-one interviews

How will the survey(s) or questionnaire(s) be administered?

In-person - Computer-administered or web/online
Online or web

Provide the URL of the survey if available. Be sure to also upload a copy of the survey/questionnaire in the Study Material section below.

If you are using Waterloo's Qualtrics, refer to guidance from IST about how to use the various features in your online data collection.

Perceived Usefulness part of Technology Acceptance Model (TAM): https://uwaterloo.ca1.qualtrics.com/jfe/form/SV_0dGmv6VLpAfU7Vs , AttrakDiff Questionnaire: https://uwaterloo.ca1.qualtrics.com/jfe/form/SV_2leYsJWaVKmitfg , NASA Task Load Index (NASA-TLX): https://uwaterloo.ca1.qualtrics.com/jfe/form/SV_1Y26N47lyM1EUrY

How will the one-on-one interviews be conducted?

In-person
Online – Video Chat

Will quotations be used in the write-up of the study

Yes

What type of quotations will be used?

Anonymous

For each of the procedures indicated above, provided a detailed, sequential description of how they will be used in the study. Provide one or more paragraphs describing how you are recruiting participants, obtaining consent, what participants are asked to do, and how the research team will be using the collected data.

A list of potential participants will be curated based on qualitative researchers with experience in Deductive Thematic Analysis at the University of Waterloo from the Computer Science and Public Health departments. Recruitment will occur through email lists (CS graduate mailing lists), poster presentation at the Research and Practice in Health Sciences Conference 2025 (using poster uploaded in the recruitment section of this application- Poster_version1_20250327.pdf), and individual contact via email (Public Health) between the student investigator and potential participants. The interview phase will commence once participants are identified and have provided their consent. The student investigator will conduct semi-structured interviews with the researchers after each phase of the study, using the attached questionnaire (InterviewQuestions_version1_20250124.pdf). Depending on the participants' convenience, interviews may be conducted on campus or via video chat. The interview questions will explore the participants' current methods and their perspectives on the proposed approach, with the gathered insights being documented. The participants will be involved in multiple semi-structured interviews in five phases. The first phase interview questionnaire will cover general topics related to thematic analysis and qualitative coding, focusing on the participants' existing knowledge and use of these methods for their tasks. From the second phase interview onwards, the focus will shift to the participants' experiences with the proposed approach and toolkit. At this stage, participants will be asked to upload their research questions and relevant literature to utilize the toolkit for identifying key concepts for deductive coding tasks. The second phase interview will address their experience with key concept identification using the toolkit. Based on the key concepts, the LLM will then generate initial codes from a portion of the data. Participants will have the opportunity to refine these initial codes by selecting, editing, and deleting them or requesting the LLM to regenerate codes, resulting in the creation of an initial codebook. The third phase interview will focus on Initial Codebook Generation. Following this, the LLM will perform a deductive coding task on the remaining data while the participant simultaneously conducts deductive coding on the same dataset. This step aims to compare the time taken by qualitative researchers to perform a typical deductive coding task using the toolkit and to calculate the Inter-Annotator Agreement (IAA) between the

participant-coded data and the LLM-coded data. Once the participant completes the deductive coding task, they will be asked to review the LLM-coded data and make any necessary modifications. The fourth phase interview will focus on Deductive Coding using the Refined Codebook. After the review process, the LLM will generate themes based on the overall coded data from both the codebook development phase and the deductive coding phase. Participants will then review these themes to finalize them. This process will result in the final set of themes and coded data. The fifth phase interview will focus on Theme Generation using the Codebook and Final Codes. Following these interviews, participants will be asked to complete three surveys sequentially: Perceived Usefulness part of the Technology Acceptance Model (TAM_version1_20250124.pdf) to evaluate the likelihood of users adopting the toolkit based on perceived usefulness; the AttrakDiff Questionnaire (AttrakDiff_version1_20250124.pdf) to assess both pragmatic (functionality) and hedonic (user experience) aspects of the toolkit; and the NASA Task Load Index (NASA-TLX_version1_20250124.pdf) to measure the efficiency and cognitive load of the toolkit on participants. The flow of the study interview questionnaire and survey can be seen in the InterviewQuestionnaireAndSurveyFlowchart_version1_20250124.pdf. At the conclusion of the study sessions, feedback letters expressing gratitude for the participants' time and insights will be distributed. The collected results will undergo analysis to iteratively improve the proposed solution, fostering continuous enhancement based on the feedback received from the qualitative researchers. Please Note: This toolkit is a desktop application (available for all operating systems) that participants can will be sent a copu. It will utilize Gemini 2.0 Flash (an LLM developed by Google) through the paid quota of the Gemini API (Google's AI service). The data input by participants will only be stored locally within the application on their desktop. The paid quota of the Gemini API explicitly states that Google will not use this data (including all data uploaded by participants in this study) for training its AI or improving any other Google service or product (https://ai.google.dev/gemini-api/terms#data-use-paid).

Please upload any study materials related to the procedure(s)

Study material

AttrakDiff_version1_20250124.pdf

Study material

NASA-TLX_version1_20250124.pdf

Study material

InterviewQuestionnaireAndSurveyFlowchart_version1_20250124.pdf

Study material

InterviewQuestions_version1_20250124.pdf

Study material

TAM_version1_20250124.pdf

Does the study involve the administration or use of an approved drug or natural health product?

