PsyTEx: A Knowledge-Guided Approach to Refining Text for Psychological Analysis

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, Gregory D. Webster, Damon L. Woodard
Florida Institute for National Security (FINS), University of Florida, USA
Correspondence: avantibhandarkar@ufl.edu

Abstract

LLMs are increasingly applied for tasks requiring deep interpretive abilities and psychological insights, such as identity profiling, mental health diagnostics, personalized content curation, and human resource management. However, their performance in these tasks remains inconsistent, as these characteristics are not explicitly perceptible in the text. To address this challenge, this paper introduces a novel protocol called the "Psychological Text Extraction and Refinement Framework (PsyTEx)" that uses LLMs to isolate and amplify psychologically informative segments and evaluate LLM proficiency in interpreting complex psychological constructs from text. Using personality recognition as a case study, our extensive evaluation of five SOTA LLMs across two personality models (Big Five and Dark Triad) and two assessment levels (detection and prediction) highlights significant limitations in LLM's ability to accurately interpret psychological traits. However, our findings show that LLMs, when used within the PsyTEx protocol, can effectively extract relevant information that closely aligns with psychological expectations, offering a structured approach to support future advancements in modeling, taxonomy construction, and text-based psychological evaluations.

1 Introduction

Large Language Models (LLMs) are transforming the field of natural language processing (NLP), performing remarkably as linguistic tools skilled in language manipulation, reasoning, explanation, and information extraction. Equipped with billions of parameters, these models excel at processing and retaining vast amounts of information, reaching state-of-the-art (SOTA) performance in a variety of tasks including text summarization (Zhang et al., 2024), Question Answering (OpenAI, 2023; DeepMind, 2023; AI@Meta, 2024), and natural language inference (NLI) (Zhong et al., 2023; Gubelmann et al., 2023; Wang et al., 2024), etc (Yang et al., 2024). which are evaluated against standard benchmarks designed to measure their zero-shot and few-shot capabilities in language understanding and information extraction (Laskar et al., 2023; Qin et al., 2023; Wang et al., 2018; Hendrycks et al., 2020; Rein et al., 2023; Zheng et al., 2024)

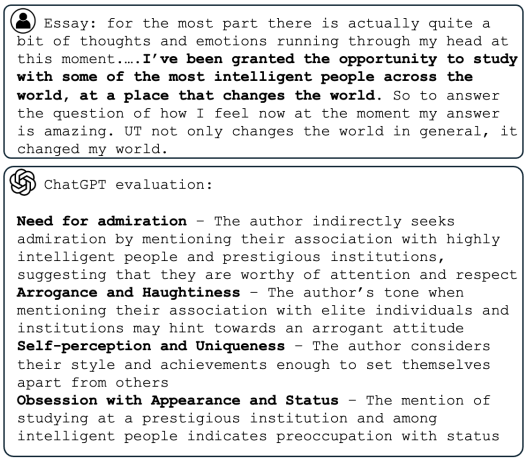


Figure 1: Narcissism Assessment from an Essay where ChatGPT Eval - High and Human Eval - Low

While near-perfect performance in these tasks showcases LLMs' ability to "understand" language, incorporating both semantic and contextual knowledge, standard benchmarks do not typically evaluate their "interpretive" capabilities. Assuming that LLMs can handle psychological evaluations and human categorization, preliminary studies using zero-shot prompting for tasks like authorship verification, author attribution, and psychological profiling, including the detection of implicit social signals such as sarcasm, personality, and implicit sentiment, reveal that their performance frequently borders on random chance (Hung et al., 2023; Bhandarkar et al., 2024b; Amin et al., 2023; Zhang et al., 2023). For example, consider a scenario (Figure 1) where ChatGPT assessed the personality trait of Narcissism from a human-authored essay. It incorrectly identified the highlighted sen-

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 151–178
May 3-4, 2025

tence as indicative of Narcissism and assigned the essay a high score, despite its actual low score. While a human might see the sentence as an expression of gratitude, a behavior typically inconsistent with Narcissism, ChatGPT misinterprets it, incorrectly identifying it as evidence of the trait.

These observations, alongside the findings from preliminary studies, suggest that current LLMs may not possess the required capabilities to effectively interpret nuanced information from text. This shortcoming is particularly critical given the potential of LLMs to revolutionize areas such as identity profiling, personalized advertising, mental health assessments, and human resources.

Highlighting the example of personality recognition where LLMs have shown notably poor performance, this work seeks to answer the question "Can LLMs effectively *interpret* psychological characteristics from text?".

2 Related Works

Personality recognition has been a longstanding area of research, with numerous studies aiming to develop models capable of personality evaluation from text (Mehta et al., 2020; Mushtaq and Kumar, 2022; Zhao et al., 2022). However, the effectiveness of these efforts is limited by the complexity of extracting subtle and often imperceptible cognitive markers from the text (Bhandarkar et al., 2024a). In recent years, there has been growing interest in utilizing the LLMs for personality assessments.

Most advanced approaches using LLMs for this purpose assume that LLMs can assess these cognitive characteristics and that their effectiveness can be enhanced by curating specialized prompts (Amin et al., 2023; Ji et al., 2023; Hu et al., 2024; Yang et al., 2023). Several techniques have been proposed in recent literature, including zero-shot prompting, chain-of-thought (CoT) prompting, and many specialized prompting methods. However, the findings remain inconsistent. While some works indicate that LLMs are not yet suitable for direct use as psychological evaluation tools, others present contradictory results (Wen et al., 2024).

Key factors contributing to this disparity are the reliance on lexical models for labeling that exhibit weak correlations with actual personality scores, synthetic datasets generated by LLMs, and questionnaire-based evaluations, where LLMs are artificially induced with personality traits and then assessed on their responses to personality questionnaires (Vu et al., 2024; Li et al., 2024; Jiang et al., 2024). While effective for evaluating AI agents and chatbots, these methods lack ground truth human data, risking overestimation of LLM capabilities and poor generalization to real-world populations. In contrast, our work evaluates LLMs on human-authored text, ensuring assessments align with natural language patterns and reinforcing both the validity and applicability of our findings for real-world psychological analysis.

More importantly, existing approaches do not assess whether LLMs possess true "interpretive" capabilities or merely rely on superficial linguistic patterns for personality assessment. Several studies suggest that LLMs can enhance their outputs through self-refinement, where models assess their own responses or follow self-generated checklists for structured reasoning (Madaan et al., 2024; Cook et al., 2024). If LLMs can apply similar internal evaluation mechanisms to psychological constructs, they may be capable of more nuanced personality assessment. However, this remains largely unexplored. Thus, it is crucial to deconstruct how LLMs might analyze psychological constructs from text to assess their interpretive capabilities.

To address this, we introduce a novel protocol named *Psychological Text Extraction and Refinement Framework* (PsyTEx) to simulate the process by which an LLM evaluates psychological characteristics. As depicted in Figure 2, this process comprehensively probes the LLM's domain knowledge and its ability to extract application-specific information and integrates evaluation capabilities using the standard prompting protocol in a standalone yet explainable step-by-step fashion. Furthermore, this framework is highly adaptable and can be seamlessly extended to incorporate other prompting techniques while maintaining the same foundational framework. This work makes the following contributions[1]:

- We introduce PsyTEx, a knowledge-guided text refinement framework to extract and amplify psychologically relevant information from text using LLMs, offering a structured methodology for evaluating the interpretive capabilities of LLMs in human categorization tasks like personality recognition.
- We present the first comprehensive zero-shot analysis of five SOTA LLMs (GPT-4o,

[1]Data and code can be accessed here.

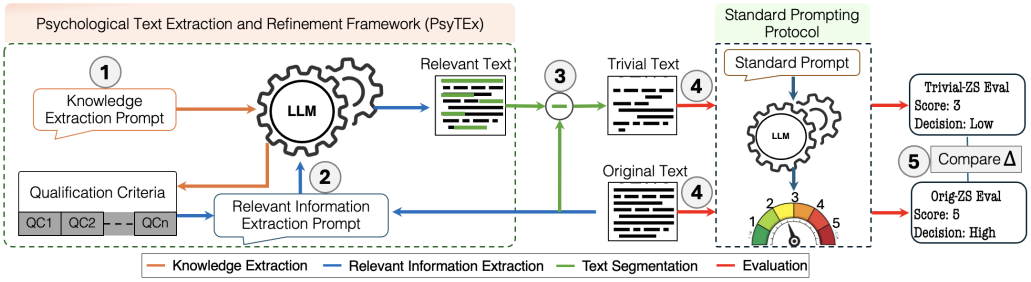


Figure 2: Overview of the PsyTEx experimental protocol. Steps are enumerated for clarity and ease of understanding.

Llama3, Mistral, OpenChat, Phi3) on two personality models (Big Five and Dark Triad) across two settings (detection and prediction).

- Our findings reveal critical limitations of LLMs in achieving SOTA results for tasks that necessitate deep textual interpretation, shedding light on the inherent challenges.
- We demonstrate that PsyTEx-refined text aligns closely with the psychological expectations, as validated by LIWC, highlighting its potential for psychological modeling, taxonomy creation, and text-based psychological assessments.

Knowledge Extraction Prompt

According to your knowledge, how is the personality trait {P} manifested in text? Can you give me an exhaustive list of textual manifestations of {P} in the order of importance and relevance to the Personality Psychology literature?

Figure 3: Prompt for Knowledge Probing

Relevant Information Extraction Prompt

Consider the following essay response carefully and evaluate each of the qualification criteria from the following list. Please refrain from making assumptions about the relevance of these qualifications to any specific personality trait(s) and disorder(s) and base your evaluations with utmost objectivity purely on the essay. When encountered, provide all relevant textual evidences of each criteria and how it manifests in the text. Finally present summary of your overall findings.
Criteria: {List of qualification criteria}

Figure 4: Prompt for Personality-relevant Information extraction

3 Methodology

The methodology for PsyTex is structured in two main steps: *knowledge extraction* and *relevant information extraction*, followed by a systematic protocol for assessing the interpretive ability of LLMs.

Knowledge Extraction Phase: The first step involves presenting the LLM with an open-ended question designed to elicit its knowledge of Personality Psychology, using a prompt outlined in Figure 3. To ensure insightful and pertinent responses, the LLM must also explain the relevance of its responses and provide examples of trait manifestations in the text. This phase assesses the LLM's foundational knowledge and its ability to retrieve and apply relevant psychological concepts for personality assessment. For each LLM-trait pair, the responses are cataloged as *Qualification Criteria*, reflecting the LLM's understanding of personality traits. Qualification criteria generated by all LLMs are presented in Tables 11 to 15 in Appendix A. Five variations of knowledge extraction prompts were tested, revealing that the generated qualification criteria remained stable across different phrasings (see Appendix A.4.3, Figure 10).

Relevant Information Extraction Phase: Next, we evaluate how LLMs utilize this knowledge in practice. Recent studies suggest that LLMs are adept at pinpointing relevant information within texts (Yuan et al., 2024; Guo et al., 2024; Goel et al., 2023). We harness this ability by using a prompt (shown in Figure 4) to guide LLMs in identifying text segments, referred to as *Relevant text*, that correspond with predetermined qualification criteria, thereby isolating text most indicative of personality traits. These tagged segments are assumed to represent portions of the text that LLMs focus on when assessing personality. To encourage deeper reasoning, the LLMs are prompted to explain their tagging decisions and how text segments meet the qualification criteria. This tagging exercise serves a dual purpose: it showcases the LLMs' ability to recognize and highlight personality-relevant text based on their knowledge and sets the stage for a critical evaluation of their performance.

Assessing the Interpretive Ability of LLMs: To assess whether LLMs effectively use their knowledge to infer personality traits, we perform *Text Segmentation*, where relevant text identified in the previous stage is removed from the original text, leaving behind *Trivial Text*, that is presumed to be irrelevant to the personality trait.

The final step evaluates the impact of remov-

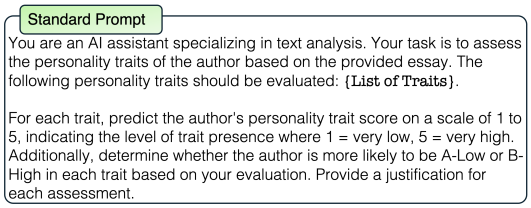


Figure 5: Standard Zero-Shot Probing Prompt

ing relevant text on personality assessments. Our protocol employs the simplest and widely used zero-shot personality evaluation approach known as the *Standard Prompting Protocol* outlined by Yang et al. (2023). The decision to use zero-shot prompting is based on two key reasons: first, existing research indicates that zero-shot prompting may outperform few-shot prompting, particularly when advanced prompting techniques are applied (Reynolds and McDonell, 2021); second, because personality traits are inherently subtle and not directly observable in text, providing few-shot examples could introduce a mismatch between the input text and the expected labels, potentially confusing the LLM and leading to a decrease in performance. Ultimately, we aim for the LLMs to rely on their intrinsic knowledge to perform personality evaluation.

We apply this evaluation separately to both the original text (*Orig-ZS*) and the trivial text (*Trivial-ZS*). This allows us to observe any changes in the LLM's performance and understand the importance of the extracted text segments. If the LLMs truly use their knowledge to assess personality, a decline in performance is expected after removing relevant text. Conversely, minimal change or improvement in performance could suggest that despite possessing relevant knowledge, LLMs are unable to apply this understanding in practice, supporting the hypothesis that LLMs might struggle to interpret complex and implicit psychological constructs like personality traits.

3.1 Datasets

To rigorously test LLMs' ability to interpret personality traits, three criteria must be met: First, data should be high-quality, scientifically robust, and tailored to reflect personality in text. Second, it should include both positive and negative traits to ensure broad LLM applicability and an accurate representation of traits found in the general population. Lastly, since personality is often assessed on a continuum (typically, a 5-point Likert scale), datasets with trait scores are crucial for evaluating LLMs' nuanced zero-shot evaluation abilities.

