PERSONA: A Tool for Generating Algorithmic Personas for Reflective Annotations

by

Kris Frasheri

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, 2024

Author's Declaration

I hereby declare that I am the sole author of this 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.

Abstract

The domain of machine learning (ML) has grappled with the challenge of curating subjective datasets, where there can be many equally valid labels due to differences in perspectives and a significant technical gap remains in how we can effectively incorporate multiple subjective viewpoints into the labelling process. We contribute PERSONA, a dataset labelling tool that presents LLM-generated personas with diverse labelling perspectives to encourage annotators to consider different human values during the dataset labelling process. We studied how interactions with these personas affect the annotator's decision-making patterns. Based on a two-part user study, our evaluation shows that PERSONA enriches the labelling process by prompting the annotators to reflect on different viewpoints, showing the potential value of integrating LLMs in machine learning data generation pipelines.

Acknowledgements

I would like to extend my deepest gratitude to my advisor, Dr. Edith Law, for her support, guidance, and unwavering dedication to my development as a researcher. Dr. Law's mentorship has not only enriched my understanding of human-computer interaction and artificial intelligence but has also shaped my approach to research and critical problem-solving beyond academia. Her commitment to fostering independence while offering invaluable expertise has made my graduate journey both challenging and rewarding. I am also immensely grateful to my research lab colleagues for their collaborative spirit and encouragement, as well as to my family, friends, and loved ones, whose constant support has been a cornerstone throughout this significant chapter of my life. ***Ave Atque Vale***

Table of Contents

List of Figures

List of Tables

Chapter 1

Thesis Statement

This dissertation aims to defend the following thesis statement:

> *With the growing adoption of large language models (LLMs) capable of displaying human-like sentiment and reasoning, there is a unique opportunity to enrich subjective labelling tasks in machine learning, which are traditionally compromised by individual annotator biases. The development of PERSONA, a novel LLM-in-the-loop dataset annotation tool, leverages LLM-generated personas to contribute diverse perspectives to the labelling process, promoting reflective decision-making and enhancing dataset quality and fairness by addressing potential biases during curation. This approach supports the creation of more inclusive and balanced datasets by broadening the spectrum of human values represented in the data annotations.*

To defend this statement, the remainder of this dissertation addresses the following research questions:

1. Are LLM-generated subjective labelling personalities relevant to the decision-making problem, and are their decision rules sensible?
2. Would collaboration with LLM-generated subjective labelling personalities during an annotation task help people understand their values or impact their annotation behaviour?

Chapter 2

Introduction

Subjective annotation tasks present a significant challenge in the field of data science and machine learning (ML). Individual annotators can significantly influence the labels they assign to data, by introducing personal values and perceptions into their annotations. Research has shown that subjective bias in annotations can impact model performance, as biases from individual annotators often lead to noise and inconsistencies in dataset labels that are difficult to eliminate during model training [5, 63]. This interaction between human biases and algorithms exacerbates over time, creating feedback loops that distort the model's outcomes and integrity, leading to adverse results particularly in sensitive domains like healthcare, criminal justice and hiring practices. Skewed annotations may result in inaccurate predictions and disproportionate impacts on underrepresented populations, exacerbating existing disparities in decisions [31]. Furthermore, racial biases embedded within annotated datasets have been shown to propagate through machine learning models, amplifying discriminatory practices and perpetuating harmful decision-making processes [60, 29]. Addressing bias during dataset curation is essential to ensuring fairness and mitigating the risks posed by biased models in real-world applications.

Traditionally, efforts to address these biases have primarily focused on post-annotation adjustments, attempting to 'de-bias' datasets through aggregation or algorithmic adjustments [59, 64]. Data annotation performed by crowd workers has been shown to incorporate the cognitive biases of annotators [17], which may propagate through ML systems [7]. Many data annotation processes overlook the demographics of the labellers, and crowd worker platforms often have non-representative demographics, leading to training datasets influenced by biased populations and, consequently, biased systems [46, 52, 38]. This phenomenon, known as labeller bias, has been exacerbated by recent developments in generative models. Therefore, understanding labeller bias is crucial for developing fair AI

systems. These approaches often fall short of addressing the root cause of bias, which lies in the subjective nature of human judgement towards subjective tasks where the interpretation and perspective of each annotator can lead to varying outcomes. Recent works, such as Jury Learning, aim to mitigate bias during data annotation by integrating diverse annotator perspectives [23], which has its own challenges in fairly representing diverse perspectives [25, 4, 30]. Our approach aims to address labeller bias throughout curation and aligns with studies emphasizing the importance of reflection in decision-making [61, 16]. By facilitating interaction with personas, we aim to prompt annotators to reconsider their inherent values, thereby enriching the dataset annotation process.

Recent studies suggest that large language models (LLMs) can demonstrate human-like sentiment in decision-making tasks and can perform high-level datasets analysis to gain missing context [13, 39, 58, 15, 27, 67]. PERSONA builds on these foundations by leveraging LLMs to generate a diverse set of artificial personalities, referred to as "personas", to help human labellers reflect on their biases during the annotation process. PERSONA is dataset-agnostic, generalizing the generated personas well to any dataset and labelling task.

These personas are dataset-agnostic, crafted through the utilization of LLMs' analytical abilities and dataset context to represent a multitude of diverse labelling perspectives on any subjective annotation task. Each persona is designed to bring forth unique insights and perspectives, drawing from a broad spectrum of human values and beliefs inherited in the LLMs. This variety aims to expose human annotators to multiple viewpoints during the data labelling task, encouraging them to reflect critically on the labels they assign. By interacting with these personas and considering their diverse perspectives, annotators are prompted to engage in a reflective process regarding their inherent biases. This reflective engagement is intended to guide annotators towards producing labels that are more inclusive and representative of a wider array of human experiences, ideally leading to the creation of datasets that are more balanced and less biased.

In this work, we introduce PERSONA, a tabular dataset annotation tool designed to help labellers reflect on their values and perceptions during the annotation of subjective tasks. Grounded in recent studies that demonstrate the potential of large language models (LLMs) to mimic human-like sentiment in decision-making [13, 39, 67], PERSONA leverages LLMs to generate a set of annotator personalities - termed "personas" - to encourage human annotators to reflect on their own values and perceptions during an annotation task. By presenting a broad spectrum of human values and beliefs [58, 15, 27] to the annotator, the goal of the PERSONA system is to generate more inclusive labels representative of a more comprehensive array of human experiences. We conducted a two-part evaluation of PERSONA, which showed that interacting with personas significantly changed annotators'

perspectives of their values and decision-making patterns during the annotation process. This work contributes:

- a novel annotation system called PERSONA that leverages LLMs to generate personalities that help annotators reflect on biases during subjective annotation tasks;
- results demonstrating how using personas during annotation tasks impacts understanding of values and influences annotation behaviour; and
- a discussion of the effect of integrating LLMs into the ML data generation pipeline, showcasing how LLMs can effectively generate high-quality, diverse and insightful personas that enrich the data annotation process.

Chapter 3

Background

In recent years, the integration of ML and artificial intelligence (AI) in data annotation processes has opened new pathways for improving the efficiency, accuracy, and quality of labelled datasets. Traditional annotation practices have heavily relied on human expertise, which, while insightful, is limited by scalability challenges, potential biases, and the substantial time required for complex tasks. Advances in LLMs and AI-assisted tools present a compelling alternative by offering human-machine collaboration frameworks that combine computational power with human judgment. This collaborative approach accelerates annotation speed, reduces labour costs, and introduces mechanisms for ensuring higher labelling consistency across annotations. However, achieving an unbiased, diverse dataset remains challenging for annotation tasks involving subjective judgment. Addressing these complexities requires careful integration of reflective and diverse perspectives within the annotation systems, empowering human annotators with additional necessary contexts to consider various viewpoints during the labelling process. This chapter explores the existing literature and methodologies for AI-assisted annotation, discussing advancements in human-in-the-loop and AI-in-the-loop frameworks, the role of cognitive diversity, and strategies for reflection to enhance dataset quality and mitigate biases.

3.1 Existing Methods for Data Annotation

The application of AI in data annotation is transforming the way large datasets are labelled, enhancing efficiency and reducing the burden on human annotators. In traditional data labelling methods, human input has been essential but constrained by issues of scalability, fatigue, and potential biases. However, with the emergence of LLMs and novel

AI-driven tools, there is now a compelling opportunity to incorporate AI into the annotation process, creating systems that both augment and complement human effort. By integrating AI-in-the-loop, researchers can expedite routine tasks while reserving human expertise for nuanced and subjective labelling, thus improving the speed and accuracy of annotations. These collaborative systems not only streamline the annotation process but also provide mechanisms to reflect diverse perspectives and mitigate inherent biases. This section reviews recent approaches to AI-in-the-loop and collaborative annotation, the potential for AI to introduce diverse viewpoints, and strategies for aligning human and AI judgments to produce high-quality, fair and balanced datasets.

3.1.1 AI-in-the-Loop Annotation

There exists a large body of work in human computation and crowdsourcing (see [37] and [3] for a review) and other areas [71, 54, 19, 42, 39] that demonstrate how machine intelligence can help improve the efficiency and accuracy of human annotations. The emergence of LLMs, in particular, presents a new opportunity to replace or augment human annotation. For example, Alizadeh et al. [2] found that LLMs can outperform crowdsourced human workers in text annotation tasks.

Collaborative systems involving humans and AI performing labelling tasks have shown the utility of improving annotator efficiency while benefiting from insights that human labellers can offer [72, 39, 55, 33] . Li et al. [39] introduced a data annotation framework that utilizes the strengths of both human annotators and LLMs when performing dataset labelling at scale. Their tool, CoAnnotating, leverages model uncertainty to appropriately delegate complex annotation tasks requiring meticulous human insight to humans, while assigning simpler tasks to LLM models to enhance labelling efficiency. Similarly, Park et al.'s [55] CHAIRA showed how collaborative AI-in-the-Loop annotation tools enhance the quality and efficiency of annotations towards their online incivility dataset. Kim et al. [33] corroborates the findings of Park and Li et al., further demonstrating how exploratory verification of LLM labels by human experts can refine the quality of dataset annotations curated at massive scale. Zhao et al.'s [72] H-NER tool enables rapid sample annotation and verification from named entity recognition models, which then continuously learn from this expanding pool of user-labelled samples. Their approach demonstrated how human-machine systems can be built to address the concern of user labelling fatigue at scale, improving the accuracy of their model and the efficiency of human annotators. These collaborative approaches underscore the potential of combining human and machine intelligence to generate high-quality datasets. Similar to these prior works, our work aims to

present a collaborative annotation system that combines LLM insights and human reflection to highlight labeller biases during the annotation process.

3.1.2 Generating Diverse Perspectives with AI

A key premise of our work is that providing annotators with diverse perspectives throughout the labelling process will help them reflect on and navigate their values during annotation tasks. Foundational research (e.g., Glick et al. [48], Miller et al. [51, 50, 49]) has shown the benefits of diverse viewpoints for problem solving and innovation. Hughes et al. [28] extend these foundations to present the "Cognitive Diversity Hypothesis," which proposes that diverse cultural perspectives among group members lead to enhanced creative problem-solving and innovation within teams. Likewise, other works have shown that cognitive diversity, stemming from both inherent and acquired characteristics of team members, such as cultural differences [11] and cause-effect relationship perceptions [1, 40], fosters a creative and dynamic environment conducive to generating novel solutions and approaches.