No

Will you be collecting any biological specimens?

No

Will you be creating or contributing to a bio-bank, bio-repository, registry, as part of the study?

No

Will you be doing any genetic testing or analysis?

No

Incidental and secondary findings

See Guideline for reporting incidental and secondary findings to study participants

Are any of the methods or procedures used likely (i.e., a real possibility and probability) to reveal an incidental finding (i.e., discoveries made in the course of research but that are outside the scope of the research and/or results that are outside the original purpose for which a test or procedure was conducted)?

No

Are any of the methods or procedures used likely to reveal a secondary finding (i.e., findings that are not the primary target of the test or procedure; rather, it is an additional result that is actively sought)?

No

Equipment use

Will there be any equipment used as part of this study?

No

Deception

Does the study involve deception or partial disclosure?

No

Risks and safeguards

Considering each method or procedure to be used in this study, indicate if participants might experience any of the following risks or harms

No known or anticipated risks

Outline the criteria for stopping the study early due to safety concerns/other issues.

None

Privacy

Will demographic and/or background information be asked of participants? If so, ensure that the demographic questions have been uploaded in the methods section.
See TCPS2, Chapter 5, Privacy and Confidentiality, and Waterloo's guideline on securing human research participant information and data.

No

Participant identification - If applicable, include how participants will be referenced in study results.

By a participant number (eg. P1 for Participant 1).

If applicable how will the key/list that links participants' codes with their actual name and/or consent forms be stored and protected? Also, outline how long the key/list will be stored.

The links to the participants' codes along with their actual names will be stored on a password-protected computer. Additionally, any confidential information will be encrypted to provide an added layer of security. The data will be retained for the duration of the required study and deleted once the study is concluded.

Are there any limitations to the promise of confidentiality?

No

Will any study data be leaving the University of Waterloo, the province, or country (e.g., member of research team is located in another institution, province, or country, etc.)?
See Waterloo's guideline on securing human research participant information and data.

No

Will any collected data or information be entered into a database for future use?

No

Are there other members of the research team who are not named on this application (e.g., co-op students, research assistants, or other temporary personnel) who may carry out specific tasks involved in your study?

No

Will individual participant identities be confidential in the publication or release of the study findings?

Yes

Data storage

What type(s) of data will be collected for this study?

Hard copy / Paper records
Audio-recordings
Electronic files

For each type of information collected, identify where the data will be stored. Specify the physical location (e.g., personal computer, UW server, server in US, a Cloud, a Google Drop box).
Describe how it will be secured to protect privacy and confidentiality (e.g., encrypted).
* Best practices are to de-identify the data at the earliest possible time.
See Waterloo's guideline on securing human research participant information and data.

The video associated with this interview will be deleted immediately, while the audio will be used for transcription and deleted thereafter. The electronic data (audio transcripts from in-person and video interviews) will be stored on a password-protected computer system. The paper records (onsent forms signed by participants if they choose to participate in person) will be stored in a locked drawer in room LHN 1706 after the records have been de-identified.

For each type of data collected, identify the minimum retention period.
See Guideline on data retention and how long researchers should keep their data.

Electronic files will be securely deleted and paper records shredded one year after their last use, or by the end of the term in which the thesis was submitted - whichever comes later - to accommodate any requests or appeals related to the work.

Data Management

Are there plans to link the data collected with other data sets, databases, or registries?

No

The Tri-Agency Open Access Policy on Publications and some journals are requesting that research data be provided to an open access repository to promote the availability of findings, to enhance transparency and share with the widest possible audience.

Do researchers plan to make the data-set available in an online repository/archive?

No

Do you have a data management plan? See the Library Research Data Management: Get Started guide.

Yes

Consent and Withdrawal

What member(s) of the research team will be responsible for obtaining informed consent?

Student investigator - Ansh Sharma

Is there a relationship between the potential participant(s) and the person obtaining consent?

Yes

Explain the nature of the relationship (e.g., professor/course instructor, teacher, supervisor, employer, health care professional/physician, etc.)

Colleagues, fellow graduate students, professors.

Describe how undue influence and/or authority to participate will be mitigated.

The participants will be recruited based on the inclusion criteria and their expertise. Professors and investigators will have the option to decline participation without any pressure or obligations from the study researchers. The researchers will explain to them that they can decline participation and withdraw from the study at any time without any consequences.

How will consent be obtained

Written signature on a consent form

Upload Information and Consent Materials - See resources and samples for creating information consent letters. Refer to the guide for creating an information consent letter on this webpage.

Upload Information and Consent Materials

ParticipantConsentForm_version1_20250124.pdf

Study group

All participants

Do you anticipate that you will need to make special accommodations for your participant group?

No

Do you anticipate needing to put in place any special procedures when obtaining informed consent?

No

Will consent need to be re-documented throughout the life of this study?

No

Describe the participants rights and processes for withdrawing consent.

The participants will be informed by e-mail of their right to withdraw from the study. During the interviews, they will again be informed of their right to withdraw at any time during the study. If they wish to withdraw their consent, they will contact the researchers via e-mail, stating their intention. A follow-up e-mail will be sent to the participants to confirm their withdrawal. This e-mail will serve as proof of confirmation that the participant has contacted the researchers regarding withdrawal, and the follow-up e-mail confirms that the participant is no longer part of the ongoing study.

