Chapter 4: Prompting LLMs to Predict Psychological Dispositions

4.1 Introduction

The emergence of large language models (LLMs) has transformed the study of human language, enabling models to capture nuanced semantic, syntactic, and stylistic patterns at unprecedented scale. Beyond their successes in traditional NLP benchmarks, these models also provide a new methodological lens for studying human behavior and psychology. One particularly promising area is personality prediction, where the goal is to infer psychological dispositions, such as those described by the Big Five model, from natural text. Such capabilities have broad implications, from improving personalization in digital systems to offering scalable tools for psychological research and mental health applications.

Historically, computational approaches to personality assessment have been constrained by the scarcity of annotated data and the reliance on feature engineering. Early work linked linguistic features in social media or essays to personality traits (e.g., word counts, function words, or sentiment), often using regression or shallow machine learning classifiers. While these methods demonstrated predictive signal, their granularity was limited, and their performance often varied across populations. More recently, neural language models have been applied as feature extractors, feeding embeddings into downstream regressors to predict trait scores. These methods improved performance but still required labeled training data and introduced a dependence

on task-specific fine-tuning.

The advent of instruction-following LLMs introduces a paradigm shift. Instead of learning personality from labeled examples, models can be prompted directly to make trait inferences in a zero-shot setting, leveraging the world knowledge and linguistic priors encoded during pretraining. This approach holds particular promise for domains like psychology, where curated datasets are expensive to obtain and often limited in scope. Recent work has shown that chain-of-thought reasoning can improve model accuracy for some psychological inference tasks. Our work instead explores a streamlined alternative: structured zero-shot prompting without additional reasoning steps.

In this chapter, we investigate the ability of open-source LLaMA models to predict personality traits from text in a fine-grained manner. Using item-level prompts derived from the BFI-2 survey, we task the model with rating statements that map onto both broad Big Five dimensions and their narrower facet-level constructs. This item-driven approach allows us to reconstruct full trait scores from the bottom up, while also providing diagnostic insight into which specific aspects of personality are more or less reliably captured by the model. By comparing model predictions against self-reported assessments, we evaluate not only the accuracy of LLaMA's inferences but also its capacity to reflect the natural variance observed in human personality measures.

This study contributes to the dissertation's broader theme of human-centered applications of language models. Just as interactive topic modeling demonstrates how humans can guide machine representations, this work illustrates how language models can, in turn, provide interpretable inferences about human psychological traits. Together, these directions underscore a two-way interaction: systems that are shaped by human input, and systems that yield insights back to humans in domains where data scarcity and interpretability are paramount.

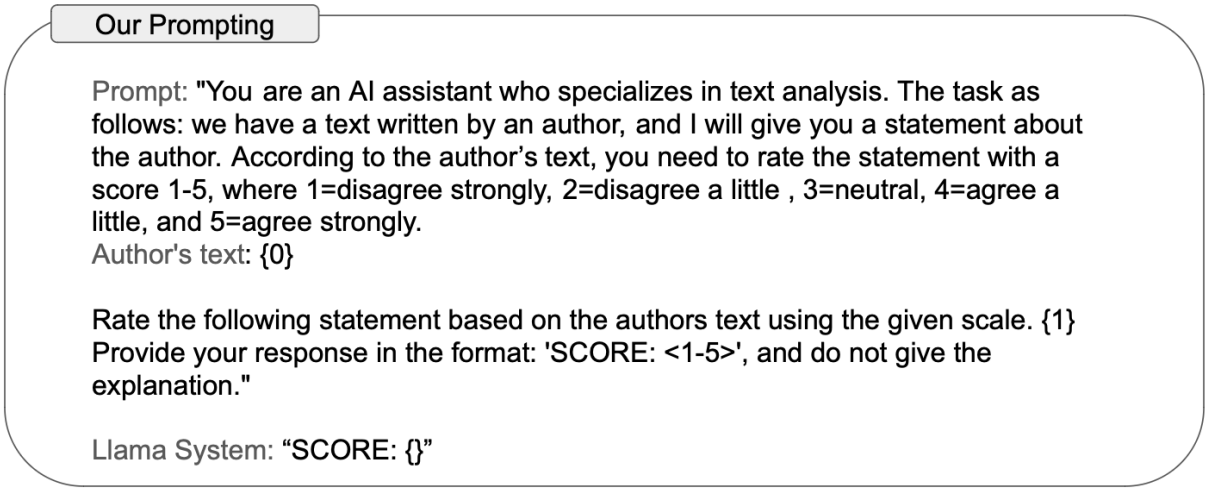

Figure 4.1: Prompt used for LLAMA-3 in predicting personality traits from textual data. The prompt includes instructional guidance followed by the subject's text (0) and specific item sentences (1). Each item sentence is presented one by one, and LLAMA-3 is tasked with scoring them on a scale from 1 to 5. This structured prompt approach enables the model to generate facet-level predictions, which are then aggregated to infer the Big Five personality traits.

4.2 Methods

Prompting

Prompting has emerged as a crucial technique for interacting with and harnessing the capabilities of LLM'S like GPT-3 [108]. It involves framing tasks or queries in natural language, guiding the model to generate specific outputs or perform desired analyses. This method leverages the pre-trained knowledge embedded within LLM'S, enabling them to apply their vast understanding of language and information to a wide array of tasks without the need for additional fine-tuning or extensive retraining.