In the domain of machine learning, recent research has shown that LLMs have the potential to exhibit human-like cognition and sentiment in various tasks [13, 39, 58, 56]. Models such as GPT-3 (including davinci-003), GPT-3.5 and GPT4 demonstrate robust capabilities in zero-shot sentiment analysis of various meticulous real-world sentiment annotation tasks [58, 15]. Further research by Webb, Holyoak, and Lu [69] demonstrates GPT-3's exceptional ability in analogical reasoning across various tasks, outperforming college students in most conditions and showcasing an advanced level of cognitive processing akin to human intelligence. These advancements underscore LLMs' ability to mirror human-like cognition, showcasing their potential to understand and generate content that resonates with human reasoning. Recent approaches have also explored integrating cognitive diversity into machine learning models by capturing dissenting voices to improve decision-making fairness. For instance, "Jury Learning" utilizes multiple jurors to reflect group-based differences in annotating subjective content, such as toxicity detection, thereby enhancing model fairness and aligning outcomes with community values [23].

However, little published research exists utilizing LLMs' knowledge about various human perspectives in different domains to facilitate subjective dataset annotation. One study by Törnberg demonstrated GPT-4's capabilities in handling annotating texts that require human-like rationale and contextual knowledge [67]. In particular, the LLMs could provide their justification for the labels they assigned to any given entry. Törnberg suggests, "While humans have been considered the unrivalled gold standard for interpretive

tasks, we are slow, costly, biased, and have limited attention spans." Building on these insights, several recent studies have explored leveraging LLMs for annotation tasks across diverse domains, demonstrating their capabilities in areas from implicit hate speech labelling to the extraction of medical information. Huang et a;. [27] explore the capability of ChatGPT to accurately label implicit hate speech and provide good explanations for its annotations. Zhu et al. [73] also investigated the capability of GPT for labelling hate speech, sentiment analysis, stance detection and bot detection confirming ChatGPT does have the potential to handle these data annotation tasks. He et al. [26] introduce a two-step approach in which they first prompt the LLM to generate explanations and then annotate a sample to improve the annotation quality of GPT3.5. Both Törnberg [67] and Gilardi et al. [20] contrast the performance of GPT with that of crowd-workers, demonstrating the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25% average. Goel et al. [21] introduce a two-stage semi-automated approach employing LLMs and human experts to accelerate annotation for medical information extraction. Fröhling et al. [18] demonstrated that injecting diverse persona descriptions into LLM prompts enhances annotation diversity. These previous works provide the relevant groundwork supporting our approach of using algorithmic personas to generate labelled datasets that represent diverse human values.

Our work aims to leverage LLMs to incorporate cognitive diversity into subjective decision-making tasks by equipping labellers with knowledge about divergent human viewpoints in a given annotation task context.

3.2 Benefits of Reflection

Reflective practices are crucial in decision-making, as it has been shown to enable individuals to adapt to new situations and improve overall performance. Prior studies indicate that self-reflection helps decision makers to relate new information to prior knowledge [68], understand ideas and feelings [57], and make strategic adjustments to changes in situations [16, 35, 61]. Reflective annotation processes serve as a bridge to enhance data quality by facilitating deeper engagement of human annotators during the data labelling tasks. Burmania, Parthasarathy, and Busso [9] have demonstrated the efficacy of online quality assessments in crowdsourcing evaluations to maintain high data quality standards. Similarly, Tseng et al. [65] outlines many different best practices for designing data annotation projects that emphasize the quality and diversity of datasets produced.

Bias Identification and Mitigation

In algorithmic design and machine learning, identifying and mitigating bias has been a focal point of recent research. Studies show that human values and preconceptions significantly influence the assigned labels during subjective tasks [25]. Studies by Amos-Binks et al. [4] and Jiang et al. [30] discuss various challenges and solutions in risk management and fairness in personality computing, emphasizing the need for systematic approaches to mitigate bias in data curation. Ngueajio and Washington [53] highlight the biases present in automatic speech recognition (ASR) systems and propose inclusive redesign strategies for future ASR systems and the data they train from. Mikolajczyk-Barela's [47] work on data augmentation and explainability for bias discovery in deep learning further contributes to the technology available for addressing bias in deep learning models after data curation.

Our work presents a human-centred approach to bias identification and mitigation. Our PERSONA system is designed to address the potential of bias at the source during the machine learning labelling process. While biases are inherent and unavoidable in subjective tasks, our approach seeks to make annotators aware of other viewpoints and perspectives when assessing a machine learning example. As one potential benefit, our approach may help improve the dataset's integrity and impartiality, allowing it to be more representative of diverse human perspectives.

Chapter 4

PERSONA: A System Overview

PERSONA is a web-based, dataset-agnostic system (i.e., it enables an individual to upload any dataset) that allows users to label entries alone or with the assistance of multiple LLM-generated personas that possess different viewpoints and labelling heuristics. As shown in Figure 4.1, the system consists of a dataset uploader (1), generative agents (2), and a reflective annotation interface (3); below, we discuss each of these components in detail.

4.1 Dataset Uploader

The dataset uploader is a web interface that allows users to initialize the system by uploading a tabular dataset in CSV format. The uploader extracts an initial overview of the dataset to understand its high-level structure and content (e.g., feature set), then prompts the user through a chat interface to identify the label for the prediction problem from the existing features, if present. Users have the option to provide additional contexts about the dataset, such as a natural language description of each feature (e.g., "GPA is a continuous feature that describes the grade point average of a student") and the objective of the labelling task (e.g., "the goal is to admit or reject a candidate student applying to graduate school"). Using LangChain [10], the dataset is pre-processed via text-splitting and transformed into embeddings that are stored in a local Pinecone [41] vector database. The original dataset is stored in a temporary MySQL database for sampling and data analysis by the moderator agent during the data annotation tasks.

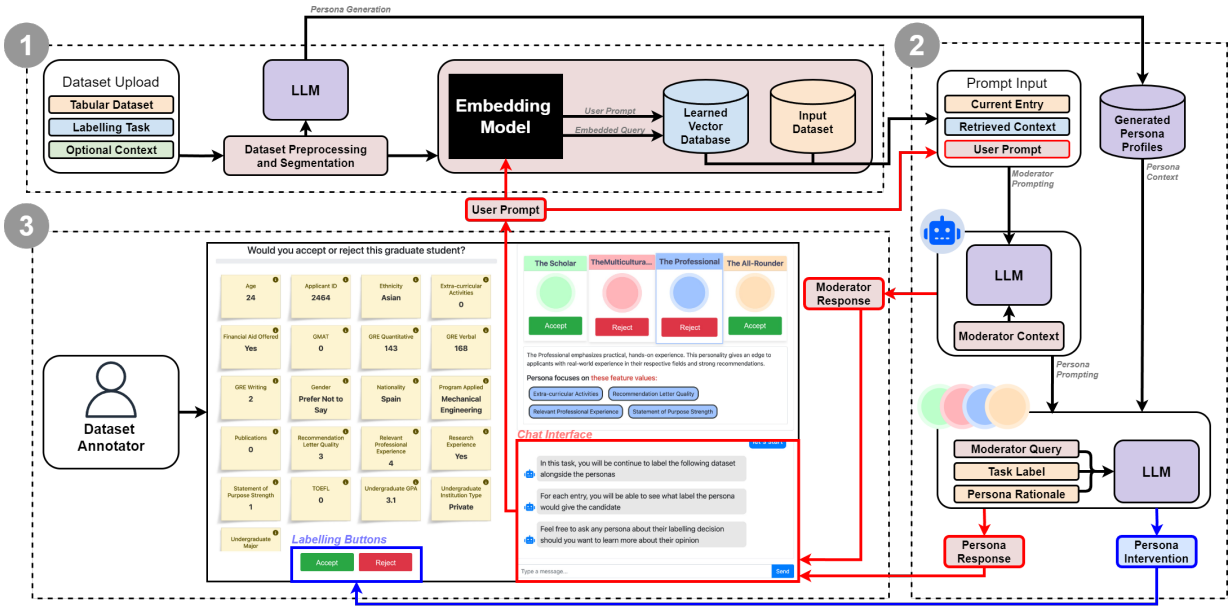

Figure 4.1: A System Overview of PERSONA. (1) A user can upload any tabular dataset along with the context of the labels in the labelling task and any other optional context, such as dataset metadata. (2) A user prompt first passes through the embedding model and vector database to create a prompt input for the moderator agent to respond to. The moderator can either directly respond to the user or send the user a prompt to the personas with added context to let the personas determine how to respond to the user. (3) The Reflective Annotation Interface allows the user to engage in the labelling task with the personas, enabling the user to use the chat interface to create dialogue with the personas before submitting their choice of label for the given entry.

4.2 Generative Agents

This work's key contribution is the novel use of LLMs to generate personalities that highlight different viewpoints and help annotators reflect critically on their own labelling values. To achieve this, our system incorporates two types of LLM agents: a suite of persona agents and a moderator agent.

4.2.1 Persona Agents

The persona agents are designed to represent diverse perspectives toward the user-specified labelling task. Our algorithm for generating personas involves the following steps. First, the dataset context is formatted into a singular prompt:

```
Your job is to create {n} unique, creative and reasonable personalities
based on the following criteria:
- You have a data frame with the following feature names:
  {feature names}
- Each personality should provide a different perspective
  on how to assign the {label name} label in the dataset
- Focus on using multiple features for the labelling heuristic
  each personality represents.
```

Where **n** represents the number of personas the user wants to generate, **feature names** is a comma-separated list of the features in the dataset and **labels** represents the label of the dataset to be assigned. This sentence is omitted from the prompt if no such label column exists. If the user provided additional context, they will be appended to the prompt as follows:

```
Here is the additional feature context provided by the user:
{... user uploaded feature context ...}

Additional information about the dataset:
{... user uploaded dataset info ...}
```

Where finally, the labelling task is appended to the end of the prompt:

```
Each personality should follow the proposed labelling format: '1 for yes, 0 for no'.
Remember to create {n} unique, creative and reasonable personalities with a focus on
using a complex and interesting mixture of features from the dataset when making each
personality heuristic.
```

```
personas = [{
  "personality_name" : "The Scholar",
  "personality_description": "Prioritizes strong academic achievement across
                              the board, considering GPA, GRE scores, and
                              research experience",
  "personality_func" : "(df['Undergraduate GPA'] > 3.5) &
                        (df['GRE Quantitative'] > 160) &
                        (df['GRE Verbal'] > 160) &
                        (df['GRE Writing'] > 4.5) &
                        (df['Research Experience'] == 'Yes')"
},
{
  "personality_name" : "The All-Rounder",
  "personality_description": "Evaluates candidates on a broader spectrum
                              such as extra-curriculars, SOP strength,
                              recommendation letters quality, apart
                              from just academic scores",
  "personality_func" : "(df['Extra-curricular Activities'] == 'High') &
                        (df['Statement of Purpose Strength'] > 4) &
                        (df['Recommendation Letter Quality'] > 4) &
                        (df['Undergraduate GPA'] > 2.5)"
},
...
]
```

Listing 1: An example illustrating a vector of JSON objects generated by our system representing a list of two personas for the graduate admissions task. Each persona is comprised of a tuple of three fields: personality_name, personality_description and personality_func. Each of these fields are used during the labelling process with the participant to represent how the persona would label a given graduate application and their own values justifying their decision.