Outline what will be done with the participant's information (data, samples, etc.) if they withdraw from the study.

The participant's data will be deleted.

Will any individuals taking part in this study be unable to provide their own informed consent?

No

Remuneration

Will there be remuneration provided to show appreciation for a participant's time, effort, skills, etc. to take part in the study?

No

Will participants incur any expenses by participating in the study?

No

Feedback and Appreciation

How will you show appreciation to participants for taking part in the study?

All participants in the study will receive a follow-up e-mail with a feedback letter thanking them for their participation and restating the goal of the study, anonymity, and the researchers' contact information. They will also receive a copy of the results when it becomes available.

When will feedback/appreciation be provided to participants (e.g., immediately after the session, at the end of a survey, mail results at time X.)?

After the successful completion of the study (interview and survey) session.

Upload Feedback/Appreciation materials

Upload Feedback/Appreciation materials

FeedbackLetter_version1_20250124.pdf

Study group

Participants in the study

How can participants learn about the study results/obtain a summary of the findings if interested?

The results of the study will be published as an MMath thesis on UWSpace and in at least one online publication.

Other Details

Provide any other information relevant to this study you wish to explain to the Research Ethics Board reviewers or to the staff in the Office of Research Ethics.

None

Other Attachments

Upload any additional study documents.

Attachments

Attestation

As the Principal Investigator/Faculty Supervisor/Local Investigator, I attest to the following:

- I will ensure all co-investigators, collaborators, and student investigators listed on this application have reviewed the application contents and will conduct the study according to the application/protocol.
- I am aware that any changes made to the research must be reviewed and provided clearance before the changes are implemented. Change requests (i.e., an amendment) are to be submitted through the system. I am also aware ethics clearance for this study is valid for only 12 months unless I renew the study prior to the ethics clearance expiry date. If an annual renewal report is NOT submitted through the system prior to the expiry date, the study will be suspended, all work on the study must stop, and Research Finance will be notified which will result in a hold being put on the funds associated with this study.
- I agree to comply with the Tri-Council Policy Statement (TCPS2) for conducting research with human participants and with University of Waterloo policies and guidelines when conducting this study (e.g., statement on human participant research, IST policies, etc.).
- I confirm I have read the University of Waterloo Research Integrity guidelines and I agree to comply with the policies and guidelines of my profession or discipline regarding the ethical conduct of research involving humans.
- Proper safeguards as to confidentiality and security of data will be maintained and data will be securely stored in accordance with University of Waterloo standards and applicable privacy legislation.

By submitting this application I agree to the above attestations and will ensure the research is conducted accordingly

Only the Principal Investigator/Faculty Supervisor can submit the application. This acts as a signature indicating approval of the application.

This is the end of the application form. Click submit in the right menu if you are ready to send it to the Research Ethics Office.

Ethics Material (Information Letter)

Interview and Survey

Information Letter – Invitation email

This email is being sent on behalf of the researchers:

Jim Wallace, University of Waterloo
Ansh Sharma, University of Waterloo

Subject: Participants needed for HCI study on Investigating Human-LLM Collaboration for Deductive Thematic Analysis through Computer-Assisted Qualitative Data Analysis Software (CAQDAS)

Dear (potential participant),

To help you make an informed decision regarding your participation, this letter will explain what the study is about, the possible risks and benefits, and your rights as a research participant. If you do not understand something in the letter, please ask one of the investigators before agreeing to participate in the study. You may wish to save a copy of this letter for your records.

This email is an invitation to participate in a research study we are conducting as part of my MMath degree at the David R. Cheriton School of Computer science and School of Public Health, University of Waterloo, under the supervision of Dr. Jim Wallace.

The goal of this study is to explore how a Large Language Model (LLM)-assisted toolkit can support researchers in deductive thematic analysis, including the identification of key concepts, generation of initial codebooks, refinement of codes, and ultimately, the production of final themes from textual data. Deductive thematic analysis is a valuable method for organizing and interpreting data based on existing theories. The process involves coding text according to predefined frameworks and systematically identifying relevant themes within the data. This study aims to leverage state-of-the-art LLMs to enhance our understanding of their usability as assistants in qualitative coding and deductive thematic analysis.

We are seeking participants who have experience in conducting deductive thematic analysis.

The participants' role in this research study will be to provide insights on how their research relates and utilizes deductive thematic analysis, their viewpoints of the proposed approach, and comments on its applicability in general and in their research.

Your participation will involve multiple semi-structured interviews in five phases in a single three hour session:

1. Interview phase 1 – Current Practices: You will share your existing approaches to qualitative research and experiences with thematic analysis and qualitative coding.
2. Interview phase 2 – Identification of Key Concepts: You will upload your research questions and relevant literature to use the toolkit to identify key concepts.
3. Interview phase 3 – Initial Codebook Generation: You will refine an initial set of LLM-generated codes for a portion of the data.
4. Interview phase 4 – Deductive Coding with the Refined Codebook: You and the LLM will both code the remaining data, enabling us to compare time efficiency and inter-annotator agreement.
5. Interview phase 5 – Theme Generation & Finalization: You will review LLM-generated themes based on the coded data and finalize them.