Most publicly available datasets for personality assessment fail to meet all three criteria. Therefore, we sourced the Sample14 dataset, which provides text samples from over 1,100 individuals across various test scenarios, featuring personality trait scores from two models: the Big Five (Openness (O), Conscientiousness (C), Extroversion (E), Agreeableness (A), Neuroticism (A)) and the Dark Triad (Machiavellianism (Mach), Narcissism (Narc), Psychopathy (Psyc)) (Carey et al., 2015). To align with existing literature and establish a comparative baseline, we also utilize the widely recognized gold-standard dataset, Essays. This dataset contains over 2,400 text samples with binary labels (Low/High) for the Big Five personality traits (Pennebaker and King, 1999). Dataset and implementation details in Appendices A.1 and A.2.

4 Results

In this section, we evaluate the performance of LLMs for personality recognition under zero-shot settings. The two datasets facilitate coarse personality detection and fine-grained personality prediction. Personality detection involves binary classification to differentiate between "high" or "low" trait categories, while personality prediction involves regression analysis to estimate precise trait scores.

4.1 Performance on Original Text

The results under the Orig-ZS setting for both paradigms are presented in Tables 1 and 2 where performance for detection is measured with the classification metric - accuracy, to enable comparison to related studies. The performance for prediction is measured with the regression metric - Root Mean Squared Error (RMSE). Close to random chance accuracy and high RMSE values for both problems is observed. Given the complex nature of zero-shot personality prediction—arguably a more intricate task than detection—these elevated RMSE values align with previous findings and are not entirely unexpected (Ganesan et al., 2023).

Further, performance variability across three dimensions was analyzed: studies for personality detection, LLMs, and personality traits. LLMs that effectively assess personality should demonstrate consistent performance across studies and traits. However, some variability among LLMs is expected due to their differing interpretative skills.

Source paper	LLM used	Strategy	O	C	E	A	N	Average
Ji et al. (2023)	GPT3.5-Turbo	Zero-shot	**0.61**	**0.56**	**0.51**	**0.59**	**0.61**	**0.58**
		Zero-shot CoT	0.66	0.53	0.49	0.61	0.6	0.58
		One-shot	0.58	0.54	0.59	0.59	0.61	0.58
		$LO-Zero-shot_{CoT\_W}$	0.59	0.57	0.5	0.59	0.61	0.57
		$LO-Zero-shot_{CoT\_S}$	0.62	0.55	0.52	0.59	0.59	0.57
		$LO-Zero-shot_{CoT\_D}$	0.64	0.57	0.51	0.6	0.6	0.58
Yang et al. (2023)		Zero-shot	**0.56**	**0.57**	**0.6**	**0.59**	**0.61**	**0.59**
		Zero-shot CoT	0.59	0.55	0.58	0.59	0.57	0.58
		PsyCoT	0.61	0.6	0.6	0.61	0.57	0.60
Our	GPT3.5-Turbo	Orig-ZS	**0.57**	**0.55**	**0.55**	**0.52**	**0.57**	**0.55**
		Trivial-ZS	0.54	0.54	0.52	0.53	0.55	0.54
	Mistral	Orig-ZS	0.54	0.49	0.5	0.52	0.56	0.52
		Trivial-ZS	0.55	0.53	0.55	0.54	0.51	0.54
	Llama3	Orig-ZS	0.56	0.54	0.54	0.54	0.58	0.55
		Trivial-ZS	0.54	0.53	0.53	0.54	0.55	0.54
	OpenChat	Orig-ZS	0.56	0.57	0.54	0.49	0.59	0.55
		Trivial-ZS	0.52	0.51	0.49	0.49	0.53	0.51
	Phi3	Orig-ZS	0.54	0.55	0.52	0.52	0.58	0.54
		Trivial-ZS	0.54	0.56	0.54	0.56	0.58	0.56
	GPT4-o	Orig-ZS	0.55	0.60	0.59	0.59	0.53	0.57
		Trivial-ZS	0.60	0.54	0.55	0.57	0.57	0.57

Table 1: Comparison of accuracy for the Essays dataset with SOTA results. Performance values from other works employing variations of zero-shot prompting are reported from source papers. Values closely matching our experimental setup are bolded, and LLMs with the highest performance in the Orig-ZS setting are underlined.

The analysis indicated variability across studies and traits, while variability among LLMs was minimal. This points to a possible element of randomness in the LLM-generated outputs. For instance, in detection using the same LLM (GPT-3.5[2]) and identical standard prompting method on the same dataset, accuracy showed a standard deviation ranging from 1-5%. Between the two reported studies, the absolute difference in accuracy when using similar zero-shot CoT prompting varied between 2 and 9%. Additionally, performance also varied across traits, with Neuroticism showing the highest average performance (0.57) and Agreeableness the lowest (0.53) across all studies.

For prediction, substantial variability between the two personality models was noted, with particularly high RMSE values for Dark Triad traits such as Psychopathy, likely due to these traits being less overtly manifested in text. For detection, the performance of open-source LLMs closely mirrors that of the most sophisticated LLM, ChatGPT (-3.5 and -4o), with a maximum difference of 5% between the highest (open-source) and lowest (closed-source) average accuracies. Similarly, despite some fluctuations, the performance across all LLMs remained relatively uniform and close to random chance. This suggests that *there are no significant differences in the ability of LLMs to assess personality traits under standard zero-shot conditions.*

4.2 Effect of Relevant Text Removal

In the Trivial-ZS scenario, removing relevant text is expected to decrease overall LLM performance compared to Orig-ZS. For detection, this would result in perfect performance for the 'low' class and significantly lower for the 'high' class. In personality prediction, the RMSE is likely to rise significantly due to the loss of crucial information.

We examine the differences (Δ) in class recall scores for detection and RMSE for prediction across the two probing settings presented in Tables 2 and 3. Numerically, Δ represents the difference calculated as Orig-ZS performance minus Trivial-ZS performance. LLMs adjusting their evaluations based on input text are likely to show a significant negative Δ value for the "low" class and a positive Δ for the 'high' class in detection. For prediction, a high negative Δ is expected. Conversely, if LLM evaluations are random, minimal or opposite-direction trends in Δ values are expected.

For detection, GPT-4o and OpenChat stand out as the only models that meet the required criteria for Δ for at least three out of five traits and show the highest Δ, especially for Openness and Conscientiousness. However, it is important to note

[2]Note that GPT-3.5 was only used for comparison with existing methods, while all other experiments employ the more recent GPT-4o model.

LLM used	Strategy	O	C	E	A	N	Mach	Narc	Psych
Mistral	Orig-ZS	0.87	0.95	1.14	1.01	0.99	1.59	1.00	2.33
	Δ	**-0.02**	**-0.07**	**-0.06**	**-0.10**	0.12	0.41	**-0.36**	0.63
Llama3	Orig-ZS	0.77	1.19	1.33	0.96	1.13	1.81	1.03	2.49
	Δ	**-0.38**	**-0.12**	**-0.18**	**-0.23**	0.07	0.16	**-0.03**	0.25
OpenChat	Orig-ZS	0.75	0.97	1.11	0.80	0.95	1.73	1.00	2.09
	Δ	**-0.11**	0.00	0.15	**-0.05**	0.08	0.52	0.03	0.36
Phi3	Orig-ZS	0.88	1.15	1.31	1.03	1.08	1.98	1.02	2.45
	Δ	0.04	0.05	0.13	0.04	0.04	0.21	**-0.03**	0.22
GPT4-o	Orig-ZS	0.84	1.11	1.26	1.01	1.11	1.65	0.96	2.39
	Δ	**-0.50**	**-0.39**	**-0.33**	**-0.10**	0.00	**-0.25**	**-0.09**	0.05

Table 2: Personality Prediction results on Sample14 dataset reported as Root Mean Squared Error (RMSE). "Δ" represents the difference between the two evaluation settings. Bold values confirm Δ expectations.

LLM	Strategy	O		C		E		A		N	
		Low	High	Low	High	Low	High	Low	High	Low	High
Mistral	Orig-ZS	0.5	0.57	0.51	0.48	0.57	0.43	0.19	0.81	0.33	0.79
	Δ	0.29	-0.29	0.25	-0.31	0.16	-0.25	**-0.07**	**0.03**	0.29	-0.2
Llama3	Orig-ZS	0.23	0.87	0.62	0.45	0.36	0.7	0.7	0.39	0.4	0.76
	Δ	**-0.07**	**0.11**	0.17	-0.16	**-0.04**	**0.05**	0.06	-0.06	0.09	-0.04
OpenChat	Orig-ZS	0.48	0.63	0.76	0.37	0.79	0.3	0.91	0.1	0.53	0.64
	Δ	**-0.35**	**0.41**	**-0.19**	**0.29**	**-0.07**	**0.14**	-0.02	0	**-0.17**	**0.28**
Phi3	Orig-ZS	0.27	0.79	0.48	0.63	0.52	0.53	0.34	0.68	0.47	0.69
	Δ	0.09	-0.09	**-0.15**	**0.12**	0.05	-0.08	**-0.04**	**-0.03**	0.1	0.0
GPT4-o	Orig-ZS	0.20	0.89	0.49	0.71	0.5	0.67	0.41	0.75	0.12	0.94
	Δ	**-0.55**	**0.43**	**-0.44**	**0.56**	**-0.32**	**0.37**	**-0.24**	**0.25**	**-0.19**	**0.11**

Table 3: Impact of Removing Relevant Text in the Essays Dataset: Recall values for 'Low' and 'High' classes are reported, with "Δ" indicating the difference between the evaluation settings. Bold values confirm Δ expectations.

that even for these models/traits, the recall scores for the "low" class are not perfect, suggesting significant potential for improvement. In prediction, three LLMs—Mistral, Llama3, and GPT-4o, satisfy the Δ criteria for at least four out of eight traits. However, in most cases, the magnitude of Δ is very low, and the overall RMSE is significantly high. Further, correlation analysis of decisions and scores by LLMs suggests that scoring is generally arbitrary (see Appendix A.5). These findings indicate that the *LLMs may assign personality trait scores to texts without substantial consideration of the actual personality-relevant content.*

4.3 Robustness of Evaluation

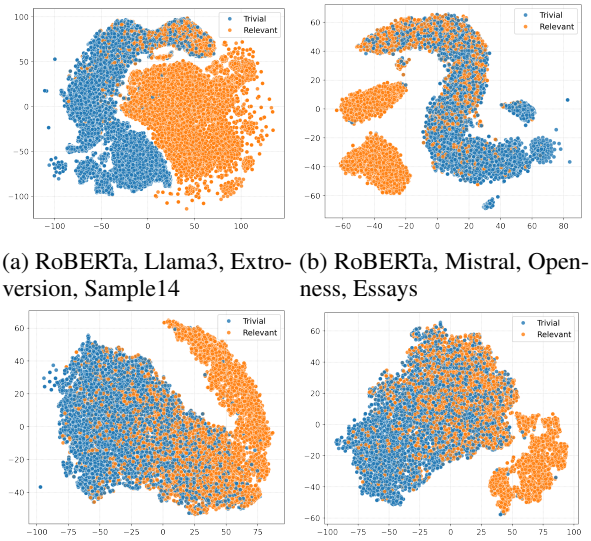
The above results suggest that perhaps LLMs show promise in utilizing their knowledge for zero-shot personality assessment, albeit for a select few LLMs. While comparisons in the previous section were based on the performance metrics (RMSE and Recall), related studies have shown that LLMs randomly change their decision at individual evaluation level (Yang et al., 2023; Shu et al., 2024). Thus, the results in the previous section could stem from this randomness. This variability could be attributed to factors such as prompt phrasing, the presentation order of traits/criteria, insufficient information, etc. Therefore, we investigated potential stability issues related to several such variables in both Orig-ZS and Trivial-ZS settings (see Appendix A.4). Our findings indicate that while LLMs modify their decisions nearly 20-40% of the time, the subsequent modifications do not consistently lead to improved performance.

This indicates that *the presence or absence of relevant text has little impact on the evaluations made by the LLMs*, corroborating the notion that LLMs may find it challenging to effectively apply their knowledge for zero-shot personality evaluation.

5 Discussions

Until now, we assumed that the text segments tagged by LLMs are personality-relevant and contain meaningful personality cues and that the presence or absence of these segments should impact subsequent evaluations.

We now shift our focus to critically examining whether the extracted text is genuinely distinct from irrelevant text and truly reflects personality-relevant content. This investigation is essential to



(a) RoBERTa, Llama3, Extroversion, Sample14

(b) RoBERTa, Mistral, Openness, Essays

(c) BERT, GPT4, Narcissism, Sample14

(d) BERT, OpenChat, Openness, Essays

Figure 6: t-SNE visualization examples of fine-tuned representations. The subfigure captions indicate Fine-tuning Transformer, LLM, Trait, Dataset

validate the usability and applicability of PsyTEx in effectively isolating and identifying personality-relevant information. To this end, we explore two key questions: Firstly, *Are there significant linguistic differences between relevant and trivial texts?* and, Secondly, *Does the LLM-extracted (tagged) text genuinely reflect personality-relevant content?* This section delves into these critical questions.

5.1 Evaluating the differences between Relevant and Trivial text

To assess the linguistic differences between trivial and relevant texts, we employ a straightforward method by fine-tuning transformer models, which have demonstrated SOTA performance across various NLP tasks. Implementation details for discriminating between trivial and relevant texts can be found in Appendix A.2.1. The results are evaluated using the Macro-F1 score, outlined in Table 6.

We observe average macro-F1 scores of 0.78 for BERT and 0.79 for RoBERTa, across all traits, LLMs, and datasets. These scores suggest significant linguistic differences between the two text groups. To further substantiate this finding, we performed qualitative validation by embedding the test sentences and visualizing the results using t-SNE projections (Van der Maaten and Hinton, 2008). Examples of this visualization are shown in Figure 6. The t-SNE projections demonstrate a *notable separation between the two groups, confirming the presence of linguistic differences.* The PsyTEx framework enables identification and tagging of text segments exhibiting linguistic separability.