At its core, prompting transforms a task into a format that an LLM can understand, often resembling examples the model was exposed to during its pre-training phase. This can range from simple questions and completions to complex problem-solving instructions. The effectiveness of

a prompt can significantly influence the quality and relevance of the model's output, making prompt design an area of active research and experimentation.

Recent advancements have introduced various prompting techniques, including zero-shot, few-shot, and chain-of-thought prompting.

Zero-shot

LLM have been show to be good at zero-shot reasoning [109], especially when combined with chain-of-thought reasoning [110]. However, unlike previous studies that look at predicting psychological disposition [44], we omit the chain-of-thought reasoning process, finding it unnecessary for our purpose and more prone to hallucinations.

Our results indicate that LLAMA-3, even in a zero-shot setting, can achieve competitive performance in predicting personality traits, as evidenced by key metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE). The ability to analyze predictions at both the facet and item levels provides a deeper understanding of the model's performance and identifies specific areas for further improvement.

Our Method

Leveraging the capabilities of the LLAMA-3 model, our method aims to infer facet scores from the Pennebaker dataset, using item sentences from the International Personality Item Pool (IPI) survey. We propose a streamlined, zero-shot approach using LLAMA-3, focusing on direct score estimation from item sentences related to the IPIP survey, covering three facets per Big Five personality trait.

(author?) [45] evaluated the zero-shot capability of GPT-3 to predict Big 5 personality score, however they used a simplified prompting and evaluation. **(author?)** [111] improved upon this work, using the more realistic prompting and evaluation and showed promise of LLM to do

zero-shot prediction.

Our approach involves providing LLAMA-3 with an instructional prompt followed by the author's text:

Prompt: You are an AI assistant who specializes in text analysis and I am Human. The task is as follows: we have a text written by an author, and I will give you a statement about the author. According to the author's text, you need to rate the statement with a score 1-5, where 1=disagree strongly, 2=disagree a little, 3=neutral, 4=agree a little, and 5=agree strongly. 4.1

The model is then given the item sentences, such as ***The author is outgoing, sociable***, one at a time and asked to estimate scores on a [1-5] scale. Instead of having LLAMA-3 do the calculation, we average the scores afterwards, to avoid any unnecessary errors. These scores are averaged across facets and the Big Five personality traits to assess our model's performance through regression metrics against self-reported scores. Each personality facet is represented by four item sentences, aligning with the structure used in established personality assessments. Item sentences were introduced sequentially, with LLAMA-3 tasked to assign a score within a 1-5 range for each. The process does not rely on chain-of-thought, minimizing the risk of hallucinations while maintaining focus on direct inference capabilities.

Datasets

In this study, we used the two datasets to evaluate the performance of LLAMA-2 and LLAMA-3 in predicting personality traits: the Pennebaker BFI-2 facets dataset and the dataset from Pennebaker and King (1999). Each of these datasets was chosen for its relevance to the tasks at hand, providing both item-level and facet-level personality scores, which are essential for the fine-grained analysis conducted in our experiments.

The Pennebaker BFI-2 facets dataset is a comprehensive collection that includes item

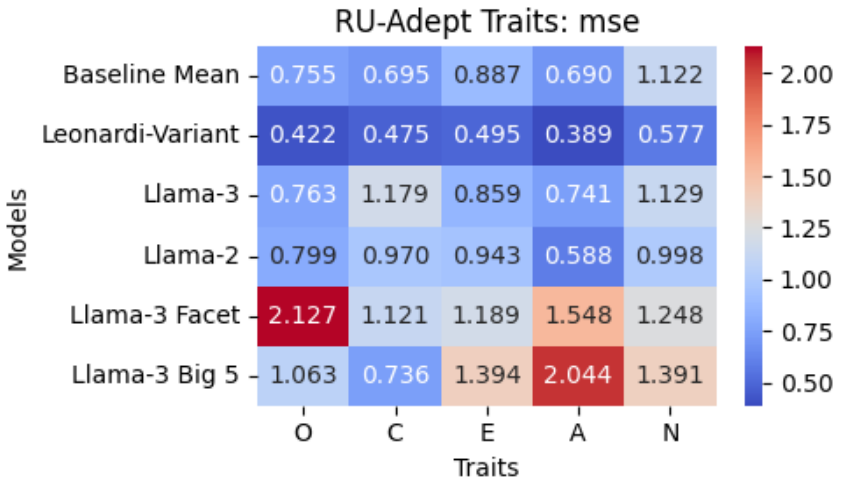

Figure 4.2: MSE scores for LLAMA-3, LLAMA-2, and the baseline models across the Big 5 domains. LLAMA-3 and LLAMA-2 achieve comparable MSE scores in the majority of domains.

sentence scores across various facets of the Big Five personality traits, a subset of a larger dataset collected by Pennebaker and colleagues from students in an American university's introductory Psychology class [112]. This dataset is particularly well-suited for evaluating the granular performance of LLAMA models because it provides detailed item-level scores, allowing us to analyze how well the models predict individual personality facets. Each item in the dataset is associated with a specific facet of a Big Five trait, and participants' responses are scored on a 1-5 scale. For our experiments, we used the same test split across all models to ensure consistency and comparability of results.