Please Note: Your data will not be used to train AI or any other Google service/product. For the study, we will use the paid quota of the Gemini API. Google does not use your prompts (including associated system instructions, cached content, and files such as images, videos, or documents) or responses to improve its products, and it will process your prompts and responses in accordance with the Data Processing Addendum for Products Where Google is a Data Processor (https://ai.google.dev/gemini-api/terms#data-use-paid).

After these interviews, we will ask you to complete three short surveys (Perceived Usefulness part of Technology Acceptance Model, AttrakDiff, and NASA-TLX) to evaluate the toolkit's usability, acceptance, user experience, and perceived workload.

For all surveys, you will be using Qualtrics that has implemented technical, administrative, and physical safeguards to protect the information from loss, misuse, and unauthorized access, disclosure, alteration, or destruction. However, no Internet transmission is ever fully secure or error-free.

Please Note: We do not collect or use internet protocol (IP) addresses or other information which could link your participation to your computer or electronic device.

You may decline to respond to any question(s) you prefer not to answer (e.g., by leaving them blank, or by skipping the question). Further, you may decide to withdraw from this study at any time without any negative consequences by advising us, the researchers. Declining to participate in the study will not affect your relationship with the investigators or with the University of Waterloo. Only researchers associated with this study will have access to the records with your information which will be securely stored at the School of Public Health Sciences and David R. Cheriton School of Computer science. We will keep our study data secure for a minimum of 7 years.

With your permission, the interviews will be audio-recorded to collect the information, and later transcribed through MS Teams for analysis. Your identity will be kept confidential. Your name will not appear in any thesis or report resulting from this study, however, with your permission, anonymous quotations from you may be used.

You may share images to help illustrate how your data analysis process happens, however this is not required. The images shared must NOT contain any confidential or identifiable information. These sketches will only be saved as a screenshot and have no connection to your identity. However, with your permission, the anonymous sketches collected may be used for future publications.

The study can be conducted over an online platform - MS Teams. MS Teams has implemented technical, administrative, and physical safeguards to protect the information provided via the Services from loss, misuse, and unauthorized access, disclosure, alteration, or destruction. However, no Internet transmission is ever fully secure or error free. The study can also be conducted in person in a mutually agreed-upon quiet room at the university. A university-provided MacBook Air, preloaded with the toolkit, will be used for the study along with MS Teams for recording and transcription. The video associated with this interview will be deleted immediately, and the audio would be used for transcription, and deleted thereafter.

Identifying information such as names, emails, etc., will be removed from your responses (data) and stored separately for **24** hours and then permanently deleted. You can withdraw your consent to participate and have your data destroyed by contacting us, within this time. All records will then be destroyed according to University of Waterloo policies. Afterwards, it may not be possible to remove your responses because we may not be able to identify which data are yours. The anonymized data set will be secured on encrypted computers for at least 7 years.

There are no direct benefits to you as a participant and no anticipated risks associated with participating in the study. Qualitative research is highly valuable, and through our research, we aim to develop a tool that will make deductive thematic analysis easier and more time-efficient for researchers. Our methodology is designed to be accessible to researchers with little to no programming knowledge, thereby making recent advancements in AI more widely available to the general public.

If you are willing to participate, please suggest a day and time that suits you and we will do our best to be available. Should you have any questions about the study, please email Ansh Sharma at a646shar@uwaterloo.ca.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB #47221). If you have questions for the Board, contact the Office of Research Ethics, toll-free at 1-833-643-2379 (Canada and USA), 1-519-888-4440, or reb@uwaterloo.ca.

We look forward to speaking with you very much and thank you in advance for your interest in this project.

Thank you,

Jim Wallace
james.wallace@uwaterloo.ca
Associate Professor
School of Public Health
Sciences & David R. Cheriton
School of Computer science
University of Waterloo, Canada

Ansh Sharma
a646shar@uwaterloo.ca
Student Investigator
David R. Cheriton
School of Computer science
University of Waterloo, Canada

Thematic Analysis Toolkit for Human-LLM Collaboration

Ansh Sharma[1], Dr. James R. Wallace[1,2]
1 David R. Cheriton School of Computer Science, 2 School of Public Health Sciences
University of Waterloo, Canada

NSERC CRSNG
Google Cloud
THE GAMES INSTITUTE

1 INTRODUCTION: **Thematic Analysis (TA)** is a qualitative "method-ish" used to identify, analyze, and interpret patterns of meaning - or themes - across a dataset. ~ Braun and Clarke (2021).

2 RESEARCH FOCUS:
- LLM effectiveness in Thematic Analysis.
- Accuracy of LLM-generated codes.
- Toolkit usability.

Thematic Analysis Toolkit (workflow)

3a Phase 1 Context Building

BACKGROUND RESEARCH
- **Researcher**: uploads literature and research questions
- **LLM**: defines research avenues

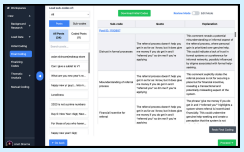
LOAD DATA
- **TA Toolkit**: retrieves data

3b Phase 2 Iterative Code Generation

GENERATING INITIAL CODES (PARTIAL DATA)
- **LLM:** generates codes
- **Researcher:** edits/adds/deletes codes
- **LLM**: generates codebook

GENERATING INITIAL CODES (REMAINING DATA)
- **LLM:** generates codes
- **Researcher:** edits/adds/deletes codes