5.2 Determining Personality-relevance of Relevant Text

We conduct a qualitative evaluation of the relevant text using the Linguistic Inquiry and Word Count (LIWC)[3] tool, a standard in psycholinguistics, to examine the relationship between psychological processes and language. Assessing the correlation of LIWC-captured psychological processes with GT trait score provides an opportunity to compare and validate characteristics of extracted relevant text with findings in Personality Psychology.

To alleviate any bias due to skew in score distribution within the dataset, we adopted a *Monte Carlo Simulation* protocol that selects one sample (with uniform probability) from each score (1-5) and calculates the Spearman Rank correlation between every LIWC category value (sum-normalized) and the trait scores. Each simulation is supported by 100,000 iterations to suppress potential instability in these correlations while only retaining statistically significant correlations ($p<0.01$). Finally, the average correlation across these iterations for each LLM-trait pair is calculated as a representative correlation value. Since this protocol necessitates trait scores, it was only performed on the Sample14 dataset.

Given the variability in LLM performance for the detection and prediction of specific traits, their ability to tag relevant text likely varies as well (see Appendix A.3). To evaluate whether LLMs generally identify personality-relevant text segments, we look for consensus among all models. The LIWC category correlation is valid if a minimum absolute correlation threshold of 0.5 is met for at least three LLMs. The median correlation from these LLMs is taken as the final representative correlation. The LIWC categories and their corresponding correlation coefficients, derived using this protocol, are presented in Tables 16 and 17 while the most informative LIWC categories sharing similarities with Psychology literature are presented in Table 4.

A considerable difference in the number of significant correlations between the Big Five and the Dark Triad traits is observed, supporting the earlier finding that LLMs struggle more with predict-

[3] https://www.liwc.app/

Trait	LIWC Categories	Explanation
Extroversion	P: socbehav, cogproc, comm, emo_pos	Tendencies for social behavior, interpersonal interactions, and positive emotional expressions
Agreeableness	P: polite, comm, tone_pos, emo_pos	Affiliative social orientation and general positive inclinations
Neuroticism	P: tentat, emo_anger, illness, conflict	Indecisiveness, excessive worry, hypersensitivity, and a propensity for conflict
Machiavellianism	P: swear; N: mental, home, need, family	Detachment from personality and emotional aspects of life and hostile demeanor
Narcissism	P: power, allnone, discrep, sexual; N: emo_anx	Need for dominance, grandiosity, assertiveness and aggressive self-presentation
Psychopathy	N: socbehav, tone_pos, insight	Lack of positive social interaction and positivity, impulsiveness or shallow thinking

Table 4: Significantly Correlated LIWC categories that share similarities with Psychology literature where. "P" and "N" represent Positive and Negative Correlations respectively. The analysis is limited to 77 categories under the broad categories "Psychological Processes" and "Expanded Dictionary" using LIWC-22

ing Dark Triad traits. However, the LIWC categories that correlate provide insights into specific linguistic patterns that may indicate these traits. The findings for both Dark Triad (Sumner et al., 2012; Holtzman et al., 2019) and Big Five (Yarkoni, 2010; Koutsoumpis et al., 2022; van der Vegt et al., 2022) are consistent with observations in existing Personality Psychology literature on trait-relevant language use. However, relying on aggregated LIWC categories for analysis can be overly broad and heavily dependent on the presence of specific words in the text, potentially invalidating correlations or preventing them from emerging if those words are absent. However, despite this limitation, *the alignment with relevant literature affirms that the relevant text effectively represents personality traits*, reinforcing PsyTEx as a valuable framework for isolating and amplifying psychological characteristics from the text.

6 Future Works

We plan to utilize the trait-relevant information identified in the PsyTEx framework for downstream personality assessment in two primary ways. Firstly, integrating attention mechanisms into existing personality detection models to focus on PsyTEx-refined text segments. These models can then be fine-tuned using existing personality detection datasets for effective assessment. However, a key limitation of this approach is the potential lack of representative data across various contexts, such as different topics, genres, or domains.

A strategy to overcome this limitation involves empowering LLMs to produce psychology-relevant insights. Efforts in this direction have included the development of taxonomies through expert-LLM teaming, categorizing information identified by LLMs into actionable insights (Shah et al., 2023). This method uses the precision of taxonomies with the LLM's ability to detect trait-relevant text instances, refined by expert analysis. We aim to refine and expand these ideas in our future work.

7 Conclusion

In this paper, we explore the question: *Can LLMs effectively "interpret" psychological characteristics from text?* To this end, we introduce a novel evaluation protocol called "Psychological Text Extraction and Refinement Framework" (PsyTEx), designed to assess the interpretive capabilities of LLMs for human categorization tasks, specifically for text-based personality recognition.

Using the simplest and most widely used LLM-based zero-shot personality evaluation, we first examine whether LLMs possess deep interpretive abilities. Our analysis of five SOTA LLMs and two personality models - Big Five and Dark Triad, revealed that LLMs frequently produce random and inconsistent outcomes regardless of the presence or absence of personality-relevant text, suggesting a lack of deep interpretive abilities. This was particularly evident in their struggle with more complex task of personality prediction and traits such as Dark Triad that require a nuanced understanding that goes beyond basic semantic processing. These results indicate that specifically tailored benchmarks are needed to evaluate LLM's interpretive abilities effectively. These benchmarks could significantly boost the efficacy of LLMs in areas such as mental health diagnosis, where a precise grasp of human psychology is essential.

While LLMs cannot be directly used to eval-

uate personality traits from the human-authored text in a zero-shot setting, our proposed framework enables them to extract personality-relevant information segments from the text. Our findings show that PsyTEx-refined text segments exhibit linguistic separability and capture meaningful patterns that align with personality psychology literature, validating its potential for enhancing personality assessment methodologies. Moreover, PsyTEx provides a foundation for downstream applications such as psychological modeling and taxonomy development, making it a valuable framework for text-based psychological analysis.

8 Limitations

We acknowledge three potential limitations in our study. Our protocol presumes that the SOTA LLMs used in this study and known for their competitive performance on standard benchmarks possess both relevant knowledge of Personality Psychology and the ability to effectively identify entailment between qualification criteria and text. However, manual review occasionally reveals instances where the text identified by the LLM as aligning with a qualification criterion actually contradicts it. Two such examples are provided below. Nonetheless, since either entailment or contradiction to certain criteria could indicate the presence or absence of a trait (for instance, the *presence* of empathy might suggest the *absence* of Psychopathy), we accept the textual evidence as valid even when the polarity of the entailment might be inverted.

ChatGPT Incorrectly Tagging Opposite Polarity
Qualification Criteria: Lack of Empathy **Text Evidence**: *"...I could tell it was taking a toll on my dad. He was hurting really bad and i wanted to help...i felt deeply for my dads pain... i wish he was still here in my life..."* **Justification**: The author exhibits empathy towards her father's feelings and mental state, indicating an awareness and understanding of his suffering.

ChatGPT's Failure to Gauge Intensity of Entailment
Qualification Criteria: Grandiosity **Text Evidence**: *"This is **my calling**, to help prevent girls and young boys from developing eating disorders.... **I know the early signs and behaviors** that developed mine and **I can now relate and apply** that to helping others."* **Justification**: The author has an elevated sense of their calling and believes they possess rare knowledge essential for helping others.

Additionally, in simulating the LLM's zero-shot evaluation process, we treat text tagged under all

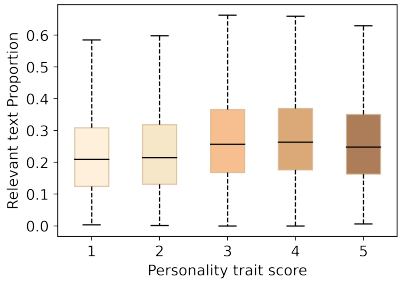


Figure 7: Proportion of tokens from original tagged as personality-relevant for Sample14 dataset

qualification criteria equally. It is possible, however, that LLMs may not weigh all criteria equally in their evaluations. Given the sub-optimal performance in detection and prediction under the original zero-shot (Orig-ZS) setting and observing little to no improvement before and after relevant text removal, we consider the importance of specific qualification criteria out of scope for this study.

Moreover, our findings indicate that personality is not uniformly represented across a text sample, as evidenced by a minimal correlation between trait scores and the proportion of personality-relevant text, as shown in Figure 7. Although this is a significant insight, our study does not account for other factors, such as the type of task that elicited the text. It is possible that certain prompts, like "Write about who you are", may evoke more personality-relevant responses than the Thematic Apperception Task. We plan to explore these dynamics in future research.

9 Ethics Statement

The primary objective of this study was to explore the limitations of LLMs in assessing personality traits from text data, aiming to encourage the development of applications that ethically and with proper permissions, evaluate human personality traits. However, we realize that the evaluation protocol introduced in this paper can be extended to assess the LLMs' capabilities for any psychological characteristics. To that end, we strongly discourage the application of our methodologies to develop LLMs that intend to covertly assess the psychological characteristics of humans without prior permission.

We secured the necessary permissions to use the Essays and Sample14 datasets, ensuring all user information was anonymized before being provided to us. We have been informed that appropriate

permissions were obtained from the participants contributing to these datasets for the use of text data explicitly for research purposes. We have rigorously adhered to the data usage policies specified for these datasets.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

AI@Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mostafa M Amin, Rui Mao, Erik Cambria, and Björn W Schuller. 2023. A wide evaluation of chatgpt on affective computing tasks. *arXiv preprint arXiv:2308.13911*.

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, Gregory D Webster, and Damon Woodard. 2024a. Bridging minds and machines: Unmasking the limits in text-based automatic personality recognition for enhanced psychology–ai synergy. *British Journal of Psychology*.

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, and Damon Woodard. 2024b. Emulating author style: A feasibility study of prompt-enabled text stylization with off-the-shelf llms. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 76–82.

Angela L Carey, Melanie S Brucks, Albrecht CP Küfner, Nicholas S Holtzman, Mitja D Back, M Brent Donnellan, James W Pennebaker, Matthias R Mehl, et al. 2015. Narcissism and the use of personal pronouns revisited. *Journal of personality and social psychology*, 109(3):e1.

Jonathan Cook, Tim Rocktäschel, Jakob Foerster, Dennis Aumiller, and Alex Wang. 2024. Ticking all the boxes: Generated checklists improve llm evaluation and generation. *arXiv preprint arXiv:2410.03608*.

Google DeepMind. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Adithya V Ganesan, Yash Kumar Lal, August Nilsson, and H Schwartz. 2023. Systematic evaluation of gpt-3 for zero-shot personality estimation. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 390–400.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al. 2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR.

Reto Gubelmann, Aikaterini-Lida Kalouli, Christina Niklaus, and Siegfried Handschuh. 2023. When truth matters-addressing pragmatic categories in natural language inference (nli) by large language models (llms). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 24–39.

Eddie Guo, Mehul Gupta, Jiawen Deng, Ye-Jean Park, Michael Paget, and Christopher Naugler. 2024. Automated paper screening for clinical reviews using large language models: Data analysis study. *Journal of Medical Internet Research*, 26:e48996.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Nicholas S Holtzman, Allison M Tackman, Angela L Carey, Melanie S Brucks, Albrecht CP Küfner, Fenne Große Deters, Mitja D Back, M Brent Donnellan, James W Pennebaker, Ryne A Sherman, et al. 2019. Linguistic markers of grandiose narcissism: A liwc analysis of 15 samples. *Journal of Language and Social Psychology*, 38(5-6):773–786.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. Llm vs small model? large language model based text augmentation enhanced personality detection model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16, pages 18234–18242.

Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Lee. 2023. Who wrote it and why? prompting large-language models for authorship verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14078–14084.

Yu Ji, Wen Wu, Hong Zheng, Yi Hu, Xi Chen, and Liang He. 2023. Is chatgpt a good personality recognizer? a preliminary study. *arXiv preprint arXiv:2307.03952*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

Antonis Koutsoumpis, Janneke K. Oostrom, Djurre Holtrop, Ward van Breda, Sina Ghassemi, and Reinout E. de Vries. 2022. The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the big five and the linguistic inquiry and word count (liwc). *Psychological Bulletin*, 148(11–12):843–868.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2024. Big5-chat: Shaping llm personalities through training on human-grounded data. *arXiv preprint arXiv:2410.16491*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2020. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, 53(4):2313–2339.

Sumiya Mushtaq and Neerendra Kumar. 2022. Text-based automatic personality recognition: Recent developments. In *Proceedings of Third International Conference on Computing, Communications, and Cyber-Security: IC4S 2021*, pages 537–549. Springer.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

James W Pennebaker and Laura A King. 1999. Linguistic styles: Language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7.

Chirag Shah, Ryen W White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Snigdha Sarathi Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Xiaochuan Ni, et al. 2023. Using large language models to generate, validate, and apply user intent taxonomies. *arXiv preprint arXiv:2309.13063*.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Chris Sumner, Alison Byers, Rachel Boochever, and Gregory J Park. 2012. Predicting dark triad personality traits from twitter usage and a linguistic analysis of tweets. In *2012 11th international conference on machine learning and applications*, volume 2, pages 386–393. IEEE.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Isabelle van der Vegt, Bennett Kleinberg, and Paul Gill. 2022. Predicting author profiles from online abuse directed at public figures. *Journal of threat assessment and management*, 9(1):17.

Huy Vu, Huy Anh Nguyen, Adithya V Ganesan, Swanie Juhng, Oscar NE Kjell, Joao Sedoc, Margaret L Kern, Ryan L Boyd, Lyle Ungar, H Andrew Schwartz, et al. 2024. Psychadapter: Adapting llm transformers to reflect traits, personality and mental health. *arXiv preprint arXiv:2412.16882*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Rethinking sts and nli in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 965–982.