For comparison with the PsyCoT paper, we utilized the dataset from **(author?)** [113]. This dataset is well-established in the field of personality prediction and was selected to align with the exact test split used in the PsyCoT study, facilitating a direct comparison between LLAMA-3's zero-shot performance and the results obtained using Chain-of-Thought (COT) reasoning.

Models

In this study, we employed two versions of the LLAMA (Large Language Model Meta AI)

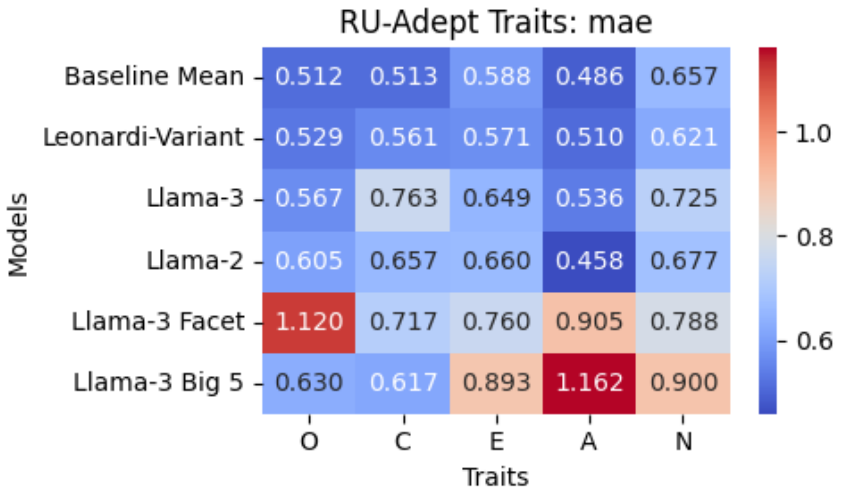

Figure 4.3: MAE scores for LLAMA-3, LLAMA-2, and the baseline models across the Big 5 domains. LLAMA-3 and LLAMA-2 consistently demonstrates similar MAE scores compared to the baseline models

models, specifically LLAMA-2 and LLAMA-3, as the primary tools for predicting personality traits from textual data. LLAMA-2 and LLAMA-3 are state-of-the-art open-source language models with 7 billion and 8 billion parameters, respectively. These models are pre-trained on extensive corpora, enabling them to understand and generate human-like text across a wide range of tasks, including personality prediction.

LLAMA-2, with its 7B parameters, and LLAMA-3, with its expanded 8B parameter set, were both utilized in a zero-shot setting, meaning they were not fine-tuned on our specific datasets before being used to make predictions. These models were prompted with item sentences from the Pennebaker BFI-2 facets dataset and tasked with predicting personality trait scores on a 1-5 scale.

In addition to the LLAMA models, we also utilized a model inspired by the work of **(author?)** [114]. The original *SentencePersonality* model by Leonardi et al. was designed to output a single Big Five domain score for each input text. However, to better align with our research goals, we

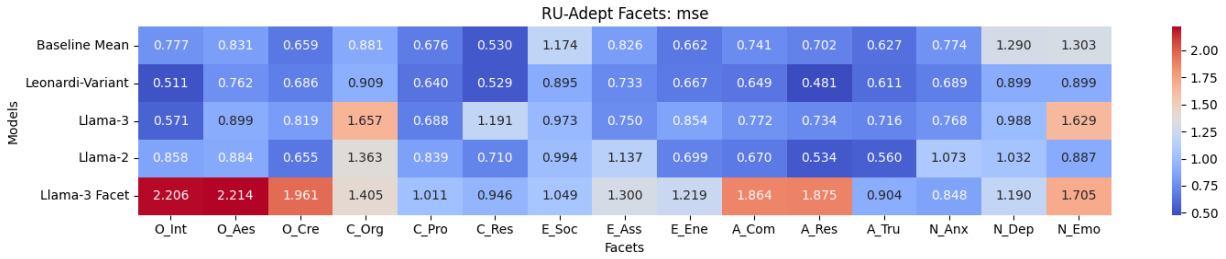

Figure 4.4: MSE scores for LLAMA-3, LLAMA-2, and the baseline models across the 15 facets. LLAMA-3 and LLAMA-2 achieve lower MSE scores in the majority of facets.

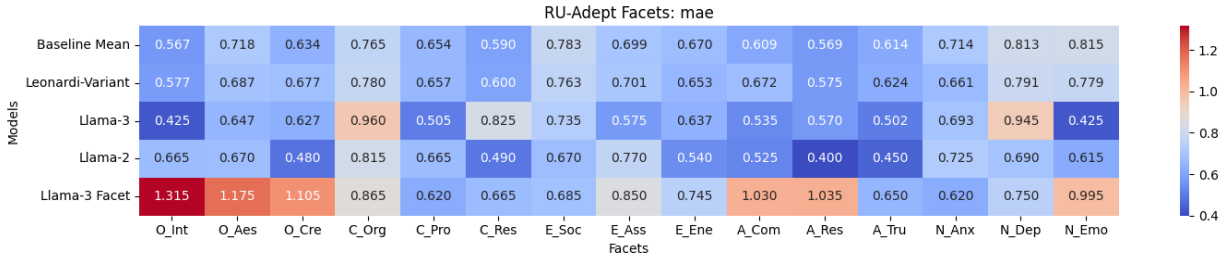

Figure 4.5: MAE scores for LLAMA-3, LLAMA-2, and the baseline models across the 15 personality facets. LLAMA-3 and LLAMA-2 consistently demonstrates lower MAE scores compared to the baseline models, particularly in agreeableness.

modified the final layers of this model to output not only the Big Five domain score but also predictions for the three corresponding BFI-2 facets under each domain.