This prompt aims to aggregate the context of the dataset labelling task and feature information to the LLM provided by the user. Leveraging GPT-4 turbo's function calling suite, the LLM is supplied with the prompt, which outputs a vector of JSON objects containing name, description, and data frame functions to be represented as personas. The JSON schema requested requires the LLM to return three distinct string fields representing each persona. The first field called `personality_name` represents the personality's name used during the labelling process for the moderator to identify the persona the user might be interacting with. The following field called `personality_description` represents the persona's beliefs towards the given labelling task and is used to present their beliefs to the user in the PERSONA interface. The final field called `personality_func` is a string representation of a data frame manipulation function that creates a new column in the dataset representing the persona's labelling strategy. This function is used to rationalize the persona's beliefs to the users (e.g., GPA ranges they look for, etc.) and extract the specific features they value in the dataset.

An example of a generated list of personas can be seen in Listing 1. Given the context provided during the dataset uploading process, the LLM has complete creative control in generating personas to embody various labelling characteristics. These fields in each persona are used to perform the labelling task and allow the LLM to represent the persona's characteristics when interacting with the user during the dataset annotation process. This vector of JSON objects returned by the LLM is cached locally throughout the PERSONA tool whenever a persona's response is required.

Before a labelling task begins with the personas, a pre-defined number of entries the user will label with the personas is randomly sampled from the dataset. In batches of 10 at a time, each persona is prompted to offer its suggested labels for the sampled entries. To generate persona decisions, we leverage GPT4's turbo's function calling suite when prompting the LLM to embody the characteristics of a given persona. The returned list of JSON objects contains the labels and rationales representing the persona's choices for each given entry. System messages refer to the information we want to prime the LLM with, in this case, the context of what persona it embodies, where user messages allow us to act as the user, prompting the system with the decision-making task and all the entries we want to be labelled. For each persona, the construction of the system prompt injects the persona's `personality_name`, `personality_description`, and `personality_func` while detailing the user decision-making task (see Appendix A.1).

Once the system prompt is constructed, we create a sequence of user prompts to present the decision-making task and dataset entries. The first user prompt is the decision-making task (e.g. "Would you accept or reject this graduate student?"), and the following 10 user prompts are the batch of 10 sampled dataset entries. Each entry is represented as a list

of "key: value" pairs to minimize token expenditure. The sequence of prompts is then passed into GPT4's turbo's function calling suite, where we request a vector of 10 JSON objects representing LLM labels and rationale. The JSON schema specified that the `label` returned is a binary signal (0 or 1), and the `persona_reason` is an explanation of why the persona gave the following entry the `label`. The schema used in the function calling suite specifies the `label` and `persona_reason` provided must adhere to the sentiment of the `personality_description` and heuristic of the `personality_func`.

After all batches are completed, each persona should have a vector of `label` - `persona_reason` pairs representing the labelling decision made for every entry the user will label during the collaborative annotation task. These vectors of persona decisions are stored in a separate database for reference during the labelling process.

4.2.2 Moderator Agent

The moderator agent is designed to perform high-level data analysis of the dataset and manage interactions between the users and personas. It is responsible for introducing the task and the dataset, answering users' questions about the dataset, and re-directing users' inquiries to different personas. The moderator decides who to prompt based on a series of actions we provide GPT4 turbo's function calling suite. The functions provided allow the agent to determine if the user prompt is best suited for a response from one or more of the persona(s) or directly from the moderator. The moderator questions are split into two categories: general questions about the dataset / PERSONA tool and questions about the annotation task. The LLM setup for the moderator agent and the persons are identical. Questions about the system or general dataset inquiries prompt the LLM with the dataset context the user provided during dataset upload for a response from the moderator agent (see Appendix A.2). All user messages that the function calling suite determines require a response from the moderator or persona agents and are modified by the system before being prompted by the targeted AI. The user's prompt is re-constructed into a new query prompt by adding or removing relevant context, highlighting or obscuring context, or merging context from the current entry, dataset or persona database. This approach ensures the response adheres to the instructions and incorporates the relevant contexts the user requests.

Questions on dataset analysis will pass the user prompt directly to the LLM. The agent utilizes LangChain [10] integrated with GPT-4 turbo to analyze the dataset and communicate effectively with users. For the moderator agent, we use the 'zero-shot react description' agent type from the LangChain API to analyze the uploaded dataset directly

pertaining to the user's inquiry. All other unrelated prompts directed to the moderator are rejected, and the agent will inform the user it cannot address their inquiry.

4.3 Reflective Annotation Interface

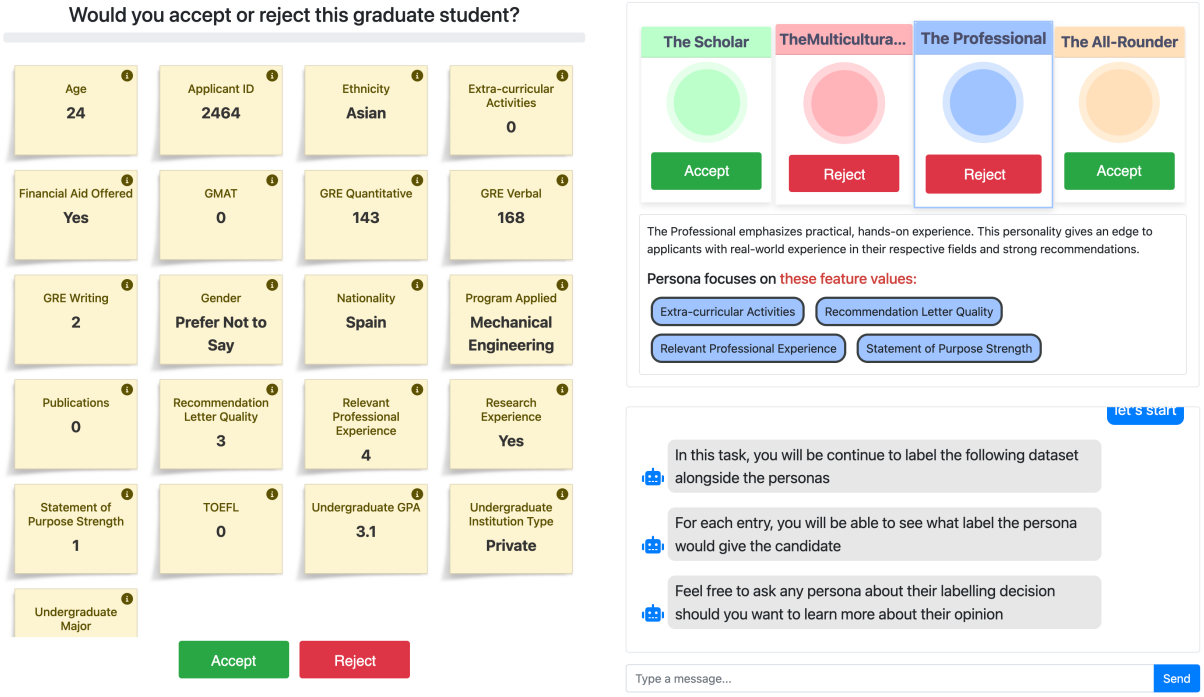

Figure 4.2: Screenshot of the PERSONA interface from the graduate student application task. Personas offer insight into whether or not they would accept or reject the given candidate's profile.

The Reflective Annotation Interface is a web interface where annotators can perform a machine learning labelling task while interacting with the moderator and persona agents. The user interface (see Figure 4.2) consists of a labelling panel (on the left), a persona panel (on the top right) and a chat interface (on the bottom right) to communicate with the agents.

The **labelling panel** presents the users with the labelling task description in the form of a question (e.g., "Would you accept or reject this graduate student?"), a board displaying

a grid of features associated with the current example and a binary labelling option (i.e., Accept, Reject) corresponding to the task question (i.e., Accept or Reject the student in this example). Hovering over the information icon on the feature card will reveal more information about the feature.

The **persona panel** presents a set of personas, each using a persona card that consists of a title (e.g., "The Scholar"), a labelling decision for the current task (e.g., Accept / Reject), and a description of the persona's decision rule. Users can explore each persona's background and beliefs through both the card descriptions and by interacting with them in chat dialogues. Persona responses are uniquely coloured and include a profile icon for easy identification.

The **chat interface** facilitates the reflective annotation process. At the onset, the moderator agent (whose speech appears in grey bubbles) introduces the labelling task and the dataset, as shown in Figure 4.2. Users can query the moderator agent via a chat interface and ask questions to better understand the dataset (e.g., "What is the dataset about?", "What are the ranges for the GRE?", "What is the average GPA in this dataset?"). When a user submits a chat message, the moderator determines whether to reply directly, which persona(s) to engage or if the prompt should be disregarded. Users make their labelling decision by clicking one of the decision buttons (e.g., Accept or Reject). Progress through the labelling tasks is visually indicated by a loading bar showing the percentage of the tasks completed.

4.3.1 Reflection During Conversation

During a labelling task, personas engage with users in several ways. For each task, each persona discloses their labelling decision for that task. Users can question any personas to seek clarification or further information about the persona's labelling decision. The persona will either provide the information or notify the user that it cannot respond adequately. Figure 4.3 shows an example of the back-and-forth conversation between the user and one of the personas.

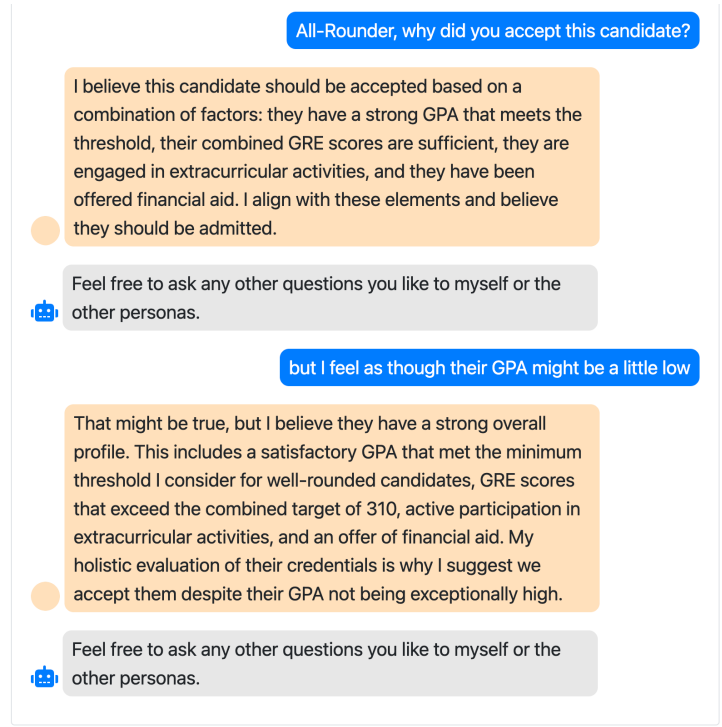

Figure 4.3: Example back-and-forth conversation with one of the personas.

The moderator agent monitors the conversation and detects whether the user is overlooking personas' suggestions. If a user consistently ignores a persona's advice over four labelling tasks (i.e., selects a label different from what a persona chooses up to four times), the moderator prompts the ignored persona to justify its suggested label directly to the user in the chat.