Zhiyuan Wen, Yu Yang, Jiannong Cao, Haoming Sun, Ruosong Yang, and Shuaiqi Liu. 2024. Self-assessment, exhibition, and recognition: a review of personality in large language models. *arXiv preprint arXiv:2406.17624*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023. Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373.

Weizhe Yuan, Pengfei Liu, and Matthias Gallé. 2024. LLMCrit: Teaching large language models to use criteria. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7929–7960, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2024. A systematic survey of text summarization: From statistical methods to large language models. *arXiv preprint arXiv:2406.11289*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Xiaoming Zhao, Zhiwei Tang, and Shiqing Zhang. 2022. Deep personality trait recognition: a survey. *Frontiers in Psychology*, 13:839619.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

A Appendix

Standard Prompt. $$$ListOfTraits$$$ is replaced with the list of traits to assess and $$$InsertAuthorText$$$ is replaced with the Author's text.

<Task Description>
You are an AI assistant specializing in text analysis. Your task is to assess the personality traits of the author based on the provided essay. The following personality traits should be evaluated: $$$ListOfTraits$$$

<Instructions>
For each trait, predict the author's personality trait score on a scale of 1 to 5, indicating the level of trait presence where 1 = very low, 5 = very high. Additionally, determine whether the author is more likely to be A-Low or B-High in each trait based on your evaluation. Provide a justification for each assessment.
Before beginning your reponse, add the marker "$$–Start of Response–$$". Please adhere to the exemplary python dictionary (JSON) format below for generating output. Ensure that formatting of the output is strictly followed without adding any additional text.

<Output Format>
"<trait1>":
"score":<score>,
"decision":"<A or B>",
"explanation":"<justification>"
,
"<trait2>":
"score":<score>,
"decision":"<A or B>",
"explanation":"<justification>"

<Input>
Author's Text: $$$InsertAuthorText$$$

Relevant Information Extraction Prompt. $$$InsertCriterias$$$ is replaced with the list of criteria specific to LLM/trait and $$$InsertAuthorText$$$ is replaced with the Author's text

<Task Description>
Consider the following essay response carefully and evaluate each of the qualification criteria from the following list. Please refrain from making assumptions about the relevance of these qualifications to any specific personality trait(s) and disorder(s) and base your evaluations with utmost objectivity purely on the essay. When encountered, provide all relevant textual evidence of each criteria and how it manifests in the text. Finally present a summary of your overall findings.

Criteria:
$$$InsertCriteria$$$

<Instructions>
Before beginning your response, add the marker "$$–Start of Response–$$". Please adhere to the exemplary python dictionary (JSON) format below for generating output. Ensure that formatting of the output is strictly followed without adding any additional text.

<Output format>
{
"<criteria-A>": {
"text evidence": ["<text evidence1>","<text evidence2>",...,"<text_evidenceN>"],
"description":"<explanation of manifestation>"
},
"<criteria-B>": {
"text evidence": ["<text evidence1>","<text evidence2>",...,"<text_evidenceN>"],
"description":"<explanation of manifestation>"
},
"summary":"<summary>",
}

<Input>
Essay: $$$InsertAuthorText$$$

Knowledge Extraction Prompt

According to your knowledge, how is the personality trait P manifested in text? Can you give me an exhaustive list of textual manifestations of P in the order of importance and relevance to the Personality Psychology literature?

<Instructions>
For each instance, please provide a short explanation in a line-separated field under the title "Description:" along with a few examples of the textual manifestation in the form of phrases or sentences in a line-separated field under the title "Examples".

A.1 Dataset Details

Sample14

This dataset includes data from 1,126 subjects and provides scores for two personality models: the Big Five (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) and the Dark Triad (Machiavellianism, Narcissism, Psychopathy), encompassing a total of eight traits and over 3,400

text samples. Subjects participated in three different tests: writing a stream-of-consciousness essay, responding to "Write about who you are" and completing a Thematic Apperception Test (Carey et al., 2015). On average, the text samples contained 3,773 characters, 829 words, and 48 sentences each.

Essays

The Essays dataset is considered the gold-standard corpus with Big Five binary labels (Low/High). The dataset includes over 2,400 text samples from subjects who were required to write a stream-of-consciousness (SOC) essay for 10 consecutive days and 20 minutes each day (Pennebaker and King, 1999). On average, the text samples contained 3,296 characters, 743 words, and 46 sentences each.

A.2 Implementation Details

While most research in zero-shot personality evaluation primarily focuses on the latest iterations of ChatGPT, the landscape of LLMs has expanded significantly, introducing a variety of models that often surpass ChatGPT in performance across numerous tasks and benchmarks. To broadly assess whether LLMs can interpret personality, our study incorporates a diverse set of both proprietary and open-source LLMs. Specifically, we utilize five models: Mistral-7B (`Mistral-7B-Instruct-v0.3`), OpenChat-7B (`openchat/openchat_3.5`), Phi3-14B (`microsoft/Phi-3-medium-128k-instruct`), Llama3-8B (`meta-llama/Meta-Llama-3-8B-Instruct`), and the latest from OpenAI, GPT-4o (`gpt-4o`) (Jiang et al., 2023; Wang et al., 2023; Abdin et al., 2024; AI@Meta, 2024; OpenAI, 2023). This selection aims to provide a comprehensive overview of the current capabilities and limitations of LLMs in interpreting personality from text.

The HuggingFace model repository[4] was used to access all open-source models, while the openAI API[5] was used for accessing the GPT-4o model. We have accepted and complied with all the usage policies for these LLMs. NVIDIA A100 Tensor Core GPUs were used for generating data from the open-source LLMs approximating 504 GPU hours.

For consistency in text generation across LLMs, top-K and top-p (nucleus) sampling with K=50 and p=0.95 is used as a decoding strategy wherever applicable. No preprocessing was performed on the author texts before being used as input to the LLMs. The detailed task descriptions, instructions, as well as output formatting requirements for each phase are outlined in the above text boxes.

The LLMs are instructed to generate output in Python JSON format. However, deviations from this format occasionally occur, leading to the addition or removal of content. To address these inconsistencies, a text-JSON extractor[6] was used to extract structured data from outputs generated by LLMs.

We used NLTK[7] to perform sentence tokenization wherever required. For generating the t-SNE projections, the scikit-learn[8] package was used while setting the perplexity to 30. For the LIWC-22[9] software, we have obtained an academic non-commercial license for research purposes.

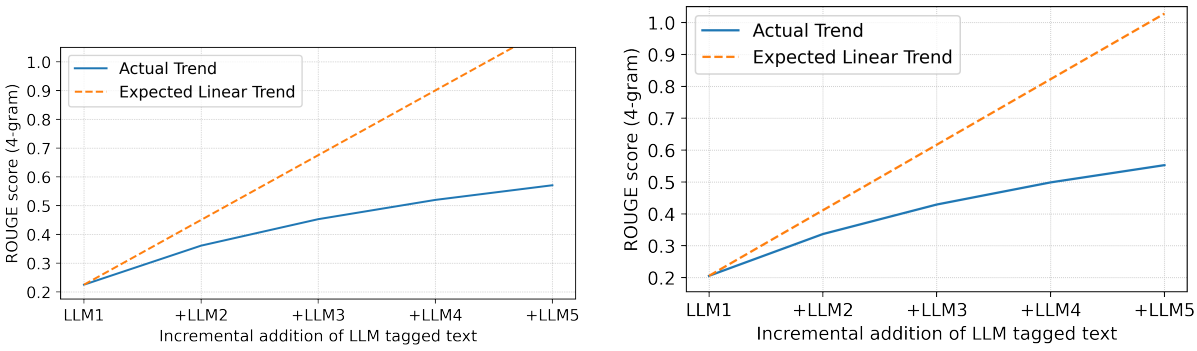
A.2.1 Finetuning Implementation

As the relevant text typically consists of sentence-like chunks, we begin by sentence tokenizing the trivial text. Following a 70:30 training to testing split, we fine-tune two transformer models, BERT and RoBERTa, on an equal number of randomly sampled sentences from both groups. The number of max_tokens and the number of epochs are set to 64 and 5, respectively.

A.3 Stability of Relevant Information

Building on the qualitative analysis suggesting that LLM-tagged "relevant" text chunks are crucial for personality assessment, it is vital to examine the information density of the original texts identified as relevant by different LLMs. This assessment will help determine the consistency with which LLMs

[4]https://huggingface.co/models
[5]https://platform.openai.com/docs/models
[6]https://github.com/mangiucugna/json_repair
[7]https://www.nltk.org/
[8]https://scikit-learn.org/stable/
[9]https://www.liwc.app/



(a) Essays

(b) Sample14

Figure 8: Variation in ROUGE Score Between Original and Relevant (LLM-tagged) Text with Incremental Inclusion of LLMs

identify similar text segments as trait-relevant, thereby evaluating the stability of relevance tagging across models. Ideally, if all LLMs are equally proficient at tagging relevant information, the density of information in the tagged segments should reach a saturation point.

To conduct this analysis, we measure the textual overlap between the segments tagged as relevant by the LLMs compared to the original text for each text sample using the ROUGE score with 4-grams. We start with a single LLM and incrementally one LLM at a time. As each new LLM is added, we combine the text chunks they have tagged, ensuring that no text chunks are repeated and just unique segments are retained. After each addition, we calculate the ROUGE score between the aggregated common text chunks tagged by the LLMs and the original text. As the order of adding LLMs influences the ROUGE scores, we evaluate all 120 permutations of the 5 LLMs and plot the average ROUGE score from all permutations in Figure 8.

If LLMs randomly tag different text chunks from the original texts regardless of the provided qualification criteria, we would expect the ROUGE score to increase linearly (as marked by a dashed line). However, we observe that the ROUGE score tends to saturate below a score of 0.6 for both datasets. This observation indicates two key points: First, there is a significant overlap in the texts commonly tagged by all LLMs, demonstrating their ability to identify personality-relevant text. Second, not all LLMs tag the same segments, suggesting that multiple LLMs may be necessary to ensure reliable tagging of personality-relevant information. However, the relevance of the combined information from multiple LLMs remains to be evaluated independently and is beyond the scope of this paper.

A.4 Prompt Stability Analysis

A well-known limitation of LLMs is their sensitivity to minor variations in prompts (Shu et al., 2024). In our study, we utilize LLMs for personality assessments using two approaches: standard zero-shot prompting (Orig-ZS) and zero-shot prompting following our PsyTEx framework (Trivial-ZS). Given this, it is crucial to evaluate the impact of prompt variations on both pipelines.

For our stability analysis, we randomly selected 100 text samples from each dataset. We then performed evaluations using both Orig-ZS and Trivial-ZS, applying the same prompts as outlined in the paper to establish a baseline for comparison. For each prompt variation considered, we assess its effect through two metrics: performance difference and unchanged rate. The performance difference measures changes at the overall performance level, while the unchanged rate examines changes at the individual decision level. These metrics are crucial for determining whether the variations in LLM evaluations and decisions are responses to changes in the prompts.

Given the resource-intensive nature of the stability analysis experiments and the high cost of using closed-source models, coupled with the observation that closed-source models performed similarly to open-source models, we opted to conduct these experiments exclusively with open-source models for efficiency.

A.4.1 Evaluation Metrics

Performance Difference

To maintain consistency with the paper and facilitate comparison, we measure performance difference by calculating the difference between the default setup to the prompt variation experiment. For Essays, this is represented by ΔF1-score and for Sample14 by ΔRMSE.

In the Orig-ZS setting, where accurate personality assessment is the goal, a positive effect of prompt variation is indicated by $\Delta < 0$ for Essays and $\Delta > 0$ for Sample14. Conversely, in the Trivial-ZS setting, which tests the LLM's performance in response to the removal of relevant information, a positive effect is shown by $\Delta > 0$ for Essays and $\Delta < 0$ for Sample14.

It is important to highlight that interpreting ΔRMSE is different from ΔF1-score. F1 scores are bounded between 0 and 1, so a ΔF1-score of 0.25 would represent a significant 25% shift in performance. However, RMSE values are unbounded, and in our case, where RMSE can range from 0 to 4, a difference of 0.25 in RMSE does not necessarily reflect a significant change in performance.

Unchanged Rate

Following and extending the re-test protocol by Yang et al. (2023), we quantify the impact of prompt variation on personality assessment by calculating the unchanged rate, $\hat{y}^i$, across the 100 samples. In the case of the Essays dataset (binary classification task), the unchanged rate refers to the number of predictions that remain the same.

There is no standard method for calculating the unchanged rate from continuous values like trait scores. Therefore, for Sample14, we slightly modify the problem to enable the calculation of the unchanged rate. First, we convert the trait scores into three broad categories: "low" for scores below 3, "high" for scores above 3, and "neutral" for scores equal to 3. We then check whether the predicted scores from the prompt variation experiment fall within the same category as those from the default baseline. If the predicted score remains in the same category, we consider the decision unchanged. Finally, similar to the Essays dataset, we calculate the proportion of samples that remain unchanged.

A low unchanged rate suggests that the prompt variation has significantly altered the predictions made by the LLM.