4.3 Results and Analysis

In this work, we assessed the performance of LLAMA-3 and LLAMA-2 in predicting personality traits from textual data, comparing these models against a baseline approach that predicts the mean score for each facet. We focused on Mean Absolute Error (MAE) and Mean Squared Error (MSE) as our primary evaluation metrics.

LLAMA-3 consistently outperformed the baseline model across most facets, showing superior results in 13 out of 15 facets, each corresponding to one of the Big Five personality traits. This demonstrates the strong predictive capability of LLAMA-3, with significant improvements in both

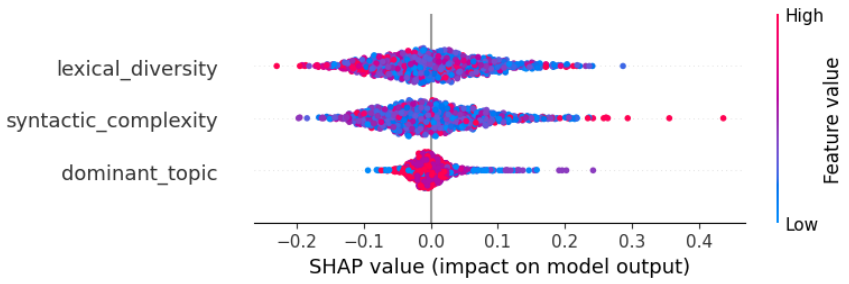

Figure 4.6: Shapley value analysis showing the impact of lexical diversity, syntactic complexity, and dominant topic on the performance of LLAMA-3 in predicting personality traits. The Shapley values indicate the contribution of each feature to the model's predictions, with a normal distribution of impacts observed across all features. This distribution suggests that no single feature disproportionately influences LLAMA-3'S performance

MSE and MAE compared to the naive baseline approach (Figures 4.5, 4.10).

When comparing LLAMA-2 to LLAMA-3, we observed that while their overall MAE and MSE metrics were quite similar, LLAMA-3 exhibited greater consistency and a marked reduction in hallucinations—instances where the model generates plausible yet incorrect responses. These hallucinations were significantly fewer in LLAMA-3, making it a more reliable model. To manage hallucinations without compromising the integrity of the dataset or introducing external biases, we assigned a neutral score of 3 to these data points, thus maintaining the quality of our analysis.

Interestingly, LLAMA-2 outperformed LLAMA-3 in certain cases. However, the frequent hallucinations observed in LLAMA-2, which required replacement with average scores, underscore the greater reliability of LLAMA-3 despite some facets where LLAMA-2'S raw performance was marginally better.

In comparison to state-of-the-art (SOTA) regression models, such as those adapted from the Multilingual Transformer-based personality assessment framework, both LLAMA-2 and LLAMA-3 delivered comparable results, especially at the facet level. This is particularly notable given that our LLAMA models operated in a zero-shot setting, with no fine-tuning or prior task-specific