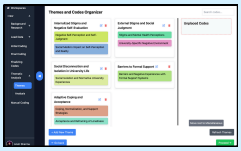
FINALIZE CODES
- **LLM:** finalizes codes
- **Researcher:** edits/adds/deletes codes




TA Toolkit: LLM Generated Codes a. Review Mode (left) b. Refine Mode (right)

3c Phase 3 Iterative Theme Generation

THEME GENERATION AND REFINEMENT
- **LLM:** generates themes
- **Researcher:** edits/adds/deletes themes



TA Toolkit: Theme Refinement

3d Phase 4 Producing Report

PRODUCE REPORT
- **TA Toolkit:** produces report
- **Researcher:** interprets report

4 TA Toolkit Trial : Join the Study!

What to expect:
Perform Thematic Analysis (and qualitative coding) using our TA toolkit!

Email me at:
Ansh Sharma
a646shar@uwaterloo.ca

Ethics Material (Consent Letter)

PARTICIPANT CONSENT FORM

By signing this consent form, you are not waiving your legal rights or releasing the investigator(s) or involved institution(s)from their legal and professional responsibilities.

Research Project Title: Investigating Human-LLM Collaboration for Deductive Thematic Analysis through Computer-Assisted Qualitative Data Analysis Software (CAQDAS)

Investigators:

Ansh Sharma, University of Waterloo

James Wallace, University of Waterloo

I have read the information presented in the invitation letter about a study being conducted by Ansh Sharma of the David R. Cheriton School of Computer science and School of Public Health, under the supervision of Professor James Wallace. I have had the opportunity to ask any questions related to this study, to receive satisfactory answers to my questions, and any additional details I wanted.

I am aware that I may allow excerpts from the conversational data collected for this study to be included in teaching, scientific presentations and/or publications, with the understanding that any quotations will be anonymous.

I am aware that I do not need to share any images with the researcher but can do so if I desire to provide more visual context to the interview. I am aware that all images that might be shared with the research team must not have any confidential or identifiable data.

I am aware that any images I might share with the research team will not be published or made available for anyone outside of the research team.

I am aware that I may withdraw my consent for any of the above statements or withdraw my study participation without penalty by advising the researcher.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Board (REB #47221). If you have questions for the Board, contact the Office of Research Ethics, toll-free at 1-833-643-2379 (Canada and USA), 1-519-888-4440, or reb@uwaterloo.ca.

Please Circle One

With full knowledge of all foregoing, I agree, of my own free will to participate in this study.	YES NO
I agree to be audio recorded, and in the case of online platform, video recorded.	YES NO
I am aware I have the option to share images with the researcher for better understanding or conveying a point. I understand this will be used for analysis and not in a publication.	YES NO
I agree to **not** share any images with confidential or Identifiable data with the researcher.	YES NO
I agree to let my conversation during the study be directly quoted, anonymously, in presentations of research results.	YES NO

Should you have any questions about the study, please contact us:

James Wallace james.wallace@uwaterloo.ca
Ansh Sharma a646shar@uwaterloo.ca

Participant Name (please print)

Participant Signature

Date:_____

Appendix D

Additional Plots and Tables

D.1 Plots

D.1.1 Metrics by expertise level

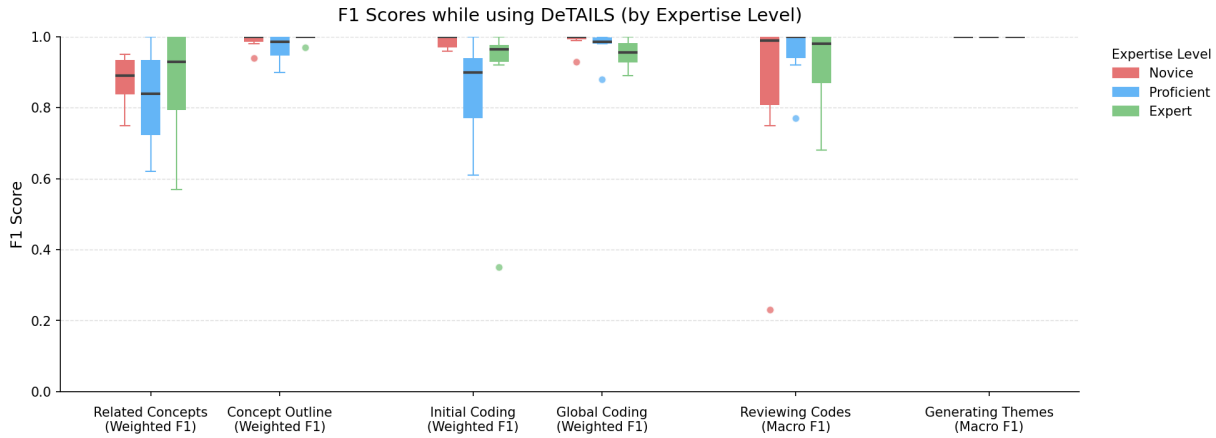


Figure D.1: F1 Scores while using DeTAILS (by Expertise Level)