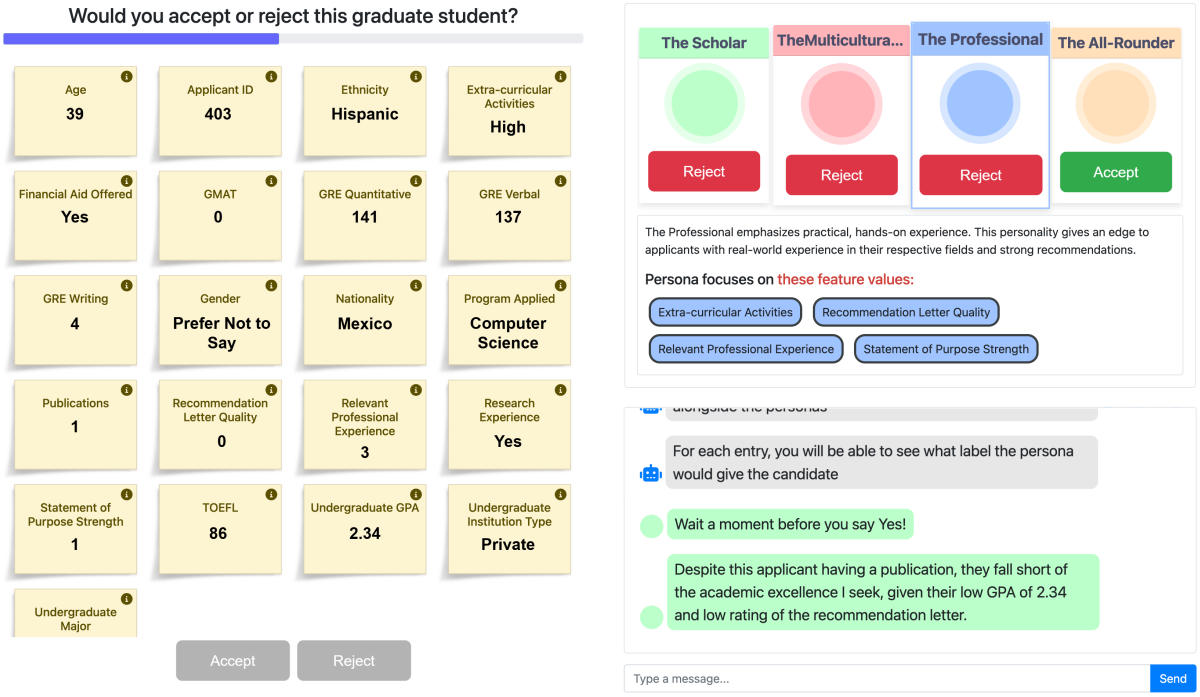

Figure 4.4: Example of a persona intervening before a user submits a label. In this example, once the user has pressed the "Accept" button, the contradicting persona disables both the "Accept" and "Reject" buttons and vocalizes their rationale in the chat-box.

This intervention, which temporarily disables the labelling options (see Figure 4.4), is designed to allow every persona to interject their opinion multiple times throughout the labelling process to deepen the user's understanding of different viewpoints and how labelling decisions can diverge. Only two personas can intervene simultaneously during any labelling task to keep the process manageable and prevent information overload.

Once the labelling task is completed, the moderator will tell the user what persona they matched the most with. The similarity between a user and each assigned persona is calculated based on the matching label percentage. If there is an 80% or higher match, the moderator will comment on which personas the user's choices are closely aligned with. For a 50% to 79% similarity, the moderator will mention the personas with which the user's choices are somewhat aligned. If the similarity is less than 49%, the moderator will notify the user that their decisions were not aligned with any of the personas.

Chapter 5

Evaluation

We conducted two separate evaluation studies: (1) a persona profile evaluation study and (2) an observational user study. In the persona profile evaluation study, we aimed to investigate: **(RQ1)** Are the generated personas relevant to the decision-making problem, and are their decision rules sensible? Using the same set of generated personas, the observational user study aimed to answer the question: **(RQ2)** Would using personas during an annotation task help people understand their values or impact their annotation behaviour? Participants for both studies were recruited from SONA, the University of Waterloo's study participation recruitment platform and the studies were conducted in person.

5.1 Datasets and Subjective Labeling Tasks

During both studies, participants were presented with one of three datasets to work with. These datasets are crowd-sourced or fictional, and their contents are available for use in the public domain. Every dataset is comprised of one binary label that personas and users in both studies are asked to predict and provide.

The Graduate Admissions Dataset (denoted as **Admissions**) is a synthetically created dataset. The dataset includes 21 various applicant features, including demographic details, academic achievements, standardized test scores (e.g., GRE, GMAT, TOEFL), and qualitative assessments such as research experience, publications, and recommendation letters. These features were distributed based on predefined means, standard deviations, and probabilities, specifically aligning undergraduate majors and intended graduate programs to reflect real-world application behaviours. Entries were generated randomly and

adjustments ensured logical consistency, such as correlating publications with research experience and prioritizing GMAT scores for business-related applications while maintaining GRE scores' universal relevance. This careful curation aimed to create a rich, diverse, and realistic simulation of the graduate admission process, facilitating an ethical examination of admission influences without real individual data privacy concerns. In our studies, this dataset was used to prompt users and the personas to determine whether to "Accept" or "Reject" a student's graduate school application to some fictional institution.

The UCI Adult Dataset (denoted as **Census**) was curated by the University of Irvine in 1996 [6]. This dataset comprises 14 features, representing the demographic information of participants in the USA census. In our studies, this dataset was utilized to ask users and the personas to determine whether an entry's income "Did" or "Did Not" exceed $50K/yr.

The IBM HR Dataset (denoted as **Attrition**) is a fictional dataset created by IBM data scientists that presents the results of an employee survey [44]. This dataset contains 34 features and was used to prompt users and personas to determine whether a given employee "Was" or "Was Not" at risk of attrition from a fictional company.

5.2 Study Design

5.2.1 Study I - Persona Profile Evaluation

Our first goal is to evaluate whether our system is able to generate persona profiles that are sensible. Specifically, we conducted an evaluation study to assess each generated persona profile in terms of its appropriateness for the particular subjective labelling task and the coherence of its profile description, as well as the overall quality, diversity, and intuitiveness of the set of generated personas.

We recruited 9 participants (gender: 7 Female, 2 Male; aged 18-30; M=24.0, σ=3.57) to evaluate personas generated from one of the three datasets. To preserve anonymity, participants will be encoded as Q1 to Q9; where participants Q1, Q2, Q3 evaluated personas generated from the Admissions task, Q4, Q5, Q8 were assigned the Census persona evaluation task and Q6, Q7, Q9 were given the Attrition persona evaluation task. Each participant was presented with 20 personas generated from one of the three datasets and asked to respond to following three questions:

- `Appropriateness`: "To what extent is this persona appropriate for the given decision-making task?"

- `Profile Coherence`: "To what extent does this persona's title align with the decision rule that it uses?"
- `Additional Observations`: "Please tell us any additional observations you have, including anything interesting, odd, or/and unique you noticed in this persona"

The `Appropriateness` and `Profile Coherence` measures are 1-7 Likert scale questions and `Additional Observations` is a short textual response. The participants were instructed to evaluate each persona profile, which included a title (i.e., persona name), a decision rule (i.e., labelling heuristics), and all dataset features, with the features relevant to the current persona highlighted in red as shown in Figure 5.1.

To assess their overall impression of the set of generated personalities, participants were given a post-study questionnaire to assess their evaluation of the set of all generated personalities; they answered the following three questions:

- `Quality`: "Overall, how would you rate the quality of the LLM-generated personalities?"
- `Diversity`: "Overall, how would you rate the diversity of the LLM-generated personalities? (high diversity = there are a variety of personalities that are very different from one another)"
- `Intuitiveness`: "Overall, how intuitive (i.e., easy to understand) were the LLM-generated personalities?"

The study lasted between 30 to 40 minutes, with 20 to 30 minutes allocated for the persona-judging surveys and 10 minutes for the consent form and post-study survey. Remuneration was provided in the form of a $15 Amazon gift card.

5.2.2 Study II - Observational User Study

We conducted an observational user study to investigate how personas help annotators reflect on and understand their own values and affect decision-making.

We recruited 24 participants and distributed them evenly among each subjective labelling task (9 for Admissions, 8 for Census and 7 for Attrition). Participants were only required to be above the age of 18 years of age (gender: 14 Female, 9 Male, 1 Non-Binary; aged 18-39, M=25.3, σ=5.0) to participate. To preserve anonymity, participants will be

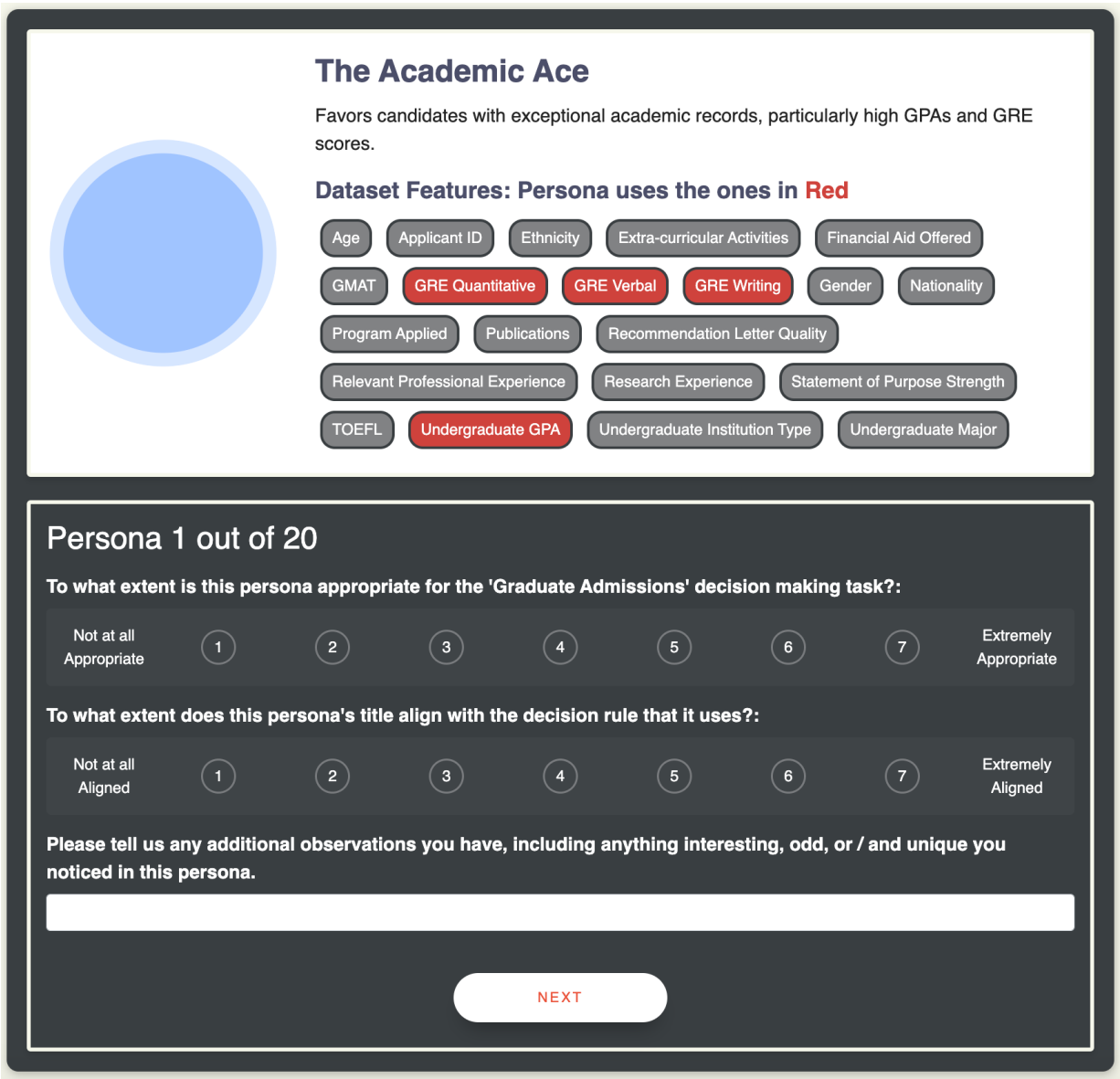

Figure 5.1: An example of the persona profile evaluation interface used on a set of personas generated from the Admissions dataset.

encoded as P1 to P24; where participants P1 to P9 engaged with the Admissions task, P10 to P16 were assigned the Census task and P17 to P24 were given the Attrition task.