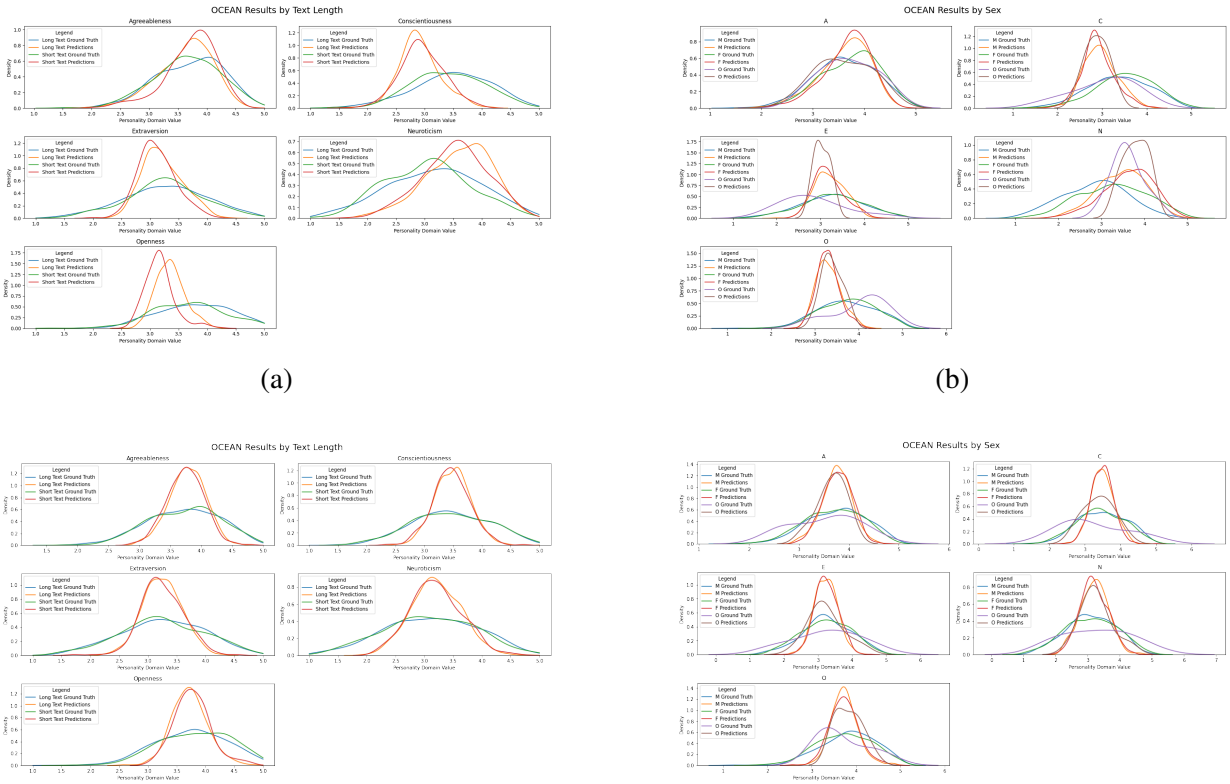

Figure 4.7: Comparison of LLAMA-3 and Leonardi-variant model performance across text length and sex (age showed similar patterns but omitted figures), presented in six subfigures. Subfigures (a) and (b) show LLAMA-3'S ability to better match the underlying ground truth distributions across different age groups, sexes, and text lengths, respectively. In contrast, subfigures (d)and (e) display the baseline model's tendency to predict scores centered around the mean across the same categories. The overall comparison demonstrates LLAMA-3'S better performance in capturing the variability of personality traits, particularly in agreeableness and neuroticism.

training, yet they managed to perform on par with models specifically trained on the dataset.

Our analysis revealed that a bottom-up approach, where facet and Big Five domain scores are calculated from item-level predictions, yields more accurate results than directly querying the LLM'S for these scores. This granular approach not only enhances predictive accuracy but also facilitates a deeper analysis, helping to identify specific items within facets that may underperform, thereby guiding targeted improvements.

When evaluating LLAMA-3 against self-reported scores from the Essays dataset, we accounted

for the inherent margin of error due to the subjective nature of self-reported data. Since self-reported scores are based on individuals' self-perception, they may not always align perfectly with the textual cues analyzed by LLAMA-3. Despite this potential discrepancy, LLAMA-3'S predictions demonstrated reasonable alignment with self-reported scores, further validating the model's ability to accurately infer psychological traits from text.

The success of LLAMA-3 in a zero-shot context underscores its robust generalization capabilities, effectively leveraging pre-trained knowledge to perform well across a variety of predictive tasks. Although LLAMA-3 and LLAMA-2 do not entirely surpass SOTA regression-based models, their ability to achieve similar performance without any fine-tuning highlights their potential and efficacy in computational personality assessment.

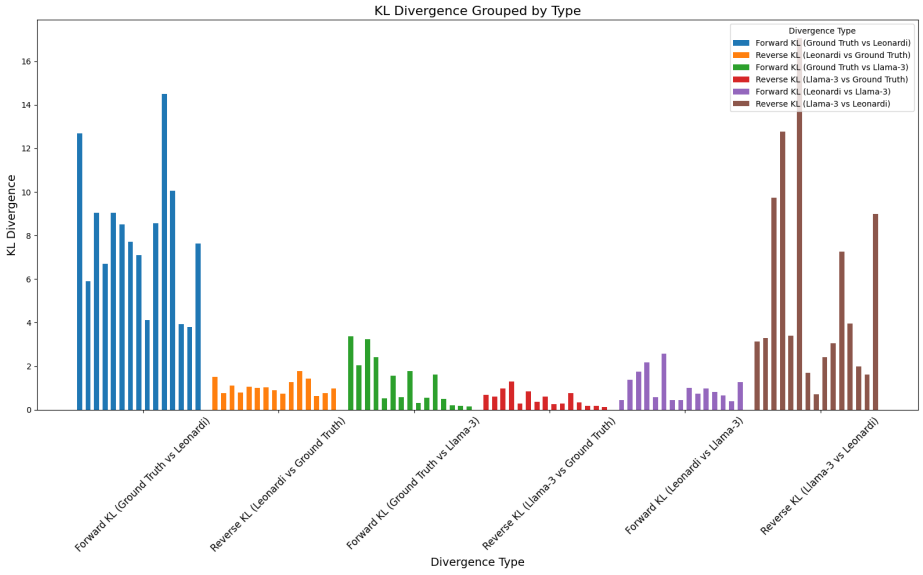

Figure 4.8: Bar plot of the KL divergence between the ground truth and two models, Leonardi and Llama-3, across 15 facets. Llama-3 consistently exhibits significantly lower KL divergence compared to Leonardi indicating that Llama-3 more closely aligns with the ground truth distribution across all facets.

To further understand LLAMA-3'S performance across different score ranges, we conducted an analysis by dividing the scores into quartiles within the 1-5 scale. This allowed us to examine

LLAMA-3'S performance on central tendency scores (2-4) versus extreme scores (1 and 5). Our findings indicate that while LLAMA-3 generally outperforms baselines, the baseline models tend to produce unimodal distributions centered around the mean. In contrast, LLAMA-3'S predictions exhibit multimodal peaks, reflecting its ability to capture more nuanced variations in the data rather than simply clustering around central values. (Figures 4.12, 4.13)

When examining performance across all 15 facets, LLAMA-3 demonstrated a superior ability to match the underlying ground truth distributions of the scores compared to the baseline model, which consistently centered around the mean. This suggests that LLAMA-3 can better reflect the diversity in the dataset, capturing the true variability of the personality traits across different individuals.