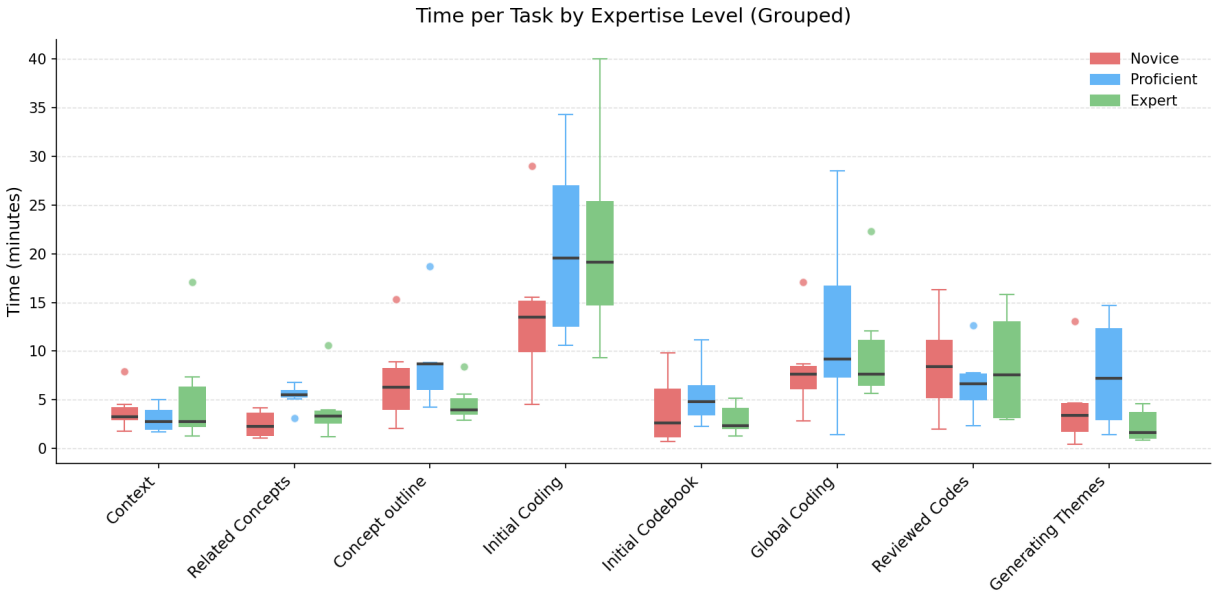


Figure D.2: Participant Timings (by phase) while using DeTAILS (by Expertise Level)

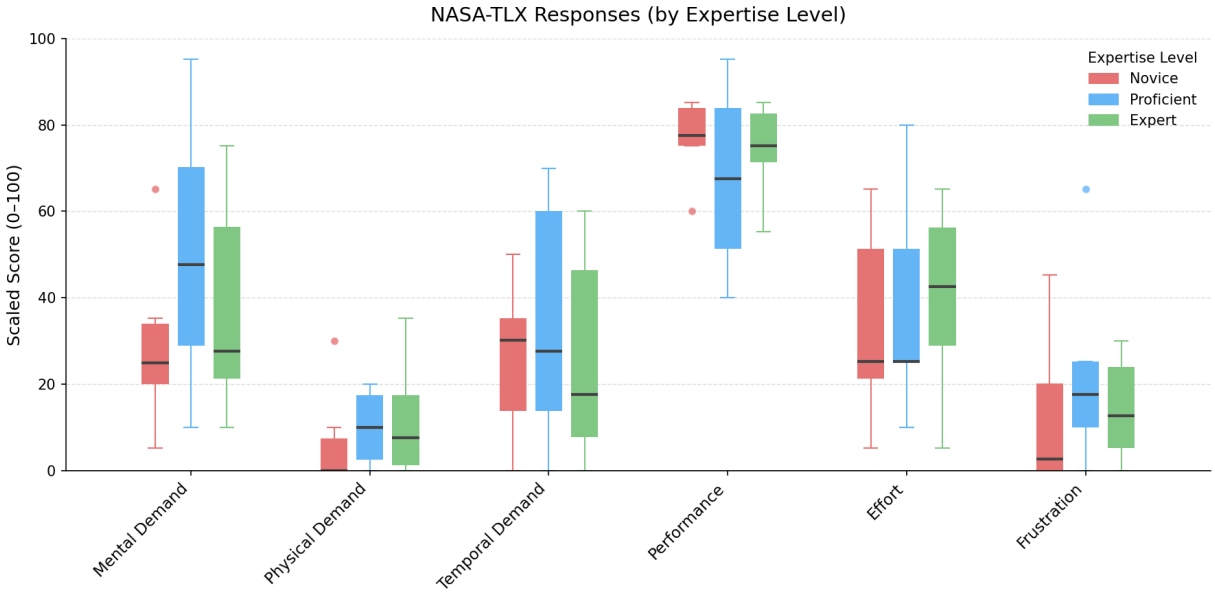


Figure D.3: NASA-TLX Responses (by Expertise Level)