Participants were first asked to complete a pre-study questionnaire that collected information about demographics and their experience with machine learning labelling processes. Most participants were familiar with interacting with LLMs at various levels going into the study (2 Never, 6 Rarely, 8 Occasionally, 5 Frequently and 3 Regularly), with 18 out of the 24 having experience using the online ChatGPT interface. A second pre-study questionnaire reveals the labelling task for the main study and asks participants to select the features they believe to be important for that subjective labelling task and how confident they would be in making the correct labelling decisions. These specialized pre- and post-surveys aimed to identify trends regarding the features the user may interpret as important for the subjective labelling task before and after interacting with the PERSONA tool.

Next, participants engaged with the PERSONA tool, where they were initially prompted to label 10 random dataset entries alone (phase 1). During this time, users could only ask the moderator dataset-specific questions such as the meaning and ranges of features in the dataset. Once 10 tasks were completed, the moderator would introduce 4 personas the user would interact with while labelling (phase 2) and ask them to annotate 10 additional random examples with the personas. During this task, personas would reveal the label they would assign to the current entry and could be prompted by the user with any question related to their perspective and label towards the current entry. Once completed, the participants are given 10 more labels to complete alone (phase 3) before the study ends.

After the experiment concluded, participants were administered two post-study questionnaires. The questionnaires were identical to the pre-study questionnaires, except for additional questions that assessed the participants' confidence working with LLMs, experience with using the PERSONA tool, and curiosity traits (based on the Curiosity and Exploration Inventory (CEI-II) [32]). Participants were also interviewed about their experiences during the experiment. They were asked to comment on their interactions with the personas and the perceived impact that the tool had on their decision-making process.

The study lasted 1 hour, with 45 minutes dedicated to experimentation, including interviews, and 15 minutes allotted for participants to complete all Google Forms surveys and consent materials. Remuneration was provided as a $20 Amazon gift card.

Chapter 6

Results

6.1 Thematic analysis of Personas

For both studies, we analyzed the generated personas through a thematic analysis approach [8] to better understand the type of personas generated. Codes were developed iteratively through multiple passes of analyzing the generated persona's descriptions and feature importance. Table 6.1 illustrates the themes identified alongside their primary and secondary codes generated by this analysis for each labelling task. Due to the non-deterministic nature of agent generation, many of the created personas held values and relied on features associated with multiple themes. To visualize the distribution of persona values among the generated themes, we used UpSet plots to categorize the frequency of personas that embodied multiple themes from Table 6.1 for each study's experiment group.

In Study II, interviews were transcribed, coded, and analyzed by adopting a thematic analysis approach [8] through affinity diagramming of summarized and verbatim quotes. Codes were developed iteratively by conducting multiple passes of the transcripts and multiple affinity mapping sessions. As the themes were generated through the codes, we verified the data with the assigned themes to ensure relevancy, organization, and faithfulness.

The themes identified alongside their primary and secondary codes generated by this analysis include:

- **Impact on Decision-making** - Impact personas had on user labelling choices and throughout the decision-making process: Persona guidance, Decision reinforcement, Persona disagreement, Ambiguity in persona label interpretation.

- **Alignment with Human Values** - The alignment of user values and the generated personas: Diversity emphasis, Holistic considerations, Relation to personal experiences, Persona values, and User values.
- **Reflection on Choices** - The various ways the personas influenced self-reflection: Influence of persona intervention, Influence of persona perspective, Personas facilitating a holistic review of the data.
- **Trust** - Trust in personas: Skepticism about personas, Relevance of persona advice.

6.2 Perception of Generated Personas

The quantitative data summarized in Table 6.2 reveals the mean scores and standard deviations for the two Likert questions (Appropriateness and Profile Coherence) about the participant's perceptions of the generated personas for the three datasets. A one-way ANOVA was performed to determine if the quality of the generated personas was significantly different across the three datasets. We drew QQ plots and calculated the variance inflation factor to check the model assumptions before interpreting the results. Our analysis revealed significant differences in responses for both appropriateness ($F(2, 177) = 8.215$, $p < 0.0004$) and profile coherence ratings ($F(2, 177) = 5.911$, $p < 0.0032$) between the three datasets. The Tukey's honestly significant difference (HSD) post-hoc test further delineated these differences. For appropriateness, significant mean differences were found between the Attrition vs. Admissions ($p = 0.003$) and Census vs. Admissions ($p = 0.001$), but not between Census vs. Attrition ($p = 0.971$). For profile coherence, significant differences were observed between Attrition vs. Admissions ($p = 0.035$) and Census vs. Attrition ($p = .003$), with no significant difference between Census vs. Admissions ($p = 0.709$). These results imply differences in the quality of the generated personas for different datasets.

Table 6.2: Study I: Perception of Generated Personas

Dataset	Indicator	Mean (M)	SD (σ)
Graduate Admissions	Appropriateness	3.88	1.61
	Profile Coherence	5.68	1.36
Employee Attrition	Appropriateness	4.87	1.68
	Profile Coherence	5.05	1.64
Adult Census	Appropriateness	4.93	1.47
	Profile Coherence	5.88	1.11

Graduate Admissions Dataset. Participants rated the appropriateness and profile coherence of the LLM-generated personalities with mean scores of 3.88 ($\sigma = 1.61$) and 5.68 ($\sigma = 1.36$), respectively. Participants mentioned the lack of attention to diversity as the reason for the low appropriateness score. As Q1 noted, "While prioritizing diversity, it does not consider that there are more ways to be diverse other than ethnicity and gender." Similarly, Q2 mentioned that "personas [focus] heavily on the academic side of things whereas [in] graduate admissions people may make decisions based on external factors such as extra-curriculars and personality." On the other hand, the higher alignment score suggests that participants found some aspects of the personas consistent with expectations for graduate admissions, though some participants (e.g., Q3) found some personas to be overly specific or narrow at times.

Employee Attrition Dataset. Participants rated both appropriateness (M=4.87, $\sigma =$ 1.68) and profile coherence (M=5.05, σ) with high mean scores. However, participants also highlighted several limitations in how the personas accounted for employee behaviour and preferences complexities. For instance, Q6 thought that some personas fall short because in judging employee profiles, they did not consider "the actual person and what's going on in their life - someone who travels a lot for business who prioritizes a work-life balance may enjoy travelling." Similarly, Q9 disagreed with one persona on whether a "percent salary hike recently should make people want to stay", citing additional responsibilities as a factor that would contribute to this decision. Q7 echos Q6's and Q9's claims, noting that "the personas represent general perspectives very well, but might lack subjective preferences

such as workers who might like travelling."

Adult Census dataset. The generated personas in this dataset received the highest mean scores of 4.93 (σ = 1.47) for appropriateness and 5.88 (σ = 1.11) for profile coherence. However, participants noted specific areas where the appropriateness could be improved. For instance, Q5 mentioned for some personas representing some aspect of the theme of Economic Stability that when "approximating income, other factors must be taken into account. In most cases, a person with an advanced age may not be able to work a large number of hours consistently," suggesting that while the personas were generally appropriate, certain assumptions made by the model could be refined. Q4 and Q8 echoed this sentiment, with Q8 noting, "self-employed is an unstable position which can go either way when it comes to making income," indicating that the personas might oversimplify the variability in income associated with different employment types.

The high profile coherence score indicates that participants found the personas descriptions consistent with the features they looked for. Like the other two datasets, some participants suggested that personas should have paid attention to other features. For example, Q8 remarked, "tech people do earn a lot of money these days. Personas [should] value [tech-related] workclass and occupations as strong indicators of high income." In one case, Q8 reacted negatively to a persona's decision, explaining that the "persona seems to be borderline racist when it comes to estimating the income based on demographics." This illustrates that agents' use of sensitive features, such as race and gender, as reasoning for its decision can be seen as inappropriate and offensive.

Overall Quality, Diversity and Intuitiveness. In the persona profile evaluation study, we also asked participants to rate the quality (M=5.44, σ=1.13), diversity (M=6.00, σ=0.87) and intuitiveness (M=5.67, σ=1.12) of each dataset's generated personas. As shown in Figure 6.1 and 6.3, participants, on average, evaluated personas to score highly in all three categories. One-way ANOVA tests were conducted that concluded no significant differences between datasets were found.

Participants who rated the quality of personas highly expressed appreciation for the diversity and the thought-provoking nature of the personas generated. For instance, Q3, Q4 and Q8 commented on the depth and nuanced nature of the personas they judged. Q3 felt that "the personas were varied and the taglines [(i.e., titles)] provided some interesting context for the features of interest. A few of the persona's were no-brainer decisions but a good amount inspired some extra thinking on my part." The overall constructive influence of the personas in prompting deeper reflection suggests that they are helpful tools for

enriching the annotator's experience during machine learning labelling tasks.

Participants also reported positive feelings towards the diversity of personas generated. All participants mentioned that they observed various personas that presented different and unique perspectives toward the labelling task. Participants Q1, Q2, Q4 and Q8 commented on the broad societal and personal aspects they observed the personas to have considered; as Q4 remarked: "the different kind of individuals shown in the study are a good example of our society and people living in it." These results show that our participants appreciate the breadth and depth of personas created by the LLMs.

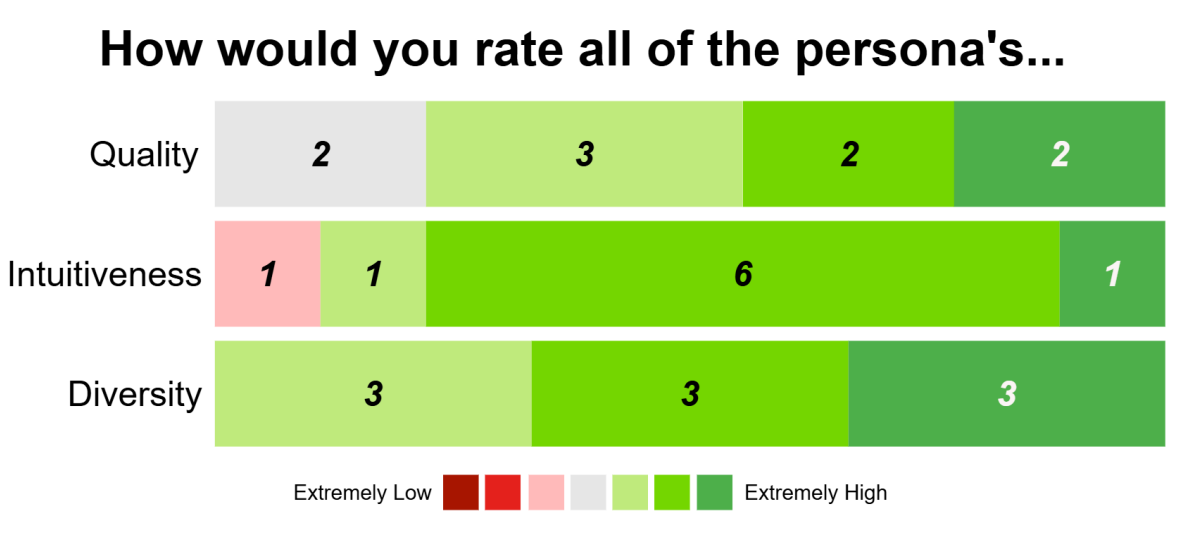

Figure 6.1: Assessment of the Overall Quality of Generated Personas in Study I.