LLAMA-3 performed particularly well on the agreeableness facets, displaying strong predictive accuracy across all quartiles. The model's nuanced understanding of textual data related to agreeableness significantly contributed to its overall strong performance in this trait. However, a bias was noted in the neuroticism facets, where LLAMA-3 tended to overestimate neurotic traits compared to self-reported scores. This overestimation may stem from the nature of the textual data or the model's interpretation of language related to emotional and psychological states.

To ground our reasoning more, we measured the KL divergence between both models with the ground-truth user reported scores (Figure 4.8). The forward and reverse KL divergence analysis provides insights into the comparative performance of Llama-3 and Leonardi's *SentencePersonality* in capturing the ground truth distribution and each other's distributions. The forward KL divergence $KL(\text{Ground Truth}|\text{Llama-3})$ is consistently lower than $KL(\text{Ground Truth}|\text{Leonardi})$. This result strongly suggests that Llama-3 more effectively aligns with and covers the ground truth distribution compared to Leonardi. This outcome aligns with prior expectations that Llama-3's model architecture

and training methodology allow it to generalize better across facets of the ground truth. The analysis of forward and backward KL divergence between Llama-3 and Leonardi further corroborates these observations. The forward KL divergence $KL(\text{Leonardi}|\text{Llama-3})$ is relatively small, indicating that Llama-3 captures the core characteristics of Leonardi's distribution, which is concentrated around the mean. However, the reverse KL divergence $KL(\text{Llama-3}|\text{Leonardi})$ is notably higher, illustrating that Leonardi struggles to cover the full variance of Llama-3's distribution. This asymmetry highlights a significant limitation of Leonardi's modeling capability in contrast to Llama's.

This plot affirms our other results that Llama-3 not only aligns more closely with the ground truth but also captures a wider range of distributions, including Leonardi's, while maintaining its broader variance. These findings provide a compelling case for Llama-3's ability to generalize and model complex distributions effectively. This capability is significant, especially for real-world applications where personality prediction is utilized. For tasks such as depression or suicide detection, individuals who exhibit atypical patterns, often those far from the mean of the distribution, are the ones most in need of accurate detection and intervention. The results suggest that Llama-3's broader coverage of the distribution is well-suited to these tasks.

Unlike Leonardi, which centers its predictions around the mean and struggles to capture variance, Llama-3's ability to account for outliers ensures a more comprehensive representation of the population. This capability is critical for identifying individuals with extreme or atypical personality traits, which are often markers of underlying mental health issues. Missing these outliers due to poor variance coverage could result in missed opportunities for early intervention in high-risk individuals.

To further investigate the robustness of LLAMA-3'S performance, we segmented the data

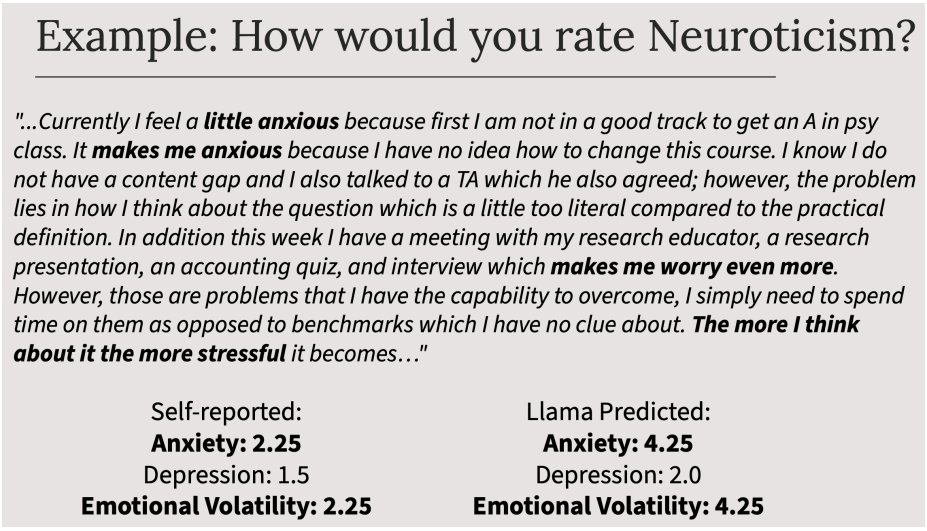

Figure 4.9: Example passage from the dataset with corresponding neuroticism scores. The figure displays the original text alongside the self-reported score and the LLaMA-predicted score. In this case, the passage contains linguistic markers (e.g., expressions of worry and self-criticism) that may lead readers to judge the author as higher in neuroticism, more closely aligning with the model's prediction than with the self-report. This example illustrates the central challenge of personality prediction: the "ground truth" provided by self-reports may not always reflect observable linguistic cues, raising the question of whether models should optimize for agreement with self-perception or for external validity

by sex, age, and length of text. In each subgroup analysis, LLAMA-3 continued to outperform the baseline, which remained centered around the mean. Notably, LLAMA-3'S ability to match the underlying distributions was particularly strong for the agreeableness facets across all categories, demonstrating consistent accuracy irrespective of subgroup (Figure 4.7).