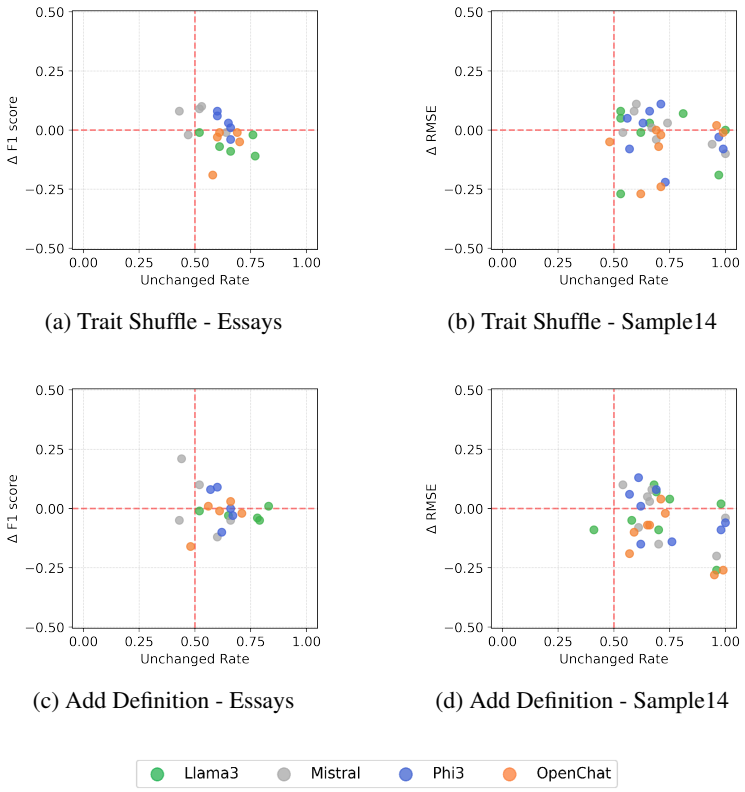
A.4.2 Standard Prompting Pipeline

In the standard prompting protocol, personality prediction relies entirely on the default prompt. To investigate potential factors that could cause LLMs to produce varying outcomes, we explore two specific scenarios. First, our protocol assumes that LLMs inherently understand personality traits and their definitions. However, when humans are tasked with annotating personality-related data, they are typically provided with definitions for each trait to guide the annotation process. Thus, incorporating these personality definitions into the prompt could potentially provide LLMs with additional context and improve their personality prediction or detection capabilities. For this first prompt variation, we add the trait definitions directly into the prompt. The definitions are borrowed and constructed from various Psychology literature as well as with the help of expert knowledge. These definitions are presented in Table 10.

Second, we examine whether the order in which personality traits are presented affects the model's predictions. Specifically, we shuffle the sequence of traits (represented by the variable $$ListOfTraits$$ in the standard prompt) to assess any impact on performance. The effects of these prompt variations are then compared to the baseline Orig-ZS performance. The detailed results are presented in Table 7 and depicted in Figure 9.

Varying the Trait Order:

From Figures 9a and 9b, it is evident that the Δ remains close to 0, with an unchanged rate around 0.6 for Essays and 0.75 for Sample14. Largely, the LLMs do not exhibit the expected positive trend. The detailed tables show only a few cases where Δ is high and in a desirable direction, such as Agreeableness for OpenChat on Essays, and Openness or Extraversion for Mistral on Sample14. Additionally, while there is some performance variation between the two runs of trait order shuffling, this variability does not consistently lead to positive outcomes and varies across LLMs and traits. Overall, altering the order in which traits are presented appears to have minimal impact on personality recognition performance.



(a) Trait Shuffle - Essays

(b) Trait Shuffle - Sample14

(c) Add Definition - Essays

(d) Add Definition - Sample14

Figure 9: Impact of Prompt Variation on the Standard Prompting Pipeline (Orig-ZS). For positive influence of prompt variation, the following is desired: Unchanged rate<0.5, ΔF1-score<0, and ΔRMSE>0

Adding Trait Definition:

It was anticipated that adding definitions to the prompt would improve personality recognition performance, but the overall trend observed in Figures 9c and 9d suggests otherwise. While there is a greater spread in values compared to the earlier trait shuffling results, it does not imply better performance. The trend for the Essays dataset moves in the opposite direction of expectations, although Sample14 shows some promise, albeit low in magnitude. The average unchanged rate for Essays remains at 0.62, with an average ΔF1 of -0.01, while for Sample14, the unchanged rate is 0.73, with an average ΔRMSE of -0.05. Desirable outcomes were observed in a few instances, such as Openness for Sample14 and Agreeableness for OpenChat, but similar to previous results, the performance changes are not significant enough to justify further investigation.

A.4.3 PsyTEx Framework

In the PsyTEx framework, multiple factors can influence personality prediction performance, beginning with the knowledge extraction phase. We introduce prompt variations at each stage of this process to assess their impact. The effects of these prompt variations are then compared against the baseline Trivial-ZS performance. Detailed results for various prompt modifications are presented in Table 8

Effect of Knowledge Extraction Prompt Phrasing

Since the qualification criteria generated during the knowledge extraction phase influence the final Trivial-ZS performance, we begin by exploring several variations in the knowledge extraction prompt. Specifically, we create four different versions of the prompt and evaluate the pairwise semantic similarity of the resulting qualification criteria against the default prompt used in our main experiments. The variations of the Knowledge Extraction Prompt are presented in Table 9.

For this analysis, we employed the `multi-qa-mpnet-base-dot-v1` model from the SentenceTransformers[10] library, which is optimized for semantic search. We began by conducting a semantic search on the criteria generated from the default prompt to establish a baseline. For each criterion, we recorded

[10]https://sbert.net/

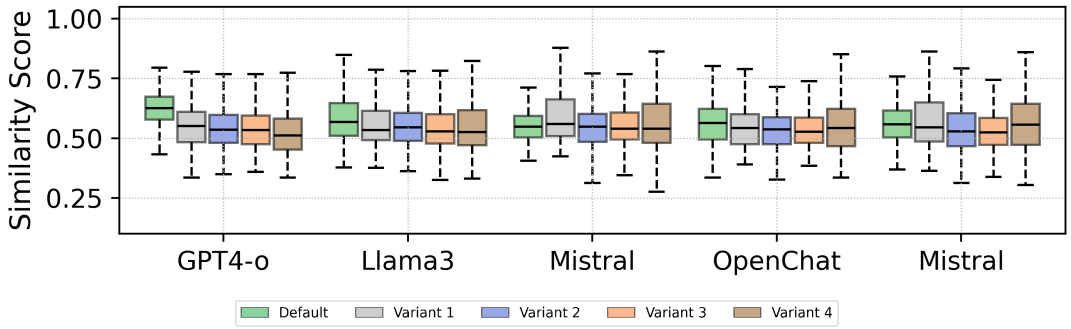


Figure 10: Semantic Similarity between Default and Knowledge Extraction Prompt variants

the similarity scores for the top three semantically similar criteria. This process was conducted for all criteria, ensuring that the criterion being analyzed was excluded from the comparison set to avoid biasing the results.

Following this baseline establishment, we compared the criteria from the default prompt to those from each prompt variant using the same methodology. The collective semantic similarities for all traits associated with a specific LLM were compiled and illustrated in Figure 10. Analysis of this data reveals that the spread of semantic similarities is consistent across different prompt variations, suggesting that the variation in prompt phrasing has minimal impact on the criteria generated by the LLMs.

Varying the Criteria Order

Similar to the trait order shuffling experiment, here we shuffle the order of the criteria that are presented to the LLM to tag relevant information. The criteria are shuffled twice, and the results from both runs are shown in Figures 11a and 11b. The Δ for both experiments remains close to 0, indicating little to no change in performance compared to the default Trivial-ZS setting. Additionally, the unchanged rate consistently stays above 0.5, suggesting that the order in which the criteria are presented has minimal impact on LLM evaluations and Trivial-ZS performance.

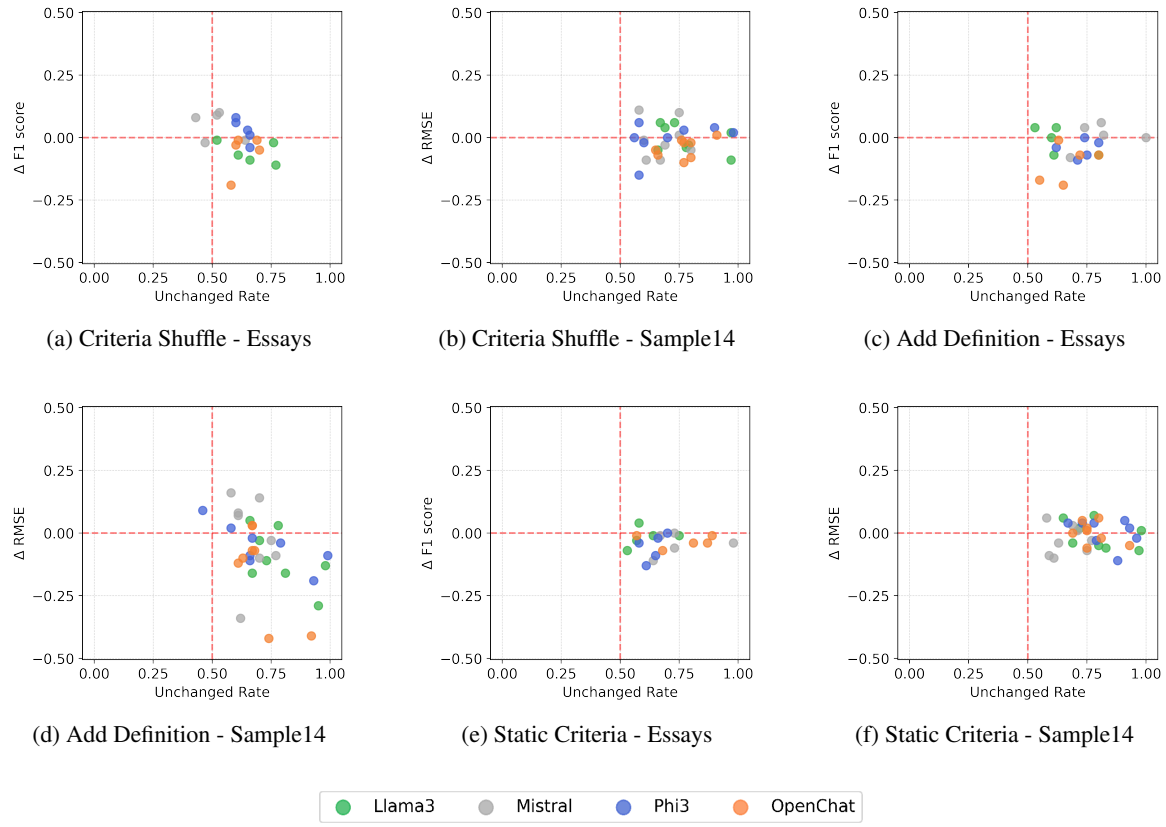
Adding Trait Definition

As with the Standard prompting pipeline, we incorporate trait definitions during the Trivial-ZS evaluation, with the key difference being that each trait is assessed individually. Observing the Figures 11c and 11d indicates that adding definitions leads to marginal performance improvements for some LLMs on the Sample14 dataset, while the Essays dataset shows an opposite trend to expectations. For instance, OpenChat shows the desired trend (ΔRMSE$<$0) for 7 out of 8 traits, although the magnitude of Δ varies across traits. However, it's important to remind readers that ΔRMSE cannot be interpreted in the same way as ΔF1-score. While the observed performance variation in the expected direction suggests that incorporating personality definitions into standard prompts may aid in personality recognition, the lack of a similar trend in the Essays dataset, combined with the fact that ΔRMSE is relative to the default value, complicates this interpretation. If the default performance is poor, even small changes can appear as improvements. Therefore, based on these results, a strong case cannot be made for using personality definitions in the prompts.

Providing Static Qualification Criteria

In the PsyTEx framework, we advocate using qualification criteria extracted independently from each LLM through the relevant information extraction prompts. This approach is driven by two key reasons. First, the process is designed to be generalizable, ensuring that even without prior knowledge of the psychological characteristic being assessed, the framework remains effective. While personality traits are well-studied in psycholinguistics, and we have predefined qualification criteria for them, this may not be the case for less established concepts, such as intent. In such instances, we may lack predefined criteria to guide LLMs in text segmentation.

Second, by relying on qualification criteria generated by the LLM itself, we assume that the model possesses both the relevant knowledge of the criteria and the ability to recognize it in text. However, it is worth considering what would happen if the qualification criteria were standardized across all LLMs. To



(a) Criteria Shuffle - Essays (b) Criteria Shuffle - Sample14 (c) Add Definition - Essays

(d) Add Definition - Sample14 (e) Static Criteria - Essays (f) Static Criteria - Sample14

Figure 11: Impact of Prompt Variation on the Trivial-ZS performance. For positive influence of prompt variation, the following is desired: Unchanged rate$<$0.5, ΔF1-score$>$0, and ΔRMSE$<$0

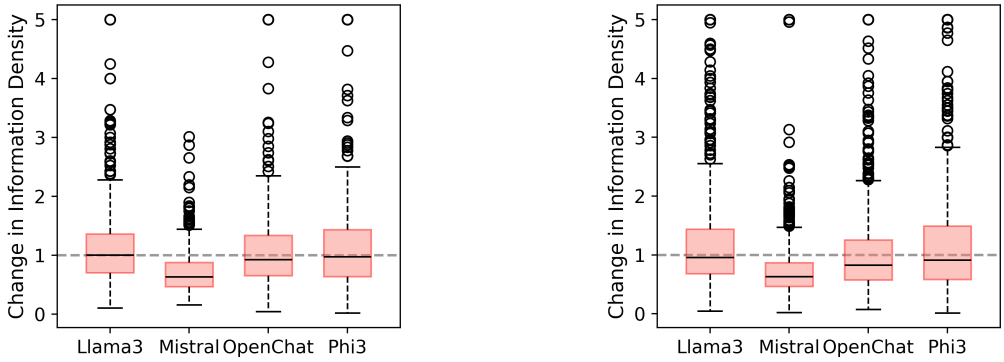
explore this, we combined the qualification criteria generated by each LLM for a specific trait, creating an all-inclusive list of criteria. We then removed any redundant phrasing and applied Trivial-ZS to assess the impact.

The hypothesis is that using a complete set of criteria will not only influence Trivial-ZS performance but also affect the text tagged as relevant by the LLMs. Ideally, this more extensive list should capture all relevant personality-related information, increasing the density of information captured from the original text samples. Thus, in addition to evaluating the impact on Trivial-ZS performance, a secondary goal is to examine changes in tagged information density. This is measured by comparing the ratio of tokens tagged by the LLM using the default setting to those tagged using the comprehensive criteria list.