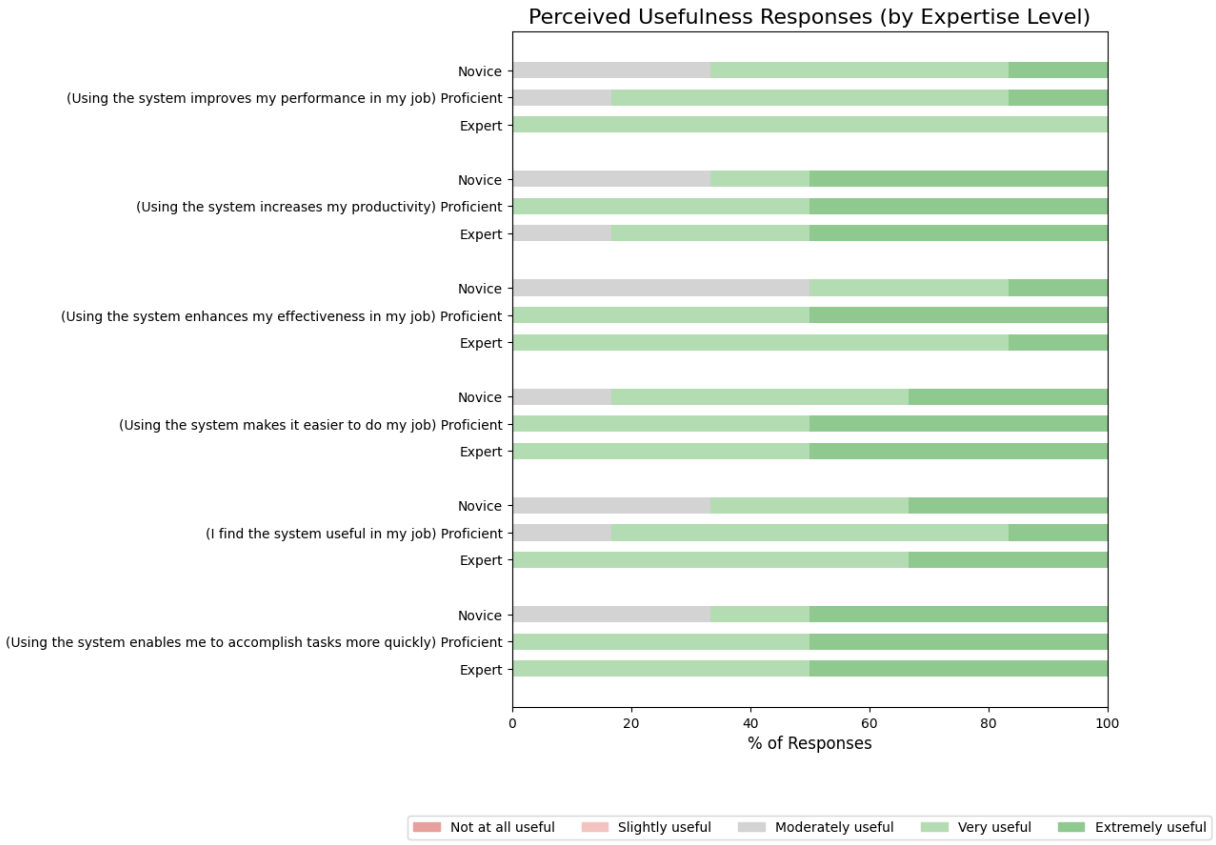


Figure D.4: Perceived Usefulness Responses (by Expertise Level)