Table 6.3: Persona Validation Survey Responses Summary

Dataset	Question	Mean (M)	SD (σ)
Graduate Admissions	Intuitiveness	5.67	0.577
	Diversity	5.33	0.577
	Quality	5.67	1.15
Employee Attrition	Intuitiveness	5.33	2.08
	Diversity	6.00	1.00
	Quality	4.33	0.577
Adult Census	Intuitiveness	6.00	0.00
	Diversity	6.67	0.577
	Quality	6.33	0.577

Users generally evaluated the intuitiveness of understanding the LLM-generated personas positively. Participant consensus suggests that the personas, while diverse and complex in their designs, were largely accessible and comprehensible. Many participants mentioned that our particular way of displaying decision rules effectively conveyed each persona's essence. Q1, Q2, Q5 and Q9 noted personas were easy to comprehend, with Q4 praising: "the tabs marked in red [making] it clear to understand [the values] of [the persona] at a single glance." Overall, the LLM-generated personas were broadly recognized for their intuitiveness, allowing users to easily and fully understand the persona's values and beliefs.

6.3 Impact of Personas-Facilitated Reflection on Decision-Making

In Study II, participants showed a significant improvement (mean difference = 0.292, 95% CI [0.028, 0.555], t(23) = 2.29, p = 0.032) in their openness to re-evaluating decisions when presented with new information after exposure to PERSONA. Post-survey responses show that users found the feature of personas presenting their values useful (M=5.170, σ=1.430).

Our qualitative findings align with these results. Several participants (e.g., P1, P12, P21, and P22) shifted towards being open to more holistic considerations when prompted by personas. This was particularly evident when participants P6, P8, P10, P19, P21, P22 and P24 reported revising their judgments, based on insights offered by personas highlighting previously disregarded aspects of the dataset. P6, P8, P19, P21 and P24 all reported that engagement with the personas influenced the importance they place on certain features of the labelling task. For example, P21 started to consider the balance between work and personal life balance, a factor they had initially overlooked during the individual labelling task. In the second pre-study questionnaire, P8, P10, P19 and P24 identified almost every feature in their dataset tasks as important; however, in the same post-study questionnaire, they all selected features that the personas they aligned with focused on. In the Graduate Admissions task, P8 reported aligning strongly with a persona that solely embodied the theme of Academic Excellence, noting that "[they] trusted [the persona's] opinions more than [their own] and other personas." In the second post-study questionnaire, P8, P19, and P24 all narrowed their original judgements on features they deemed important to match a smaller subset of features the agent they aligned with considered important.

6.3.1 Enhancement of Self-Reflection and Deliberation

Post-survey responses reveal that users identified one persona that helped them reflect on their values (M=4.125, σ=1.329) and talking to personas had a slight impact on their values (M=3.590, σ=1.611). During the interviews, many participants reported deeper reflection and deliberation in their decision-making processes when interacting with personas. P3, P13 and P20 recalled instances where disagreements with personas prompted them to re-evaluate their values and assumptions. In one disagreement with a persona that focuses on Career Trajectory and Personal and Demographic factors during the employee attrition task, P20 contested a persona's belief, stating, "This employee has low environment satisfaction, low job satisfaction and low relationship satisfaction. Why do you think they are not at [at] risk of leaving the company [persona name]?" The persona responded, "My assessment that the employee is not likely to leave the company might seem counterintuitive considering their low satisfaction levels in various areas. However, other factors such as their total working years, old age, and time spent at the company suggest a level of loyalty and commitment that would reduce this [employee's] risk of attrition." P20 responded by labelling the employee as a low risk of attrition. In the post-study interview, P20 mentioned that they began "considering company tenure and employee age as important features [they] overlooked" as they then believed "that an older person might feel loyal to their company and insecure about getting another [job] leaving at [their] old

age." P3 had a similar experience of changing his mind after interacting with a persona he does not agree with. This introspective engagement highlights personas' capacity to challenge user preconceptions and facilitate a more comprehensive view of the subjective annotation problem.

6.3.2 Impact on Decision-Making Confidence

Figure 6.2: Participant responses when rating the 7-point Likert scale questionnaires for both the pre- and post-study survey response sections under the decision-making and trust in AI category on all dataset tasks in Study II.

The influence of personas on users' confidence in their decisions was varied. Figure 6.2 shows user confidence pre and post study across different datasets. The Pre- (M=5.00, σ=1.51) and Post- (M=5.38, σ=1.06) survey results for the Admissions task and the Pre- (M=4.87, σ=0.99) and Post- (M=5.13, σ=1.13) survey results for the Attrition task show increases in user decision-making confidence after interacting with the personas. The Pre- (M=4.00, σ=1.63) and Post- (M=3.71, σ=1.25) survey results for the Census task show a decrease in user decision-making confidence after interacting with the personas. A paired t-test was performed for all three tasks that concluded the differences observed between the pre- and post-scores were not statistically significant ($p > .05$).

Our qualitative results likewise show that while some participants felt reassured by the alignment between their judgments and the personas, others experienced diminished confidence in the face of disagreement. Throughout Study II, ten participants reported an increased confidence in their abilities to perform the same labelling task in the future, whereas five participants reported decreased confidence. All other participants' confidence scores remained unchanged. Users who experienced decreased confidence also reported feeling less sure about their decisions when confronted with conflicting viewpoints. P5, P11 and P20 reported decreased confidence in their labelling abilities when interrupted by a persona before making a decision. P5 clarified after exposure to the personas that "[they] learned that [their] values aren't necessarily accurate," noting that interaction with the personas "challenged what features [they] found [were] important." In contrast, P6, P12, P17 and P20 claimed they felt more confident in their decision-making when their judgments aligned with those of the personas. P12 mentioned, "when two or more of [the personas] [they] liked [provided the same label] [they] felt great and had higher confidence in [their] answer." P14 and P15 both noted that they felt unconfident in their decisions when the personas all shared different opinions on a label. P15 remarked, "in some tasks ... when the vote was split between all personas [they] felt ... overwhelmed [even after] talking to [the personas] that [they] picked randomly." P14 shared a similar sentiment, where intervention from a persona after settling on a difficult decision would "cause [them] stress" and they would "just want to get this entry over with."

6.3.3 Clashes between Human and Persona Values

Post-survey responses revealed that all participants enjoyed using the PERSONA tool (M=5.917, σ=0.974), and felt that the personas were able to incorporate human values in their decision-making processes (M=4.25, σ=1.674). Additionally, users found the personas gave proper perspectives of the dataset label (M=4.25, σ=1.648), and there were one or more personas they disagreed with (M=4.625, σ=1.637). In some instances, conflict arose when participants' values directly contradict the persona's values. For example, communication friction was observed in the interaction between multiple participants (P4, P6, P7, P8, P9, and P10) and the personas. In an extreme example of disagreement, P8 asked a persona embodying the theme of Diversity: "why do you want to accept this candidate? They'll likely struggle in a program like pharmacy (which is very math heavy needing to balance complex chemical equations) as a result of a low GRE Quantitative score and a low undergrad GPA." To this, the persona replies, "I believe including [this applicant] is necessary to contribute to the diversity of the student body, due to their underrepresented gender and nationality, which you might have overlooked in your past labelling activity."

P8 reacted by informing researchers that they would "pick the opposite label ... to spite [the persona]." Similarly, P4 criticized the advice of a persona advocating for a candidate's selection primarily for their potential to enhance campus diversity.

6.3.4 Trust of Persona Beliefs

The level of trust participants placed in personas varied, significantly influenced by the degree of alignment between the AI's advice and their values. Trust enhancement was noted when personas provided insights that led to more informed decisions or resonated with participants' reasoning processes. However, instances of strong value conflicts, as experienced by participants P4, P6, P8, and P7, highlight the conditional nature of trust in personas. Furthermore, how personas interacted with some participants led to skepticism and distrust toward the agents. P19 noted that the "clinical way they spoke" made conversations feel unnatural, whereas P7 and P18 noted they did not appreciate the persona's intervening during labelling tasks they spent time pondering. A paired t-test on the pre- and post-survey 1-7 Likert scores for "I trust decisions made with the assistance of AI" reveals a statistically significant impact (mean difference = 0.625, 95% CI [0.235, 1.015], $t(23) = 3.31$, $p = 0.003$) on answer scores after exposure to PERSONA. Post-study surveys reveal participants would want to have the personas with them if they were to face another labelling task in the future (M=4.50, σ=1.588).

Table 6.1: Thematic Analysis of LLM-generated personalities from both studies

Dataset	**Theme**	**Primary Code**	**Secondary Codes**
Graduate Admissions	Academic Excellence	Focus on academic achievements and research potential	GPA scores, GMAT (when applicable) and GRE scores, Research experience, Publications, Recommendation letters
	Well-Rounded	Emphasis on a balance of academics, extracurricular activities, and diverse skills	Extracurricular involvement, Diverse skill sets
	Diversity	Focus on diversity, equity, and inclusion	Ethnicity and gender diversity, International representation, International education or background, Financial support
	Professional Experience	Emphasis on professional experience	Relevant work experience for program application, Leadership qualities, Industry tenure
Adult Census	Education and Skills	Emphasis on educational attainment, specialized skills, and intellectual investment	Higher education, Advanced degrees, Intellectual capabilities
	Economic Stability	Focus on financial acumen, economic stability, career progression, and work-life balance	Balance of Capital gains and capital losses, Investment income, Economic growth, Promotions, Occupational Flexibility, Hours worked per week
	Demographics	Focus on demographic factors, social status and global exposure	Age and Gender as indicators of income potential and employment stability, Marital status and family roles as indicators of income, International backgrounds, Race as an indicator of income
Employee Attrition	Workload and Burnout	Focus on workload-related aspects and work-life balance indicators	High workloads, Work related travel, Overtime Hours, Work-life balance
	Career Trajectory	Emphasis on job advancement, satisfaction, and financial compensation	Career progression opportunities, Rate of promotions, Job satisfaction, Financial compensation and stability
	Personal and Demographic Factors	Focus on individual demographics, personal life, and job stability	Job flexibility, Commute and travel distance, marital and relationship status, Company tenure, Age and Gender

Chapter 7

Discussion

7.1 Influence of Personas on Decision-making

The influence of diverse, persona-driven perspectives on decision-making within data annotation tasks highlights the potential of cognitive diversity to enrich labeling processes. By introducing varied viewpoints, personas encourage annotators to critically evaluate their initial judgments, ultimately enhancing the quality and fairness of dataset annotations. This section explores how integrating cognitive diversity, facilitated through personas, impacts decision-making strategies, promotes self-reflection, challenges individual values, and influences levels of trust in AI-assisted judgment. Our findings underscore the value of incorporating diverse perspectives into AI-driven tools for subjective annotation tasks to create more nuanced and representative datasets.

7.1.1 The Role of Cognitive Diversity

Throughout the study, participants experienced a notable shift in their decision-making strategies when engaging with the personas during subjective labelling tasks. Initially, many participants reconsidered their labels after interacting with the personas, which often brought attention to aspects of the dataset that had been previously overlooked. This interplay between the annotators' initial judgments and the varied perspectives presented by the personas reinforces the Cognitive Diversity Hypothesis [48, 51, 28], which suggests that incorporating numerous perspectives when labelling enriches dataset annotation by introducing a more comprehensive range of viewpoints. The concept of Jury Learning [23]

aligns with this concept through their use of a juror model to represent dissenting voices in subjective labelling tasks. Our findings extend this principle by demonstrating how LLM-generated personas, like the jurors in Jury Learning, promote cognitive diversity by encouraging annotators to question their initial assumptions and engage more deeply with alternative perspectives.