If the comprehensive list increases information density, the ratio will be less than 1, indicating that more personality-relevant information was identified. However, a notable reduction in Trivial-ZS performance should also be observed indicating that the removal of information tagged using the comprehensive criteria list affects LLM personality evaluation performance. The results of this experiment aggregated for all traits for an LLM are presented Figure 12. To facilitate interpretation, the y-axis has been capped at 5.

From the information density analysis, we observe that while the median density ratio hovers around 1, indicating that both the default and comprehensive prompts produce similar token counts, the upper whiskers and outliers (>1) suggest that, in general, the default prompt tags more words. This reinforces two key points: first, the LLM-generated qualification criteria extracted using the Knowledge Extraction Prompt are valid, and second, introducing unfamiliar criteria can reduce the LLM's ability to identify relevant information, likely leading to confusion.

Moreover, the results show minimal to no change in Trivial-ZS evaluation depicted in Figures 11e and 11f, indicating that providing a static, all-inclusive list of qualifications does not improve the models' ability to tag personality-relevant information and, consequently, does not affect the LLM's performance in Trivial-ZS evaluations.



(a) Essays

(b) Sample14

Figure 12: Comparison of Information Density: Proportion of tokens tagged using the default criteria versus those tagged by the static comprehensive criteria list. Values less than 1 indicate more personality-relevant information identified using static criteria list.

A.5 Performance variation between detection and prediction

In Section 4.1, we observed that LLMs performed relatively better for binary classification (detection) tasks than in the fine-grained task of assigning personality scores (prediction). This disparity may stem from the LLMs' insufficient nuanced understanding of personality traits, which could lead to seemingly arbitrary assignments of trait scores. Consequently, it is crucial to evaluate whether LLMs accurately understand and respond to both the tasks - labeling of traits (high or low) and the assignment of numerical trait scores.

To evaluate the consistency of LLM outputs, we conducted statistical tests assessing the stability between binary decisions (labels) and assigned scores (ranging from 1 to 5) from the Orig-ZS evaluations. Following Yang et al. (2023), we computed the Spearman Rank correlation coefficient between the decision labels and scores. For additional validation, we also calculated the point-biserial correlation coefficient to examine the relationship between these binary and continuous outputs. The results of these tests are presented in Table 5 which will illuminate the extent to which LLMs comprehend the task and follow instructions.

LLM	O		C		E		A		N		Mach		Narc		Psyc	
	PB[1]	SR[2]	PB	SR	PB	SR	PB	SR	PB	SR	PB	SR	PB	SR	PB	SR
Essays																
Mistral	0.36	0.34	0.1	0.1	0.36	0.38	-0.07	-0.04	0.58	0.6						
Llama3	0.61	0.63	0.82	0.8	0.89	0.9	0.54	0.55	0.83	0.83						
OChat[3]	0.57	0.56	0.45	0.47	0.41	0.43	0.14	0.15	0.75	0.73						
Phi3	0.42	0.42	0.49	0.48	0.59	0.58	0.18	0.19	0.68	0.67						
GPT4	0.58	0.6	0.86	0.81	0.87	0.82	0.76	0.76	0.53	0.65						
Avg	0.51	0.51	0.54	0.53	0.62	0.62	0.31	0.32	0.67	0.70						
Sample14																
Mistral	0.47	0.44	0.31	0.3	0.36	0.38	0.12	0.12	0.71	0.71	0.49	0.55	0.6	0.58	0.52	0.62
Llama3	0.67	0.67	0.8	0.78	0.87	0.87	0.69	0.69	0.87	0.85	0.39	0.65	0.71	0.72	0.78	0.94
OChat	0.63	0.61	0.69	0.66	0.5	0.51	0.41	0.4	0.79	0.76	0.23	0.31	0.36	0.4	0.22	0.3
Phi3	0.38	0.39	0.45	0.46	0.55	0.54	0.2	0.23	0.68	0.67	0.44	0.6	0.52	0.52	0.48	0.62
GPT4	0.41	0.46	0.62	0.68	0.87	0.81	0.73	0.77	0.67	0.73	0.6	0.76	0.86	0.82	0.6	0.82
Avg	0.51	0.51	0.57	0.58	0.63	0.62	0.43	0.44	0.74	0.74	0.43	0.57	0.61	0.61	0.52	0.66

[1] Point Biserial Correlation; [2] Spearman Rank Correlation; [3] OpenChat

Table 5: Decision to Label Correlation obtained from Orig-ZS evaluations. All the correlations are significant at p-value<0.01.

LLM	O		C		E		A		N		Mach		Narc		Psyc	
	BT[1]	RoB[2]	BT	RoB	BT	RoB	BT	RoB	BT	RoB	BT	RoB	BT	RoB	BT	RoB
							Essays									
Mistral	0.79	0.80	0.81	0.82	0.78	0.80	0.78	0.79	0.8	0.82						
Llama3	0.83	0.84	0.83	0.84	0.92	0.93	0.81	0.82	0.81	0.82						
OChat[3]	0.75	0.77	0.76	0.78	0.76	0.78	0.75	0.77	0.77	0.79						
Phi3	0.73	0.74	0.75	0.77	0.76	0.78	0.75	0.76	0.76	0.78						
GPT4	0.79	0.81	0.76	0.79	0.77	0.79	0.82	0.83	0.79	0.81						
Avg	0.78	0.79	0.78	0.80	0.80	0.82	0.78	0.79	0.79	0.80						
							Sample14									
Mistral	0.78	0.79	0.8	0.81	0.79	0.81	0.77	0.79	0.8	0.82	0.77	0.79	0.75	0.77	0.76	0.78
Llama3	0.8	0.82	0.81	0.82	0.92	0.93	0.8	0.82	0.81	0.82	0.78	0.79	0.81	0.82	0.8	0.82
OChat	0.74	0.74	0.74	0.75	0.75	0.77	0.74	0.75	0.77	0.78	0.73	0.74	0.76	0.76	0.74	0.75
Phi3	0.72	0.74	0.75	0.77	0.75	0.77	0.74	0.76	0.77	0.78	0.74	0.75	0.75	0.77	0.73	0.74
GPT4	0.78	0.8	0.78	0.8	0.77	0.8	0.81	0.83	0.81	0.83	0.76	0.77	0.8	0.81	0.77	0.79
Avg	0.76	0.78	0.78	0.79	0.79	0.81	0.77	0.79	0.79	0.80	0.76	0.77	0.77	0.79	0.76	0.78

[1] BERT-Finetuned; [2] RoBERTa-Finetuned; [3] OpenChat

Table 6: Macro-F1 scores from Transformers finetuned to discriminate between relevant and trivial text

The results indicate that while most LLM-trait pairs exhibit a significant positive correlation, the degree of correlation varies significantly both within and across different LLMs. On average, the correlation across LLMs and traits is approximately 0.5, indicating considerable inconsistency in how LLMs assign scores and make decisions. This variability suggests that the range of scores the LLMs use to label a text sample as ‘high” and “low” for a certain trait may change significantly or that these models simply assign random trait scores or labels. Notably, the variation in decision-to-score stability also differs among models; for instance, Mistral exhibits the lowest overall stability, whereas Llama3 and GPT-4o demonstrate the highest. These observations suggest that certain LLMs may be more adept at adhering to instructions, a capability that could potentially extend to their effectiveness in recognizing personality traits. Future studies should investigate this hypothesis- exploring whether some LLMs are inherently better suited to identify particular traits than others.

Essays											
LLM	Trait	Default	Definition			Trait-Shuffle1			Trait-Shuffle2		
		F1	F1	UC	ΔF1	F1	UC	ΔF1	F1	UC	ΔF1
Llama3	O	0.42	0.41	0.83	0.01	0.39	0.82	0.03	0.53	0.77	-0.11
	C	0.47	0.5	0.65	-0.03	0.55	0.57	-0.08	0.54	0.61	-0.07
	E	0.48	0.52	0.78	-0.04	0.61	0.63	-0.13	0.57	0.66	-0.09
	A	0.55	0.56	0.52	-0.01	0.58	0.52	-0.03	0.56	0.52	-0.01
	N	0.56	0.61	0.79	-0.05	0.49	0.82	0.07	0.58	0.76	-0.02
Mistral	O	0.55	0.45	0.52	0.1	0.52	0.57	0.03	0.46	0.52	0.09
	C	0.61	0.4	0.44	0.21	0.56	0.57	0.05	0.51	0.53	0.1
	E	0.53	0.65	0.6	-0.12	0.63	0.49	-0.1	0.55	0.47	-0.02
	A	0.48	0.53	0.43	-0.05	0.58	0.45	-0.1	0.4	0.43	0.08
	N	0.49	0.54	0.66	-0.05	0.36	0.77	0.13	0.5	0.64	-0.01
OpenChat	O	0.54	0.53	0.56	0.01	0.57	0.67	-0.03	0.57	0.6	-0.03
	C	0.49	0.5	0.61	-0.01	0.54	0.56	-0.05	0.5	0.61	-0.01
	E	0.61	0.58	0.66	0.03	0.55	0.65	0.06	0.6	0.69	0.01
	A	0.37	0.53	0.48	**-0.16**	0.59	0.48	**-0.22**	0.56	0.58	**-0.19**
	N	0.57	0.59	0.71	-0.02	0.53	0.74	0.04	0.61	0.7	-0.04
Phi3	O	0.51	0.61	0.62	-0.1	0.47	0.71	0.04	0.48	0.65	0.03
	C	0.54	0.45	0.6	0.09	0.58	0.64	-0.04	0.53	0.66	0.01
	E	0.61	0.53	0.57	0.08	0.55	0.5	0.06	0.53	0.6	0.08
	A	0.58	0.58	0.66	0	0.61	0.58	-0.03	0.52	0.6	0.06
	N	0.53	0.56	0.67	-0.03	0.58	0.53	-0.05	0.57	0.66	-0.04
Avg.		0.52	0.53	0.62	-0.01	0.54	0.61	-0.02	0.53	0.61	-0.01

Sample14											
LLM	Trait	Default	Definition			Trait-Shuffle1			Trait-Shuffle2		
		R	R	UC	ΔR	R	UC	ΔR	R	UC	ΔR
Llama3	O	0.82	0.75	0.69	0.07	0.73	0.75	0.09	0.75	0.81	0.07
	C	1.04	1.09	0.58	-0.05	1.27	0.5	-0.23	1.31	0.53	-0.27
	E	1.41	1.37	0.75	0.04	1.45	0.51	-0.04	1.33	0.53	0.08
	A	0.94	1.03	0.41	-0.09	1.06	0.62	-0.12	0.9	0.53	0.04
	N	1.18	1.27	0.7	-0.09	1.12	0.66	0.06	1.18	0.62	0
	Mach	1.85	2.11	0.96	-0.26	1.95	0.97	-0.1	2.03	0.97	-0.18
	Narc	0.97	0.87	0.68	0.1	0.99	0.64	-0.02	0.95	0.66	0.02
	Psyc	2.4	2.38	0.98	0.02	2.4	0.99	0	2.4	1	0
Mistral	O	0.92	0.82	0.54	0.1	0.84	0.59	0.08	0.81	0.6	**0.11**
	C	0.91	0.99	0.61	-0.08	1.1	0.66	-0.19	0.87	0.74	0.04
	E	1.19	1.11	0.67	0.08	0.92	0.6	**0.27**	1.11	0.59	0.08
	A	0.99	0.96	0.66	0.03	1	0.69	-0.01	1	0.54	-0.01
	N	0.95	1.1	0.7	-0.15	0.93	0.76	0.02	0.99	0.69	-0.04
	Mach	1.63	1.83	0.96	-0.2	1.47	0.92	0.16	1.68	0.94	-0.05
	Narc	1.03	0.98	0.65	0.05	1	0.66	0.03	1.02	0.67	0.01
	Psyc	2.2	2.24	1	-0.04	2.16	0.99	0.04	2.3	1	-0.1
OpenChat	O	0.72	0.79	0.65	-0.07	0.7	0.73	0.02	0.75	0.71	-0.03
	C	0.88	0.95	0.66	-0.07	0.79	0.69	0.09	0.93	0.48	-0.05
	E	1.11	1.13	0.73	-0.02	1.11	0.78	0	1.37	0.62	-0.26
	A	0.66	0.85	0.57	-0.19	0.86	0.68	-0.2	0.89	0.71	-0.23
	N	1.07	1.17	0.59	-0.1	1.11	0.68	-0.04	1.12	0.7	-0.05
	Mach	1.64	1.92	0.95	-0.28	1.43	0.91	0.21	1.64	0.96	0
	Narc	0.94	0.9	0.71	0.04	0.87	0.74	0.07	0.96	0.69	-0.02
	Psyc	1.95	2.21	0.99	-0.26	2.22	0.99	-0.27	1.98	0.99	-0.03
Phi3	O	0.99	0.86	0.61	**0.13**	0.9	0.68	0.09	0.88	0.71	**0.11**
	C	1.04	1.18	0.62	-0.14	1.09	0.65	-0.05	1.12	0.57	-0.08
	E	1.24	1.23	0.62	0.01	1.16	0.61	0.08	1.19	0.56	0.05
	A	1.01	0.96	0.57	0.05	0.98	0.59	0.03	0.99	0.63	0.02
	N	1.2	1.13	0.69	0.07	1.06	0.63	0.14	1.13	0.66	0.07
	Mach	1.94	2.02	0.98	-0.08	1.98	0.96	-0.04	1.97	0.97	-0.03
	Narc	1	1.12	0.76	-0.12	1.05	0.75	-0.05	1.22	0.73	-0.22
	Psyc	2.31	2.36	1	-0.05	2.36	0.99	-0.05	2.38	0.99	-0.07
Avg.		1.25	1.30	0.73	-0.05	1.25	0.76	0.00	1.29	0.73	-0.03

Table 7: Effect of Prompt Variation on Standard Prompting Pipeline. Most desirable outcomes are bolded. **UC** stands for Unchanged rate and **R** stands for RMSE.