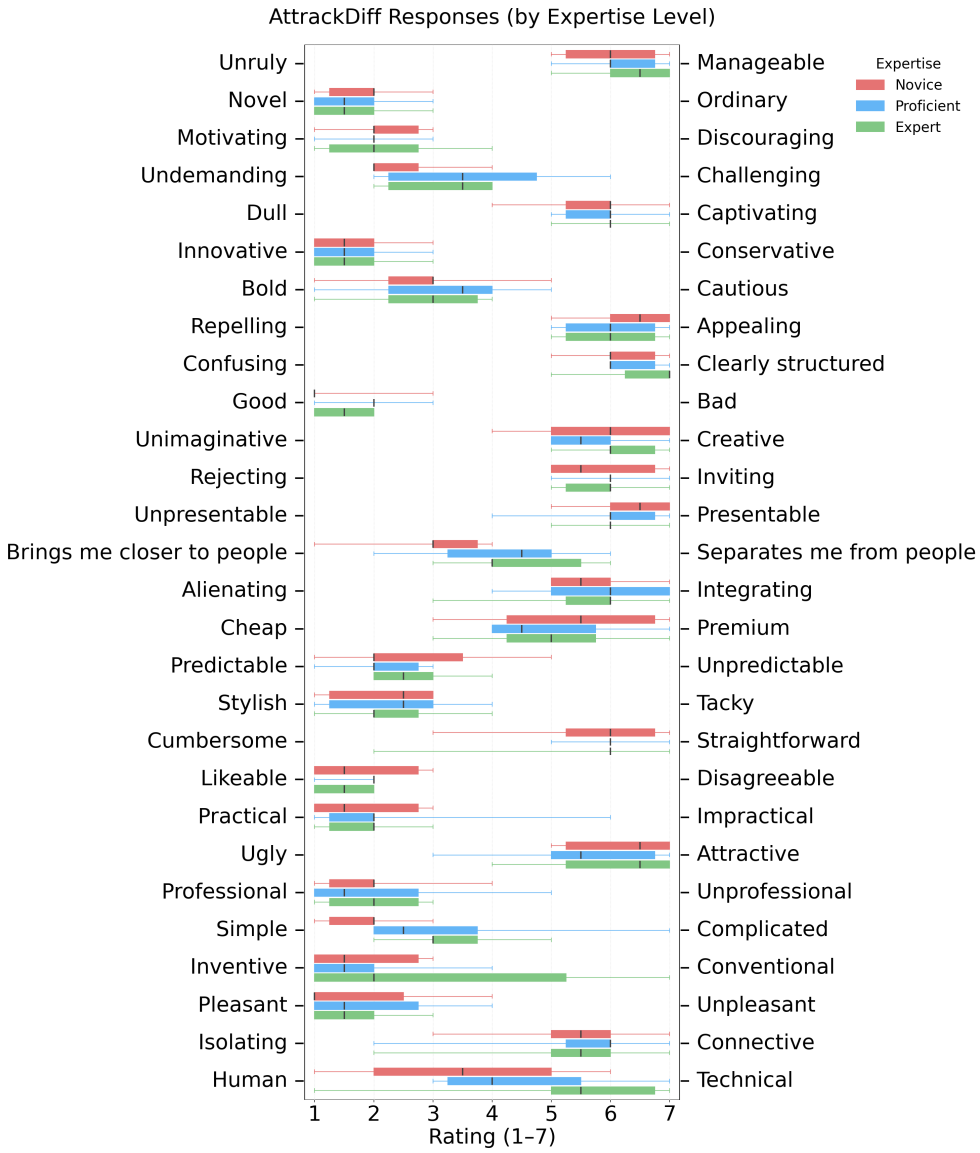


Figure D.5: AttrakDiff Responses (by Expertise Level)

D.2 Tables

D.2.1 Metrics by expertise level

ID	Related Concepts			Concept Outline		
	W. Precision	W. Recall	W. F1 Score	W. Precision	W. Recall	W. F1 Score
Novicep;						
P1	0.6	1	0.75	1	1	1
P2	0.8	1	0.89	1	1	1
P7	0.7	1	0.82	0.98	0.98	0.98
P8	0.9	1	0.95	1	1	1
P9	0.8	1	0.89	0.89	1	0.94
P10	0.9	1	0.95	1	1	1
Proficient						
P3	0.85	0.94	0.89	1	1	1
P4	0.53	1	0.7	0.89	1	0.94
P5	0.75	0.84	0.79	1	1	1
P13	1	1	1	1	1	1
P14	0.46	0.92	0.62	0.83	0.99	0.9
P16	0.95	0.95	0.95	0.97	0.97	0.97
Expert						
P6	1	1	1	1	1	1
P11	0.78	0.97	0.86	1	1	1
P12	1	1	1	0.97	0.97	0.97
P15	1	1	1	1	1	1
P17	0.4	1	0.57	1	1	1
P18	0.65	0.93	0.77	1	1	1

Table D.1: Weighted Precision, Weighted Recall and Weighted F1 Score of Participants while using DeTAILS at the Background Research Phase

ID	Initial Coding			Initial Codebook	Global Coding		
	W. Precision	W. Recall	W. F1 Score	Avg. Cosine Similarity	W. Precision	W. Recall	W. F1 Score
Novice							
P1	1	1	1	1	0.87	1	0.93
P2	1	0.93	0.96	1	1	1	1
P7	0.93	1	0.96	1	1	1	1
P8	1	1	1	1	1	1	1
P9	1	1	1	1	1	1	1
P10	1	1	1	1	0.98	1	0.99
Proficient							
P3	0.63	0.88	0.73	0.91	0.97	1	0.98
P4	0.44	1	0.61	1	0.96	1	0.98
P5	0.84	1	0.91	1	0.79	1	0.88
P13	1	1	1	1	1	0.98	0.99
P14	0.8	1	0.89	1	1	1	1
P16	0.93	0.98	0.95	1	1	1	1
Expert							
P6	1	0.21	0.35	1	0.93	0.98	0.95
P11	1	0.92	0.96	1	1	1	1
P12	0.94	1	0.97	1	0.99	1	0.99
P15	0.85	1	0.92	1	0.8	1	0.89
P17	1	1	1	1	0.93	1	0.96
P18	0.98	0.98	0.98	1	0.92	0.92	0.92

Table D.2: Weighted Precision, Weighted Recall, Weighted F1 Score and Average Cosine Similarity of Participants while using DeTAILS at the Coding Phase

ID	Reviewing Codes			Generating Themes		
	M. Precision	M. Recall	M. F1 Score	M. Precision	M. Recall	M. F1 Score
Novice						
P1	0.7	0.8	0.75	1	1	1
P2	1	1	1	1	1	1
P7	0.33	0.18	0.23	1	1	1
P8	1	1	1	1	1	1
P9	1	1	1	1	1	1
P10	0.98	0.99	0.98	1	1	1
Proficient						
P3	1	1	1	1	1	1
P4	0.94	0.9	0.92	1	1	1
P5	1	1	1	1	1	1
P13	0.75	0.8	0.77	1	1	1
P14	1	1	1	1	1	1
P16	1	1	1	1	1	1
Expert						
P6	0.96	0.96	0.96	1	1	1
P11	0.82	0.86	0.84	1	1	1
P12	1	1	1	1	1	1
P15	1	1	1	1	1	1
P17	1	1	1	1	1	1
P18	0.63	0.75	0.68	1	1	1

Table D.3: Macro Precision, Macro Recall and Macro F1 Score of Participants while using DeTAILS at the Reviewing Codes and Generating Themes Phase