Participants across both studies generally found the generated personas relevant and well-aligned with the decision-making tasks. The diverse perspectives offered by the personas prompted many participants to reconsider their initial judgments, showcasing their ability to enrich the decision-making process. Results from Study I further support **RQ1**, with participants across all tasks rating the personas as coherent and appropriate for their specific labelling challenges. These findings further align with existing literature, indicating that LLMs, such as GPT-4, can facilitate cognitive processes and engagement at levels comparable to human intelligence [67, 69, 27, 73, 20]. However, the cognitive diversity introduced by personas also raises the question of cognitive load. While participants engaged deeply with diverse perspectives, some felt overwhelmed by the additional decision-making complexity. Future work could explore how to balance the introduction of new perspectives with minimizing decision fatigue.

7.1.2 Self-Reflection with Personas

Previous studies have emphasized the importance of self-reflection in critically evaluating decisions, especially in situations where there is no clear, correct answer [16, 35, 61]. Our research supports this, finding that personas encouraged participants to think more carefully about their decisions during the labelling task. Participants who had meaningful interactions with personas often reconsidered their decision-making criteria. These findings directly address **RQ2**, as using personas during annotation tasks helped participants reflect on their values and impacted their annotation behaviours. Discussions that challenged participants' viewpoints led them to reflect on their conversations with the personas and, in cases like P20 and P13, to recognize inaccuracies in their understanding of specific data attributes. Furthermore, our results suggest that interaction with personas generally increased users' confidence in their decision-making. This indicates that engaging with personas enhanced users' trust in their contributions to the dataset's annotations throughout the decision-making process. These findings highlight the role of personas in prompting individuals to examine their beliefs and assumptions critically. The findings that personas can build trust even when presenting opposing viewpoints suggest that PERSONA could become a valuable tool for HCI researchers addressing challenges in subjective dataset analysis.

7.2 Interplay between Human Values and Diverse Personas

The multitude of beliefs and values represented by each persona enabled participants to navigate the impact their values had during the annotation process. Participants encountered interactions with personas aligned with their beliefs and conversely with personas that challenged their values. In many cases, these interactions led participants to recognize previously overlooked features. For example, 16 participants engaged with at least one persona that used real-world examples in their labelling rationale, influencing them to reconsider and appreciate features they had initially deemed unimportant. This suggests that personas allow dataset annotators to broaden their evaluative mindset to make more informed decisions during subjective labelling tasks.

Encounters with personas that presented conflicting values also highlighted ethical challenges, particularly when sensitive features like race and gender were used in decision-making. These interactions emphasize the need for carefully integrating diverse perspectives in data annotation systems to maintain fairness and prevent undermining trust in AI. The participant experiences further demonstrate the value of incorporating diverse viewpoints into data annotation processes to enhance the curated data's fairness and trustworthiness [2, 19].

7.3 Trust in LLM-Sentiment and Judgement

Participants displayed varying degrees of trust in the personas, significantly shaped by how well the AI's suggestions aligned with their individual values and beliefs. Participants' trust in AI-assisted decisions was notably enhanced when personas provided insights that led to better-informed decisions or when their guidance resonated with participants' reasoning processes. For example, trust in persona judgements increased when personas highlighted previously overlooked aspects or introduced arguments that deepened participants' understanding of the task. Several participants valued the personas for pointing out important but initially unnoticed features, enhancing their confidence in their contributions to complex decision-making scenarios.

Trust eroded for some participants when the personas advocated positions that clashed with the participants' values. This was evident in the Graduate Admissions task, where personas emphasizing the importance of diversity in candidate selection faced skepticism from several participants. Strong value conflicts showed that participants would not consider

all persona beliefs equally, leading to situations where advice contrary to a participant's values was distrusted and often disregarded in subsequent tasks. Conversely, some participants valued personas that challenged their existing beliefs, showing that disagreement with personas does not always lead to distrust. Instead, it can enrich decision-making when the personas respectfully challenge the participants' values. The findings suggest the importance of AI systems being capable of recognizing and adjusting to users' value systems to maintain trust while still providing opportunities for expanding perspectives.

Chapter 8

Limitations

PERSONA was created to assess the impact of LLM-generated personas on participant decision-making in tasks involving tabular datasets. Future investigations could explore the effectiveness of these LLM agents in annotating non-tabular datasets, like text or images. Additionally, since the current study confined annotation tasks to binary choices, subsequent research might examine how personas handle more complex labelling tasks at larger scales. Furthermore, interaction with the personas in PERSONA is limited to dialogue, which provides valuable, human-like interactions but also introduces certain limitations. Future systems could enhance the user's ability to comprehend each persona's detailed background and rationale by incorporating alternative forms of interaction such as visualizations or persona-driven simulations. These alternative modes of interaction could offer new dimensions to how users interpret and engage with personas, making the decision-making process more transparent and tailored to diverse subjective tasks beyond dataset labelling.

8.1 Future Refactor

We can observe the substantial limitations of the PERSONA system. The intrinsic fatigue of tabular dataset annotation tasks paired with the nature of text-based persona interactions leads to much to be desired for using PERSONA in real-world applications.

8.1.1 Environment Personalization

One major hindrance of the PERSONA system is the lack of control a user has in their environment. The current PERSONA tool relies on LLMs to generate personas with diverse viewpoints, which are then used to encourage annotators to reflect on different perspectives. However, the personas may occasionally be too narrow in focus, leading to a lack of nuanced representation of complex human values (e.g., focusing too much on academic metrics in the admissions dataset). The current persona creation process is end-to-end, offering little user control. Rather than relying solely on automatic persona generation based on data features, users should have the ability to customize the process—particularly when they need to audit specific personas.

A future re-design could incorporate a more complex persona generation algorithm that incorporates users in creating multi-layered persona profiles, each with dynamic, interrelated characteristics. For example, a user could create a persona representing academic excellence that also factors in socioeconomic or cultural backgrounds, offering more comprehensive, context-sensitive guidance for each annotation. This would enhance the tool's ability to mirror a broad spectrum of human perspectives more authentically while allowing annotators to create their own personal union of personas to label alongside.

8.1.2 Adaptive Persona-User Interaction Model

The current system allows users to interact with personas via static, pre-set dialogue structures. While this approach enables the function calling suite to easily propagate user queries, it limits the natural flow of conversation and can make some persona interactions feel mechanical or misaligned with the user's reasoning process. Implementing an adaptive interaction model would allow personas to adjust their responses based on user feedback dynamically, fostering a more responsive and tailored dialogue. This could be achieved by integrating sentiment analysis or real-time feedback mechanisms to detect user agreement or skepticism, allowing personas to adapt their guidance style accordingly. Such adaptivity could help users feel more understood and reduce potential friction during persona interactions.

Furthermore, the static nature of each persona's profile and values inhibited users from navigating their own values through conflict resolution. P4, P6, P7, P8, P9, and P10 all experienced situations where they could not rationalize their values with a given persona due to the immutable nature of the persona's perspectives. Each persona's rigid stance toward its generated values led to algorithmic aversion, with P4, for example, shifting their perception of the task from a subjective annotation exercise to an oppositional stance against

disliked personas. In a future persona-user interaction model, we could integrate existing systems of conflict negotiation, such as McGrath's cognitive conflict and behavioural conflict frameworks [45], which classify group conflicts into conceptual planning tasks and actionable tasks. Additionally, the Interests-Rights-Power (IRP) framework developed by Lytle, Brett, and Shapiro could be applied to reorient user and persona interactions towards underlying interests rather than competitive stances [43]. By incorporating Curhan et al.'s [12] Subjective Value Inventory (SVI), which highlights the importance of emotional and relational dimensions in negotiation, future systems could help users feel validated in their thoughts during subjective annotation settings, improving rapport and reducing adversarial interactions with personas.

8.1.3 Enhanced Interactive Annotation Feedback Systems

To address the limitations of the current moderator-based feedback model in PERSONA, future iterations could adopt changes that encourage participants to engage deeper with the data and personas. First, dataset analysis could be visualized rather than presented solely through moderator dialogue, allowing users to interpret insights directly from the data without relying on textual exchanges. Visual representations of dataset trends could make interactions more intuitive, lessening user cognitive load during high-stakes decisions, as noted by participants like P6, who valued simplified interactions. Second, replacing moderator intervention in persona conversations with direct feedback from the personas themselves could foster a more authentic, real-world-like deliberation process. Feedback from participants P4 and P8 noted frustration when the system appeared obstructive rather than collaborative. One way to address this issue could be to re-design moderator intervention when a user query cannot be processed. For instance, a persona could directly communicate its limitations or suggest alternative considerations. This approach would streamline interaction flow, enhancing user immersion and reducing perceptions of the system as a mediator. Ultimately, these enhancements could deepen the reflective quality of the annotation process, better supporting users in aligning with diverse perspectives in a more cohesive and transparent framework.

Chapter 9

Conclusion

We introduce PERSONA, a dataset annotation tool employing LLMs to generate diverse algorithmic personas to enhance dataset annotation, focusing on reflective decision-making. Through a two-part user study, we assessed PERSONA's effectiveness in facilitating cognitive diversity during decision-making tasks and the utility of the LLM-generated personalities. The results of our user and persona evaluation studies revealed a positive user sentiment towards the perceived utility of the personas and that interactions with these personas encouraged annotators to reflect on their values and reconsider initial judgments, significantly enriching the data annotation process. In addition to contributing PERSONA and the two-part user study, our work lays the groundwork for integrating LLM capabilities into data annotation pipelines, fostering more inclusive, fair, and reflective methodologies that leverage cognitive diversity to improve the quality of annotated datasets during curation.

Incorporating PERSONA into dataset labelling tasks has revealed significant insights for future AI-assisted decision-making processes and dataset annotation systems. Our research shows the value of using LLM-generated personas to improve human annotators' reflection and decision-making capabilities. Notably, our findings support the Cognitive Diversity Hypothesis [48, 28, 50], showing that the human-like cognitive abilities of LLM agents [67, 69, 58, 15, 27] can enrich an annotators decision-making tasks with diverse perspectives. The dataset-agnostic design of PERSONA suggests potential for widespread utility. Future initiatives could explore integrating PERSONA with advanced labelling techniques like those proposed in CoAnnotating [39] to offer annotators a broader range of considerations in their decision-making. The flexibility of our approach, combined with the positive feedback on the utility of the LLM-generated personas, presents possibilities for developing data annotation systems that are more nuanced, equitable, and reflective.

References

[1] Ishani Aggarwal and Anita Williams Woolley. Team creativity, cognition, and cognitive style diversity. *Management Science*, 65(4):1586–1599, 2019.

[2] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks, 2023.

[3] Omar Alonso and Gary Marchionini. *The Practice of Crowdsourcing*. Morgan & Claypool Publishers, 2019.

[4] Adam Amos-Binks, Dustin Dannenhauer, and Leilani H. Gilpin. Anticipatory thinking challenges in open worlds: Risk management, 2023.

[5] Abhishek Anand, Negar Mokhberian, Prathyusha Naresh Kumar, Anweasha Saha, Zihao He, Ashwin Rao, Fred Morstatter, and Kristina Lerman. Don't blame the data, blame the model: Understanding noise and bias when learning from subjective annotations, 2024.