Essays														
LLM	Trait	Default	Definition			Static Criteria			Criteria-Shuffle1			Criteria-Shuffle2		
		F1	F1	UC	ΔF1	F1	UC	ΔF1	F1	UC	ΔF1	F1	UC	ΔF1
Llama3	O	0.45	0.52	0.61	-0.07	0.52	0.53	-0.07	0.58	0.56	-0.13	0.54	0.58	-0.09
	C	0.51	0.47	0.63	0.04	0.54	0.57	-0.03	0.6	0.58	-0.09	0.52	0.62	-0.01
	E	0.54	0.51	0.53	0.03	0.5	0.58	**0.04**	0.55	0.53	-0.01	0.55	0.56	-0.01
	A	0.5	0.5	0.6	0	0.51	0.64	-0.01	0.57	0.58	-0.07	0.56	0.65	-0.06
	N	0.5	0.56	0.8	-0.06	0.5	0.75	0	0.45	0.78	**0.05**	0.48	0.77	0.02
Mistral	O	0.5	0.44	0.81	**0.06**	0.5	0.73	0	0.52	0.82	-0.02	0.57	0.74	-0.07
	C	0.45	0.44	0.82	0.01	0.51	0.73	-0.06	0.49	0.77	-0.04	0.46	0.69	-0.01
	E	0.52	0.48	0.74	0.04	0.53	0.67	-0.01	0.56	0.74	-0.04	0.56	0.64	-0.04
	A	0.48	0.57	0.68	-0.09	0.6	0.64	-0.12	0.54	0.64	-0.06	0.48	0.64	0
	N	0.3	0.3	1	0	0.34	0.98	-0.04	0.3	1	0	0.34	0.95	-0.04
OpenChat	O	0.35	0.59	0.55	-0.24	0.42	0.68	-0.07	0.41	0.71	-0.06	0.41	0.72	-0.06
	C	0.36	0.42	0.8	-0.06	0.4	0.87	-0.04	0.31	0.89	**0.05**	0.38	0.87	-0.02
	E	0.47	0.52	0.72	-0.05	0.51	0.81	-0.04	0.42	0.79	**0.05**	0.49	0.82	-0.02
	A	0.33	0.52	0.65	-0.19	0.34	0.81	-0.01	0.33	0.87	0	0.33	0.87	0
	N	0.52	0.55	0.63	-0.03	0.53	0.57	-0.01	0.6	0.62	-0.08	0.52	0.65	0
Phi3	O	0.48	0.5	0.8	-0.02	0.48	0.7	0	0.47	0.81	0.01	0.42	0.72	**0.06**
	C	0.54	0.54	0.74	0	0.56	0.66	-0.02	0.54	0.68	0	0.58	0.71	-0.04
	E	0.55	0.59	0.62	-0.04	0.64	0.65	-0.09	0.6	0.59	-0.05	0.61	0.56	-0.06
	A	0.57	0.64	0.75	-0.07	0.7	0.61	-0.13	0.64	0.67	-0.07	0.54	0.63	0.03
	N	0.47	0.56	0.71	-0.09	0.51	0.58	-0.04	0.47	0.68	0	0.49	0.62	-0.02
Avg.		0.47	0.51	0.71	-0.04	0.51	0.69	-0.04	0.50	0.72	-0.03	0.49	0.70	-0.02

Sample14														
LLM	Trait	Default	Definition			Static Criteria			Criteria-Shuffle1			Criteria-Shuffle2		
		R	R	UC	ΔR	R	UC	ΔR	R	UC	ΔR	R	UC	ΔR
Llama3	O	1.12	1.07	0.66	0.05	1.06	0.65	0.06	1.16	0.71	-0.04	1.17	0.66	-0.05
	C	1.31	1.47	0.67	-0.16	1.24	0.78	0.07	1.3	0.68	0.01	1.25	0.67	0.06
	E	1.47	1.63	0.81	-0.16	1.53	0.83	-0.06	1.47	0.87	0	1.51	0.78	-0.04
	A	1.18	1.21	0.7	-0.03	1.23	0.8	-0.05	1.19	0.8	-0.01	1.21	0.79	-0.03
	N	1.16	1.13	0.78	0.03	1.14	0.72	0.02	1.13	0.8	0.03	1.1	0.73	0.06
	Mach	1.64	1.93	0.95	-0.29	1.71	0.97	-0.07	1.64	0.98	0	1.73	0.97	-0.09
	Narc	1.01	1.12	0.73	-0.11	1.05	0.69	-0.04	1	0.73	0.01	0.97	0.69	0.04
	Psyc	2.2	2.33	0.98	-0.13	2.19	0.98	0.01	2.15	0.97	0.05	2.18	0.97	0.02
Mistral	O	0.81	0.91	0.7	-0.1	0.88	0.75	-0.07	0.82	0.74	-0.01	0.84	0.69	-0.03
	C	1.04	0.88	0.7	0.16	1.03	0.71	0.01	0.94	0.75	0.1	0.93	0.75	0.11
	E	1.11	1.04	0.61	0.07	1.21	0.61	-0.1	1.09	0.7	0.02	1.2	0.67	-0.09
	A	1.08	0.92	0.58	0.16	1.12	0.63	-0.04	1.01	0.68	0.07	1.09	0.6	-0.01
	N	0.87	0.96	0.77	-0.09	0.9	0.77	-0.03	0.83	0.74	0.04	0.92	0.8	-0.05
	Mach	1.11	1.45	0.62	-0.34	1.2	0.59	-0.09	1.27	0.59	**-0.16**	1.2	0.61	-0.09
	Narc	1.31	1.23	0.61	0.08	1.25	0.58	0.06	1.39	0.64	-0.08	1.2	0.58	0.11
	Psyc	1.66	1.7	0.75	-0.04	1.63	0.69	0.03	1.6	0.72	0.06	1.65	0.75	0.01
OpenChat	O	0.8	0.87	0.67	-0.07	0.82	0.81	-0.02	0.84	0.81	-0.04	0.82	0.8	-0.02
	C	0.94	1.01	0.68	-0.07	0.92	0.75	0.02	0.94	0.8	0	1.01	0.8	-0.07
	E	0.97	1.09	0.61	-0.12	0.96	0.75	0.01	1.02	0.81	-0.05	0.99	0.77	-0.02
	A	0.86	0.85	0.67	0.01	0.8	0.8	0.06	0.84	0.75	0.02	0.95	0.66	-0.09
	N	0.97	1.08	0.63	-0.11	0.92	0.73	0.05	1.02	0.75	-0.05	1.01	0.76	-0.04
	Mach	1.15	1.6	0.74	**-0.45**	1.21	0.75	-0.06	1.22	0.74	-0.07	1.28	0.77	-0.13
	Narc	0.94	0.95	0.67	-0.01	0.94	0.69	0	1.02	0.67	-0.08	1.03	0.65	-0.09
	Psyc	1.63	2.05	0.92	-0.42	1.68	0.93	-0.05	1.66	0.92	-0.03	1.63	0.91	0
Phi3	O	0.85	0.87	0.67	-0.02	0.8	0.91	0.05	0.86	0.66	-0.01	0.84	0.7	0.01
	C	1.12	1.09	0.58	0.03	1.08	0.78	0.04	1.08	0.52	0.04	1.06	0.58	0.06
	E	1.24	1.15	0.46	0.09	1.2	0.67	0.04	1.22	0.6	0.02	1.24	0.56	0
	A	0.83	0.95	0.66	-0.12	0.94	0.88	**-0.11**	0.91	0.74	-0.08	0.98	0.58	**-0.15**
	N	1.11	1.2	0.66	-0.09	1.14	0.79	-0.03	1.17	0.61	-0.06	1.13	0.6	-0.02
	Mach	1.82	2	0.93	-0.18	1.8	0.93	0.02	1.76	0.93	0.06	1.77	0.9	0.05
	Narc	1.05	1.09	0.79	-0.04	1.01	0.73	0.04	1.06	0.77	-0.01	1.02	0.77	0.03
	Psyc	2.21	2.29	0.99	-0.08	2.23	0.96	-0.02	2.24	0.98	-0.03	2.17	0.98	0.04
Avg.		1.21	1.29	0.72	-0.08	1.21	0.77	-0.01	1.21	0.76	-0.01	1.22	0.73	-0.02

Table 8: Effect of Prompt Variation on Trivial-ZS evaluation. Most desirable outcomes are bolded. **UC** stands for Unchanged rate and **R** stands for RMSE.

Prompt Variant	Prompt Text
Default	According to your knowledge, how is the personality trait P manifested in the text? Can you give me an exhaustive list of textual manifestations of P in the order of importance and relevance to the Personality Psychology literature? For each instance, please provide a short explanation in a line-separated field under the title "Description" along with a few examples of the textual manifestation in the form of phrases or sentences in a line-separated field under the title "Examples".
Variant 1	How is the personality trait P represented in written text according to current research? Please offer a detailed list of textual indicators or features of P, ordered by their significance and relevance in Personality Psychology. For each indicator, provide a concise description under "Description" and include a few examples of the indicator in text under "Examples"
Variant 2	How is the personality trait P exhibited in written communication based on existing literature? Provide a thorough list of textual signs or traits associated with P. For each sign, include a short description under "Description" and several sample phrases or sentences under "Examples".
Variant 3	How does the personality trait P typically appear in the text according to Personality Psychology studies? Provide a detailed and prioritized list of textual characteristics or indicators of P. For each characteristic, include a succinct description under "Description" and a set of examples under "Examples".
Variant 4	If I ask you to conduct personality trait evaluation from text, what are the key characteristics that you would assess to evaluate P from text? For each characteristic, include a description under "Description" and a set of examples under "Examples".

Table 9: Variations of Knowledge Extraction Prompt. In each prompt P is replaced with a specific personality trait and a subsequent criteria list for each trait is obtained.

Trait	Definition
O	Openness denotes receptivity to new ideas and new experiences. People with high levels of openness are more likely to seek out a variety of experiences, be comfortable with the unfamiliar, and pay attention to their inner feelings more than those who are less open to novelty. They tend to exhibit high levels of curiosity and often enjoy being surprised.
C	Conscientiousness reflects the tendency to be responsible, organized, hard-working, goal-directed, and to adhere to norms and rules. People with high levels of conscientiousness are good at setting and keeping long-range goals, self-regulation and impulse control and take obligations to others seriously.
E	Extraversion is typically characterized by outgoingness, high energy, and/or talkativeness. People with high levels of extraversion tend to thrive in social situations, enjoy engaging with others, and often seek out stimulating environments.
A	Agreeableness can be described as cooperative, polite, kind, and friendly. People high in agreeableness are more trusting, affectionate, altruistic, and generally displaying more prosocial behaviors than others.
N	Neuroticism is defined as a tendency toward anxiety, depression, self-doubt, and other negative feelings. Highly neurotic individuals tend to be labile (that is, subject to frequently changing emotions), anxious, tense, and withdrawn.
Mach	Machiavellisnism is characterized by manipulativeness, deceitfulness, high levels of self-interest, and a tendency to see other people as means to an end. People with high levels of Machiavellianism lack empathy and take a cynical, unemotional view of the world; their primary interests center on power and status, and they'll do whatever is necessary to achieve their goals.
Narc	Narcissism is characterized by a grandiose sense of self-importance, a lack of empathy for others, a need for excessive admiration, and the belief that one is unique and deserving of special treatment. People with high levels of narcissism exhibit an inflated sense of self-importance, a deep need for excessive admiration, a lack of empathy, an exaggerated sense of entitlement, and a tendency to exploit others to maintain their self-image.
Psyc	Psychopathy is a condition characterized by the absence of empathy and the blunting of other affective states. People with high levels of psychopathy exhibit a pervasive pattern of antisocial behavior, a lack of empathy and remorse, shallow emotions, manipulativeness, impulsivity, and a tendency toward reckless and often criminal behavior without regard for the consequences or the harm inflicted on others.