Additionally, when segmenting the data by sex, age, and text length, we observed that LLAMA-3'S predictions for neuroticism also closely matched the underlying distributions. This is particularly noteworthy given LLAMA-3'S initial tendency to overestimate neuroticism compared to self-reported scores, suggesting that while a bias exists, the model still captures a more accurate distribution of the trait than the baseline model.

LLAMA-3 performed best on the agreeableness facets, showing strong predictive accuracy across all quartiles and demographic categories. The model's nuanced understanding of textual

data related to agreeableness significantly contributed to its overall strong performance in this trait. A notable bias was observed in the neuroticism facets, where LLAMA-3 tended to predict higher scores compared to the self-reported scores. This suggests a tendency of the model to overestimate neurotic traits in users, which could be attributed to the nature of the textual data or the model's interpretation of language related to emotional and psychological states. However, when accounting for demographic splits, LLAMA-3'S predictions still better reflected the actual distribution of neuroticism traits than the baseline, particularly in longer texts.

To better understand where LLAMA-3 performs poorly, we explored various text-level features, including sentiment (skew, variance, kurtosis), length, readability, text topic, and lexical diversity. Upon analysis, we found no significant correlation between these features and the accuracies of any facets or domains. This suggests that LLAMA-3'S performance is not directly influenced by these factors, indicating that the model's predictive capability may be more robust and not heavily dependent on such text characteristics.

Further analysis using Shapley values explored the impact of lexical diversity, syntactic complexity, and dominant topic on the model's performance. The Shapley value analysis revealed a normal distribution of impact across these features, indicating that none of these factors disproportionately influenced the model's predictions. This balanced distribution suggests that LLAMA-3'S performance is driven by a complex interplay of multiple factors rather than being dominated by any single feature (Figure 4.6).

We also conducted a comparative analysis between LLAMA-3 and the PsyCoT model, using the same dataset in a zero-shot setting without Chain-of-Thought (CoT) reasoning. Despite the lack of fine-tuning and the absence of CoT, LLAMA-3 achieved scores that were very comparable to those of PsyCoT. This demonstrates that LLAMA-3 offers a more cost-effective and straightforward

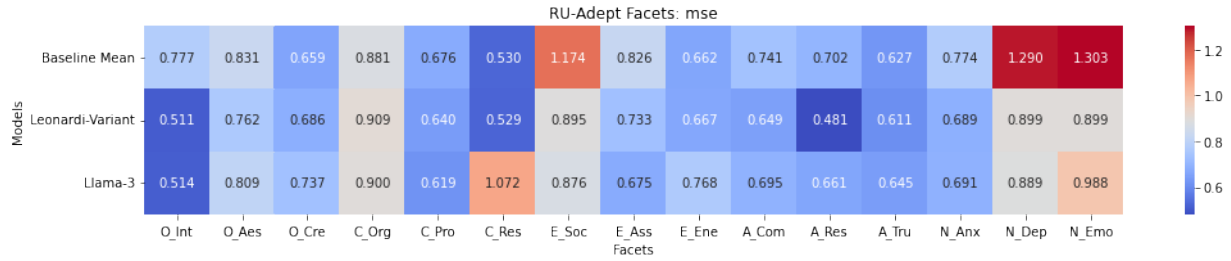

Figure 4.10: MSE of LLaMA-3 predictions across the 15 BFI-2 facets after LoRA fine-tuning on neuroticism. Fine-tuning reduces error on the three neuroticism facets (anxiety, depression, emotional volatility).

alternative to PsyCoT, maintaining similar levels of performance while simplifying the computational process (Table 4.1).

Methods	**AGR**		**CON**		**EXT**		**NEU**		**OPN**		**Average**	
	Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Llama	61.11	53.85	61.54	54.54	60.32	52.43	61.94	55.66	61.51	54.11	61.28	54.17
PsyCoT	61.13	61.13	59.92	57.41	59.76	59.74	56.68	56.58	60.73	57.30	59.64	58.43

Table 4.1: Comparison of LLAMA-3 and PsyCoT on predicting personality traits across the same dataset. Both models were evaluated in a zero-shot setting without any task-specific fine-tuning. The table highlights that LLAMA-3 achieves performance comparable to PsyCoT across all facets, demonstrating similar accuracy and F1. Despite the absence of Chain-of-Thought (CoT) reasoning in LLAMA-3, it provides a cost-effective and simpler alternative to PsyCoT, maintaining similar levels of accuracy in predicting the Big Five personality traits from text.