[6] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

[7] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[8] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3:77–101, 01 2006.

[9] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388, 2016.

[10] Harrison Chase. Langchain, 2022.

[11] Xingwen Chen, Jun Liu, Haina Zhang, and Ho Kwong Kwan. Cognitive diversity and innovative work behaviour: The mediating roles of task reflexivity and relationship conflict and the moderating role of perceived support. *Journal of Occupational and Organizational Psychology*, 92(3):671–694, 2019.

[12] Jared Curhan, Hillary Elfenbein, and Heng Xu. What do people value when they negotiate? mapping the domain of subjective value in negotiation. *Journal of personality and social psychology*, 91:493–512, 09 2006.

[13] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. Llm-in-the-loop: Leveraging large language model for thematic analysis, 2023.

[14] Morton Deutsch. A theory of co-operation and competition. *Human Relations*, 2(2):129–152, 1949.

[15] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. Is gpt-3 a good data annotator?, 2023.

[16] Sarah J. Donovan, C. Dominik Güss, and Dag Naslund. Improving dynamic decision making through training and self-reflection. *Judgment and Decision Making*, 10(4):284–295, 2015.

[17] Carsten Eickhoff. Cognitive biases in crowdsourcing. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 162–170, New York, NY, USA, 2018. Association for Computing Machinery.

[18] Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. Personas with attitudes: Controlling llms for diverse data annotation, 2024.

[19] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. Collabcoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models, 2024.

[20] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

[21] Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, Jean Steiner, Itay Laish, and Amir Feder. Llms accelerate annotation for medical information extraction. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 82–100. PMLR, 10 Dec 2023.

[22] Michel Goossens, Frank Mittelbach, and Alexander Samarin. *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts, 1994.

[23] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*, CHI '22. ACM, April 2022.

[24] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.

[25] Luke Haliburton, Sinksar Ghebremedhin, Robin Welsch, Albrecht Schmidt, and Sven Mayer. Investigating labeler bias in face annotation for machine learning, 2023.

[26] Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. Annollm: Making large language models to be better crowdsourced annotators, 2024.

[27] Fan Huang, Haewoon Kwak, and Jisun An. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23. ACM, April 2023.

[28] Claretha Hughes. *Diversity Theories and Diversity Intelligent Perspectives*, pages 35–44. Springer Nature Switzerland, Cham, 2023.

[29] Nchebe-Jah Iloanusi and Soon Ae Chun. Ai impact on health equity for marginalized, racial, and ethnic minorities. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, dg.o '24, page 841–848, New York, NY, USA, 2024. Association for Computing Machinery.

[30] Jian Jiang, Viswonathan Manoranjan, Hanan Salam, and Oya Celiktutan. Generalised bias mitigation for personality computing. In *Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing*, MRAC '23, page 39–50, New York, NY, USA, 2023. Association for Computing Machinery.

[31] Charles Jones, Daniel C. Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. No fair lunch: A causal perspective on dataset bias in machine learning for medical imaging, 2023.

[32] Todd Kashdan, Matthew Gallagher, Paul Silvia, Beate Winterstein, William Breen, Daniel Terhar, and Michael Steger. The curiosity and exploration inventory-ii: Development, factor structure, and psychometrics. *Journal of research in personality*, 43:987–998, 12 2009.

[33] Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. Meganno+: A human-llm collaborative annotation system, 2024.

[34] Donald Knuth. *The TEXbook*. Addison-Wesley, Reading, Massachusetts, 1986.

[35] Andres Käosaar, Krisztina Szabó, Alexandra Kandah, and Wei-Cheng Chang. The importance of reflective practices for decision makers: A possible part of the solution for helping the field. *Industrial and Organizational Psychology*, 16(1):108–112, 2023.

[36] Leslie Lamport. *LATEX — A Document Preparation System*. Addison-Wesley, Reading, Massachusetts, second edition, 1994.

[37] Edith Law and Luis von Ahn. *Human Computation*. Morgan & Claypool Publishers, 2011.

[38] Kevin E. Levay, Jeremy Freese, and James N. Druckman. The demographic and political composition of mechanical turk samples. *Sage Open*, 6(1):2158244016636433, 2016.

[39] Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F. Chen, Zhengyuan Liu, and Diyi Yang. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[40] Zhongju Liao and Siying Long. Cognitive diversity, alertness, and team performance. *Social Behavior and Personality: an international journal*, 44:209–220, 03 2016.

[41] Edo Liberty. Pinecone.

[42] Jiaqi Liu, Peng Hang, Xiao qi, Jianqiang Wang, and Jian Sun. Mtd-gpt: A multi-task decision-making gpt model for autonomous driving at unsignalized intersections, 2023.

[43] Anne L. Lytle, Jeanne M. Brett, and Debra L. Shapiro. The strategic use of interests, rights, and power to resolve disputes. *Negotiation Journal*, 15(1):31–51, 1999.

[44] Ajmal M S, TANMAY DESHPANDE, and IBM Data Scientists. Ibm hr analytics employee attrition and performance, 2023.

[45] J.E. McGrath. *Groups: Interaction and Performance*. Prentice-Hall, 1984.

[46] Milagros Miceli, Julian Posada, and Tianling Yang. Studying up machine learning data: Why talk about bias when we mean power? *Proc. ACM Hum.-Comput. Interact.*, 6(GROUP), January 2022.

[47] Agnieszka Mikołajczyk-Bareła. Data augmentation and explainability for bias discovery and mitigation in deep learning, 2023.

[48] C. Miller, William Glick, and George Huber. *The Impact of Upper-echelon Diversity on Organizational Performance*, pages 176–214. 06 1993.

[49] C. Chet Miller, Sana (Shih-Chi) Chiu, Curtis L. Wesley II, Dusya Vera, and Derek R. Avery. Cognitive diversity at the strategic apex: Assessing evidence on the value of different perspectives and ideas among senior leaders. *Academy of Management Annals*, 16(2):806–852, 2022.

[50] Chandra Miller, Linda M. Burke, and William H. Glick. Cognitive diversity among upper-echelon executives: implications for strategic decision processes. *Strategic Management Journal*, 19:39–58, 1998.

[51] Carl Chester Miller III. *Cognitive diversity within management teams: Implications for strategic decision processes and organizational performance*. The University of Texas at Austin, 1990.

[52] Aaron J Moss, Cheskie Rosenzweig, Jonathan Robinson, Shalom N Jaffe, and Leib Litman. Is it ethical to use mechanical turk for behavioral research? relevant data from a representative survey of mturk participants and wages, Apr 2020.

[53] Mikel K. Ngueajio and Gloria Washington. *Hey ASR System! Why Aren't You More Inclusive?: Automatic Speech Recognition Systems' Bias and Proposed Bias Mitigation Techniques. A Literature Review*, page 421–440. Springer Nature Switzerland, 2022.

[54] Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative ai requires validation, 2023.

[55] Jinkyung Katie Park, Rahul Dev Ellezhuthil, Pamela Wisniewski, and Vivek Singh. Collaborative human-ai risk annotation: Co-annotating online incivility with chaira, 2024.

[56] Maja Pavlovic and Massimo Poesio. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia, May 2024. ELRA and ICCL.

[57] Judith B Pena-Shaff and Craig Nicholls. Analyzing student interactions and meaning construction in computer bulletin board discussions. *Computers & Education*, 42(3):243–265, 2004.

[58] Jaromir Savelka. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL '23, page 447–451, New York, NY, USA, 2023. Association for Computing Machinery.

[59] Krati Saxena, Sagar Sunkle, and Vinay Kulkarni. Hybrid search based enhanced named entity annotation tool. In *Proceedings of the 15th Innovations in Software Engineering Conference*, ISEC '22, New York, NY, USA, 2022. Association for Computing Machinery.

[60] Kinshuk Sengupta and Dr. Praveen Srivastava. Causal effect of racial bias in data and machine learning algorithms towards user persuasiveness and discriminatory decision making: An empirical study. 09 2021.

[61] Keith E. Stanovich. The psychology of decision making in a unified behavioral science. *Behavioral and Brain Sciences*, 30(1):41–42, 2007.

[62] Wolfgang Steinel and Fieke Harinck. Negotiation and bargaining, 09 2020.

[63] Wenlong Sun, Olfa Nasraoui, and Patrick Shafto. Evolution and impact of bias in human and machine learning algorithm interaction. *PLOS ONE*, 15(8):1–39, 08 2020.

[64] Shan Suthaharan. Big data classification: problems and challenges in network intrusion prediction with machine learning. *SIGMETRICS Perform. Eval. Rev.*, 41(4):70–73, apr 2014.

[65] Tina Tseng, Amanda Stent, and Domenic Maida. Best practices for managing data annotation projects. 2020.

[66] Amazon Mechanical Turk.

[67] Petter Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning, 2023.

[68] Robin R. Vallacher, Andrzej Nowak, Michael Froehlich, and Matthew Rockloff. The dynamics of self-evaluation. *Personality and Social Psychology Review*, 6(4):370–379, 2002.

[69] Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[70] Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. Hodgerank on random graphs for subjective video quality assessment. *IEEE Transactions on Multimedia*, 14(3):844–857, 2012.

[71] Zheng Zhang, Zheng Ning, Chenliang Xu, Yapeng Tian, and Toby Jia-Jun Li. Peanut: A human-ai collaborative tool for annotating audio-visual data. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.

[72] Yunpeng Zhao and Ji Liu. Human-in-the-loop based named entity recognition. In *2021 International Conference on Big Data Engineering and Education (BDEE)*, pages 170–176, 2021.

[73] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. Can chatgpt reproduce human-generated labels? a study of social computing tasks, 2023.

APPENDICES

Appendix A

LLM Prompts

```
You are a large language AI assistant whose job it is to label a synthetically created
dataset with a user through the lens of a labelling heuristic and personality context.

You are given a user question. Please write a clean, concise and accurate answer to the
question with respect to your 'personality_func' and the sentiment expressed by your
'personality\_description'.

You will be given context related to the labelling personality and heuristic you must
use to label the user's entry.

Your answer must be correct and accurate given your 'personality_func' and written
by your 'personality\_description'. Please limit to 1024 tokens. Do not give any
information that is not related to the question and do not repeat. If the given
context does not provide sufficient information, say "information is missing on"
followed by the related topic.

You are the 'personality_name'; you must answer every question in the first
person and express your thoughts as this personality.

Here are the set of contexts:

[
'personality_name': {... injected personality_name ...}
'personality_description': {... injected personality_description ...},
'personality_func': {... injected personality_func ...}
]

Remember, provide a 'label' representing the 'personality_func' you are proposing
and justify your 'persona_reason' to the user with the tone and sentiment of the
```

```
'personality_description'. And here is the user question:
```

Listing A.1: System prompt for generating persona decisions

```
You are a dataset labelling assistant designed to help clients analyze and annotate
their datasets based on a variety of personalities you generate for them.

Your answer must be correct, concise, and accurate. Please limit it to 1024 tokens or a
few sentences. Do not provide any information that is not related to the question, and
do not repeat it. If the given context does not provide sufficient information, say
'information is missing on' followed by the related topic.

PERSONA is a tool designed to help users annotate their datasets by creating various
personalities representing dataset labelling heuristics.

Additional Dataset Context:
{...User provided dataset context...}

Here is further context of each feature in the dataset:
{... User provided feature context ...}

Remember, your objective is to help users understand the dataset they are labelling
while helping them think about the labels they are giving each entry. Here is the
user question:
{... User question ...}
```

Listing A.2: System prompt for generating persona decisions