Table 10: Definition of Personality Traits

Trait	Qualification Criteria
O	Imagination and Creativity, Intellectual Curiosity, Preference for Novelty and Variety, Appreciation for Arts and Aesthetics, Open-mindedness and Tolerance, Innovation and Inventiveness, Complexity
C	Organization and Planning, Dependability, Perfectionism, Self-Discipline, Adherence to Rules and Norms, Cautiousness, Efficiency, Punctuality
E	Sociability (Interacting with others), Talkativeness (Verbal communication), Assertiveness (Confident expression of ideas and feelings), Excitement-seeking (Desire for thrilling experiences), Positive emotionality (Experience and expression of positive emotions), Activity (Energetic engagement), Optimism (Expecting good outcomes), Impulsivity (Acting on whims
A	Empathy and Compassion, Trust and Altruism, Cooperativeness and teamwork, Politeness and consideration, Forgiveness and tolerance, Modesty and humility
N	Expressions of Negative Emotions, Avoidance of Emotional Topics, Fear and Anxiety, Impulsiveness, Self-Consciousness, Mood Swings, Sensitivity to Criticism, Perceived Lack of Control, Insecurity, Emotional Volatility
Mach	Cunning and Deceit, Self-Interest, Manipulation and Influence, Grandiosity, Amoral/Antisocial Tendencies, Cynicism, Calculation and Strategic Thinking, Lack of Empathy
Narc	Grandiosity, Self-centeredness, Manipulative behavior, Lack of empathy, Arrogance, Envy, Lack of intimacy, Superficiality
Psyc	Grandiosity and Self-Centeredness, Lack of Remorse or Guilt, Callousness and Lack of Empathy, Manipulation and Deceit, Shallow Emotions, Parasitic Lifestyle, Impulsivity and Irresponsibility, Criminal or Antisocial Behavior

Table 11: Manifestations of Personality Traits Identified by **Mistral**

Trait	Qualification Criteria
O	Intellectual curiosity, Artistic and creative expression, Appreciation for beauty and aesthetics, Open-mindedness and tolerance, Love of learning and exploration, Imagination and fantasy, Love of nature and the outdoors, Appreciation for complexity and nuance, Love of travel and exploration, Appreciation for tradition and heritage
C	Perfectionism, Planning and Organization, Self-Discipline, Responsibility, Punctuality, Attention to Detail, Goal-Oriented, Proactivity, Reliability, Self-Monitoring
E	Assertive language, Social references, Active verbs, Emotional expressions, Storytelling, Conversational tone, Humor, Self-promotion, Enthusiasm, Word choice
A	Cooperation, Empathy, Altruism, Compassion, Tolerance, Politeness, Avoidance of Conflict, Social Harmony
N	Anxiety and Worry, Emotional Instability, Self-Consciousness, Irritability, Hypervigilance, Self-Pity, Rumination, Social Withdrawal, Perfectionism, Emotional Reactivity
Mach	Manipulative language, Exploitative language, Dishonest language, Superficial language, Aggressive language, Passive-aggressive language, Self-promotional language, Flattery language, Blame-shifting language, Gaslighting language
Narc	Grandiosity, Self-Aggrandizement, Self-Celebration, Lack of Empathy, Entitlement, Exploitation, Grandiose Fantasies, Envy, Self-Promotion, Defensiveness, Lack of Accountability, Manipulation
Psyc	Lack of empathy and remorse, Superficial charm and wit, Manipulation, and exploitation, Impulsivity and recklessness, Grandiosity and entitlement, Lack of intimacy and emotional connection, Antisocial behavior and disregard for authority, Callousness and lack of emotional depth

Table 12: Manifestations of Personality Traits Identified by **Llama3**

Trait	Qualification Criteria
O	willingness to explore new ideas, experiences, and perspectives., preference for variety and novelty, as well as a curiosity about the world., higher tolerance for ambiguity and uncertainty, leading to a more flexible mindset., preference for creativity and artistic expression., willingness to question and challenge established norms and beliefs
C	Attention to detail and accuracy, Dependability and reliability, Adherence to rules and regulations, Perfectionism and high standards, Future-oriented thinking, Self-discipline and self-control, Punctuality and time management, Neatness and cleanliness, Responsibility, and accountability
E	Sociability, Assertiveness, Enthusiasm, Energized by social situations, Talkativeness, Outgoing nature, Expressiveness, Dominance, Activity level, Positive affect
A	Cooperation and Harmony, Empathy and Compassion, Altruism and Generosity, Trust and Forgiveness, Politeness and Consideration, Adaptability and Flexibility, Positive and Optimistic, Warmth and Affection, Conscientiousness and Responsibility, Modesty and Humility
N	Anxiety, Emotional instability, Depression, Irritability, Impulsivity, Vulnerability to stress, Low self-esteem, Social anxiety, Substance abuse, Health problems
Mach	Manipulation and Deception, Self-Interest, Cynicism, Emotional Detachment, Sense of Humor
Narc	Grandiose self-esteem, Need for admiration, Lack of empathy, Arrogance, Exploitative behavior, Envy, Entitlement
Psyc	Callousness, Grandiose self-worth, Need for stimulation, Manipulation and deceit, Antisocial behavior, Lack of responsibility, Shallow affect

Table 13: Manifestations of Personality Traits Identified by **OpenChat**

Trait	Qualification Criteria
O	Curiosity, Imagination, Creativity, Originality, Open-mindedness, Intellectualism, Aesthetics, Diversity, Adventure-seeking, Nonconformity, Intellectual humility
C	Organization and planning, Responsibility, and dependability, Goal-directed behavior, Attention to detail, Punctuality and time management, Proactivity and initiative, Diligence and hard work, Honesty and integrity, Responsibility towards others, Environmental consciousness, Health and self-care, Financial responsibility
E	Direct and Assertive Communication, Use of first-person singular pronouns, Emphasis on Social Interactions, Emphasis on Positive Emotions, Use of Expressive Language, Desire for Novelty
A	Empathy, Altruism, Cooperativeness, Friendliness, Trustworthiness, Conciliation, Forgiveness, Helpfulness, Generosity, Positivity
N	Self-doubt, Negative Emotions, Mood Instability, Pessimism, Overreaction to Stress, Hypersensitivity to Criticism, Emotional Exhaustion, Ruminating, Insecurity, Social Anxiety, Intensified Emotional Responses
Mach	Manipulative behavior, Emotional detachment, Deceitfulness, Use of flattery, Lack of remorse, Cunningness, Use of fear, Selfishness, Grandiose sense of self, Charisma
Narc	Self-enhancement and grandiosity, Lack of empathy, Manipulative behavior, Need for admiration, Inflated sense of self-importance, Lack of accountability, Sensitivity to criticism, Entitlement, Jealousy, Lack of authenticity
Psyc	Lack of empathy, Shallow affect, Superficial charm, Grandiose self-worth, Pathological lying, Manipulativeness, Impulsivity, Lack of remorse or guilt, Failure to accept responsibility, Parasitic lifestyle, Poor behavioral controls, Early behavioral problems

Table 14: Manifestations of Personality Traits Identified by **Phi3**

Trait	Qualification Criteria
O	Use of Imaginative and Creative Language, Preference for Variety and New Experiences, Intellectual Curiosity and Inclination Towards Learning, Open-Mindedness and Tolerance for Unconventional Ideas, Aesthetic Sensitivity and Appreciation for Art and Beauty, Expressiveness and Richness in Emotional Descriptions, Philosophical and Reflective Thinking, Use of Figurative and Metaphorical Language, Interest in Diverse Topics and Cross-Disciplinary Thinking, Use of Descriptive and Detail-Rich Narratives
C	Organization and Orderliness, Dependability and Reliability, Persistence and Perseverance, Attention to Detail, Self-Discipline and Control, Goal-Setting and Achievement Orientation, Responsibility and Accountability, Punctuality, Hard-Working and Industrious, Planning and Foresight, Achievement-Striving
E	Sociability and Social Interaction, Talkativeness and Expressiveness, Enthusiasm and Positivity, Assertiveness and Leadership, Preference for Stimulation and Activity, Friendliness and Approachability, Outgoing Nature and Willingness to Meet New People, High Activity Levels and Liveliness, Preference for Group Work, Risk-Taking and Adventurousness
A	Compassion and Empathy, Politeness and Manners, Cooperation and Willingness to Help, Positive and Encouraging Language, Conflict Avoidance, Trust and Faith in Others, Supportive and Reassuring Statements, Compliments and Praise, Consideration of Others' Opinions, Expressions of Gratitude
N	Expressions of Anxiety, Expressions of Emotional Instability, Expressions of Negative Affect, Expressions of Self-Consciousness, Expressions of Vulnerability, Expressions of Guilt, Expressions of Pessimism, Expressions of Hypersensitivity, Expressions of Indecisiveness, Expressions of Excessive Self-Concern
Mach	Manipulation and Exploitation, Strategic Planning and Cunning, Lack of Morality and Ethics, Cynicism and Distrust, Manipulative Charm, Emotional Detachment, Focus on Self-interest, Deceptiveness and Lying, Noncompliance with Social Norms, Control over Others
Narc	Self-Aggrandizement, Lack of Empathy, Need for Admiration, Sense of Entitlement, Exploitativeness, Enviousness, Arrogance and Haughtiness, Preoccupation with Fantasies, Interpersonal Manipulation, Self-Perception of Uniqueness, Defensive Reactions to Criticism, Obsession with Appearance and Status
Psyc	Lack of Empathy, Superficial Charm, Manipulativeness, Grandiosity, Pathological Lying, Impulsivity, Irresponsibility, Lack of Remorse or Guilt, Shallow Emotions, Parasitic Lifestyle, Callousness, Poor Behavioral Controls, Criminal Versatility, Promiscuous Sexual Behavior, Early Behavioral Problems

Table 15: Manifestations of Personality Traits Identified by **GPT4**

LIWC Categories	O	C	E	A	N	Mach	Narc	Psyc
Drives	-	0.76	-	-	-0.97	-	-	-
affiliation	-	-	-	-	-	-	-	-
achieve	-	-	-0.74	-0.69	-0.89	-	-	-
power	0.97	0.92	-	1.00	-0.98	-	**0.91**	-
Cognition	-	-	0.99	-	0.96	-	0.84	-
allnone	-0.92	-0.88	**-0.95**	-1.00	-0.96	-	**0.97**	-
cogproc	-	-	1.00	-	**0.97**	-	-	-
insight	-	-	-	-	-	-	-0.65	-0.58
cause	-	-	-	-	-	-	-	-
discrep	-1.00	**-0.89**	-0.92	-0.99	-	-	0.95	-
tentat	-	-	0.86	**-0.86**	1.00	-	-	-
certitude	0.82	0.98	-	-	0.99	-	-0.78	-
differ	-	-	0.82	0.90	-	-	-	-
memory	**0.97**	**1.00**	-	1.00	-	-	**0.73**	0.98
Affect	0.89	-	-	0.88	-1.00	-	-	-
tone_pos	**0.98**	0.94	0.97	0.86	-	-	-	-0.70
tone_neg	-	-	-	0.54	-1.00	-	0.78	-
emotion	0.877	-	0.662	0.920	-	-	-	-
emo_pos	0.733	-	0.989	0.895	-	-	-	-
emo_neg	-	-	-	0.88	-0.79	-	-	-
emo_anx	0.98	-	-0.89	1.00	**-0.98**	-	**-0.91**	-
emo_anger	-	**1.00**	0.98	**1.00**	**0.98**	-	-0.52	-
emo_sad	0.98	1.00	0.97	-1.00	**-0.98**	-	0.72	-
swear	0.97	-	-	-1.00	-	0.72	**0.94**	-
Social	0.81	0.64	-	-	1.00	-	-	-
socbehav	0.95	0.83	**1.00**	-	**1.00**	-	-	-0.72
prosocial	-	-	-0.92	-	**-0.97**	-	-	-
polite	0.97	**1.00**	-	1.00	-	-	-0.98	-
conflict	**0.98**	1.00	0.98	**1.00**	**0.97**	-	-	-
moral	-	-	-	-	-0.92	-	-0.79	-
comm	-	-	0.98	1.00	**0.98**	-	-	-
socrefs	-	-	-0.89	-	1.00	-	-0.69	-
family	0.83	-	-0.80	-	-0.79	-0.55	**-0.95**	-
friend	-0.93	1.00	-	-	-0.98	-	-	0.85
female	**-0.98**	-0.97	**-0.98**	**-1.00**	-	-	-0.95	-
male	0.98	**0.98**	0.98	**1.00**	-	-	-	-
Culture	-	1.00	-0.79	-0.85	-	-	**-0.98**	0.93
politic	0.97	1.00	-	-	-	-	**-0.97**	0.97
ethnicity	**-0.66**	-	-	-	-0.97	-	**-0.97**	0.97
tech	0.98	**1.00**	-	1.00	-	-	-0.97	0.88
Lifestyle	0.99	-	0.93	-	0.95	-	-	-
leisure	-0.94	-0.86	-0.83	-	-0.98	-	**0.57**	-
home	-	-	0.98	**0.96**	0.96	-0.80	-0.98	-
work	0.90	0.83	**0.96**	-	0.99	-	-0.70	-
money	**0.98**	-	-	1.00	-	-	-	-
relig	**0.98**	**1.00**	0.98	**1.00**	-	-	0.59	-
Physical	-0.81	-0.96	-0.87	-	**-0.99**	-	-	-

Table 16: Median resultant LIWC Correlations across valid LLMs from Monte Carlo Simulation (Part 1/2)

LIWC Categories	O	C	E	A	N	Mach	Narc	Psyc
health	-	-0.85	-0.97	1.00	-0.94	-	-	-
illness	**0.97**	**1.00**	0.98	1.00	0.98	-	-0.98	-
wellness	-	-	-	-	-	-	**-0.97**	0.97
mental	0.97	-	-	-	-0.97	-0.97	0.80	-
substances	0.97	1.00	-	-	-	-	-	-
sexual	**0.97**	1.00	-	-	0.97	-	0.91	-
food	-0.90	-	-0.96	1.00	-0.98	-	-0.84	0.80
death	**0.97**	**1.00**	-	1.00	-	0.62	0.86	0.91
need	-0.79	-	0.89	-	0.93	-0.65	-0.69	-
want	-0.96	**-0.76**	-0.99	-	-0.99	-	-	-
acquire	-	-	-	0.84	-0.99	-	-	-
lack	0.98	**1.00**	0.97	1.00	-	-	**-0.89**	-
fulfill	-0.90	0.72	-	**1.00**	-0.89	-	-0.98	-
fatigue	0.98	-	-0.88	-	-0.97	-	-0.97	-
reward	**0.97**	**1.00**	-	1.00	-	-	-	-
risk	0.98	1.00	-	-	-	-	-	-
curiosity	-0.90	-	0.98	-	-	-	-	-
allure	-0.98	-0.75	**-1.00**	-0.80	**-1.00**	-	-	-
Perception	-	-0.91	-	-	-1.00	-	-	-
attention	-	-	0.76	1.00	0.98	-	-	-
motion	-	**-0.77**	-0.82	-	-1.00	-	-	-
space	-	-	1.00	-	-0.96	-	-	-
visual	-	-0.89	-	-0.99	-	-	-	-
auditory	-	-	-	0.99	-0.86	-	0.68	-
feeling	-0.80	-	-0.95	-	**-0.99**	-	-	-
time	-	**-0.83**	-0.81	-0.93	-0.90	-	-	-
focuspast	-	-0.60	**-1.00**	-0.72	-1.00	-	-	-
focuspresent	-	-	-	-0.94	1.00	-	-	-0.57
focusfuture	-0.72	-0.66	-	-	-	-	-	-
Conversation	-0.93	-0.63	-0.98	-0.93	-	-	-	-

Table 17: Median resultant LIWC Correlations across valid LLMs from Monte Carlo Simulation (Part 2/2)