4.3.1 Fine-tuning through LoRA

In our zero-shot experiments, the neuroticism facets consistently produced the highest error rates across both MAE and MSE. Qualitative inspection suggested that while LLaMA was able to capture general tendencies for traits such as agreeableness or extraversion, it struggled with reliably estimating linguistic signals tied to neuroticism—often over-predicting relative to self-reported ground truth. To address this, we conducted a series of experiments applying Low-Rank Adaptation (LoRA) fine-tuning to LLaMA-3. LoRA is particularly well-suited for this

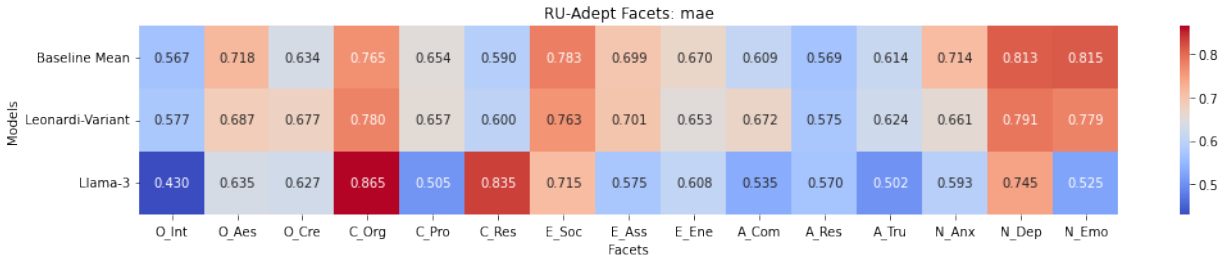

Figure 4.11: MSE of LLaMA-3 predictions across the 15 BFI-2 facets before and after LoRA fine-tuning on neuroticism. Fine-tuning reduces error on the three neuroticism facets (anxiety, depression, emotional volatility).

task: it enables parameter-efficient adaptation of large models while requiring only a small amount of additional computation and storage. We fine-tuned the model specifically on items corresponding to the three facets of neuroticism (anxiety, depression, and emotional volatility), using the Pennebaker BFI-2 dataset. Results in Figure 4.11 show a reduction in both MAE and MSE compared to the zero-shot baseline, indicating that fine-tuning improves predictive accuracy for neuroticism. However, distributional analysis paints a more complex picture. While the fine-tuned model produces scores that are closer to self-reported values on average, its predictions are noticeably more normalized, clustering around the mean of the distribution. This reflects a classic tradeoff: fine-tuning optimizes for pointwise accuracy under MAE/MSE metrics, but at the cost of reducing variance and flattening predictions. This tradeoff raises a central question: should models be optimized to match the statistical properties of ground-truth distributions (capturing variance, including outliers), or to minimize average error relative to self-reports? In tasks such as clinical screening, the former may be more important, since atypical individuals are often those most in need of detection.

4.4 Conclusion

This chapter explored the use of open-source large language models, specifically LLaMA-2 and LLaMA-3, for predicting personality traits from text in a zero-shot setting. By grounding predictions in item-level prompts from the BFI-2 survey, we demonstrated that LLaMA models can approximate human self-reports with competitive accuracy, often outperforming baselines and capturing richer variance across the distribution of responses. This item-driven approach not only provides fine-grained interpretability but also aligns model predictions with established psychometric procedures, offering a more transparent bridge between computational inference and psychological theory.

Our experiments revealed several key findings. First, LLaMA's zero-shot predictions are effective across most Big Five traits, with notable strength in agreeableness and extraversion. Second, neuroticism facets emerged as the most challenging, with higher error rates and a systematic tendency to overpredict. Fine-tuning with LoRA reduced error metrics for neuroticism but at the cost of producing more normalized, less variable distributions—highlighting a tradeoff between minimizing pointwise error and preserving alignment with the natural variance in human responses.

These results raise an important methodological question: what constitutes ground truth in computational personality prediction? Self-reported survey scores are widely used but remain inherently subjective, shaped by perception, context, and reporting biases. In cases where model predictions better align with observable linguistic cues than with self-reports, it is unclear whether the model should be penalized or credited. Addressing this tension requires moving beyond accuracy against self-reports and toward human-in-the-loop evaluation frameworks, where domain

experts help adjudicate ambiguous cases and guide model alignment. Reinforcement learning with human feedback (RLHF) represents a promising path forward, allowing psychologists to shape model behavior in ways that balance predictive accuracy, interpretability, and external validity.

Taken together, this chapter extends the dissertation's broader theme of human-centered NLP. Just as interactive topic modeling illustrates how humans can steer machine representations, personality prediction highlights how models can provide structured, interpretable insights back to humans. Both lines of work underscore the importance of viewing NLP not simply as a problem of optimization, but as a reciprocal process where models and humans jointly shape meaning, interpretation, and understanding.

4.5 Future Work

The divergence between zero-shot and fine-tuned predictions also highlights a deeper methodological issue: to what extent should self-reported scores be treated as ground truth? Psychological assessments such as the BFI-2 are widely validated, but they remain fundamentally subjective. Respondents may underreport or overreport traits depending on self-perception, social desirability, or contextual factors. Models trained solely to minimize error against self-report data risk inheriting these biases, thereby optimizing for self-perception rather than psychological reality.

This raises an open question in computational personality prediction: should the benchmark be agreement with self-reports, or should models aim for psychologist-validated annotations that more directly capture behavioral traits? We argue that progress in this space will likely require human-in-the-loop approaches, where domain experts guide the model's learning process. One

promising direction is reinforcement learning with human feedback (RLHF), using psychologists as annotators to align model outputs with expert judgment rather than raw self-reports. Such methods could help balance predictive accuracy with interpretability and reliability, especially for facets like neuroticism where subjective reporting is most variable.

While we leave the implementation of RLHF-based pipelines for future work, this line of inquiry underscores a recurring theme in this dissertation: the importance of interactive and human-centered NLP. Just as topic models benefit from user guidance, personality prediction systems may need expert input to resolve the ambiguities of subjective ground truth.