PADO 🌊: Personality-induced multi-Agents for Detecting OCEAN in human-generated texts

Haein Yeo[1], **Taehyeong Noh**[1], **Seungwan Jin**[2], **Kyungsik Han**[1,2,*]
[1]Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea
[2]Department of Data Science, Hanyang University, Seoul, Republic of Korea
{haeinyeo, yestaehyung, seungwanjin, kyungsikhan}@hanyang.ac.kr

Abstract

As personality can be useful in many cases, such as better understanding people's underlying contexts or providing personalized services, research has long focused on modeling personality from data. However, the development of personality detection models faces challenges due to the inherent latent and relative characteristics of personality, as well as the lack of annotated datasets. To address these challenges, our research focuses on methods that effectively exploit the inherent knowledge of Large Language Models (LLMs). We propose a novel approach that compares contrasting perspectives to better capture the relative nature of personality traits. In this paper, we introduce PADO (**P**ersonality-induced multi-**A**gent framework for **D**etecting **O**CEAN of the Big Five personality traits), the first LLM-based multi-agent personality detection framework. PADO employs personality-induced agents to analyze text from multiple perspectives, followed by a comparative judgment process to determine personality trait levels. Our experiments with various LLM models, from GPT-4o to LLaMA3-8B, demonstrate PADO's effectiveness and generalizability, especially with smaller parameter models. This approach offers a more nuanced, context-aware method for personality detection, potentially improving personalized services and insights into digital behavior. The code is available at `https://github.com/haaaein/PADO`.

1 Introduction

Personality often has a significant impact on our daily lives and work. Since an individual's personality manifests in various aspects such as behavioral patterns, interpersonal relationships, and stress management abilities, one's lifestyle and work outcomes can vary greatly depending on one's personality (Youyou et al., 2017; Štajner and Yenikent, 2020). With the widespread creation and sharing of information in online environments, research is actively underway to develop models for recognizing individual personalities, to understand the role of personality in given contexts, and to develop personalized services (Alshouha et al., 2024; Dhelim et al., 2022; Yang et al., 2020).

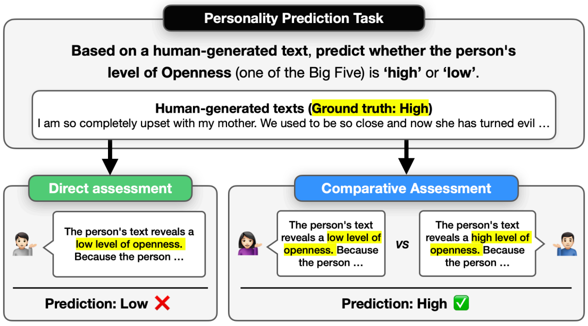

Personality Prediction Task
Based on a human-generated text, predict whether the person's level of Openness (one of the Big Five) is 'high' or 'low'.
Human-generated texts (Ground truth: High)
I am so completely upset with my mother. We used to be so close and now she has turned evil ...
Direct assessment
The person's text reveals a low level of openness. Because the person ...
Prediction: Low ❌
Comparative Assessment
The person's text reveals a low level of openness. Because the person ...
vs
The person's text reveals a high level of openness. Because the person ...
Prediction: High ✅

Figure 1: Our research motivation. For latent and relative characteristics like personality, comparing different perspectives may be more effective than a single assessment approach.

In terms of personality detection, the task of reliably extracting personality from user-generated text is complex and challenging due to the inherently latent and context-dependent relative nature of personality (Fang et al., 2022). Relying on specific words or word combinations to identify personality can limit the utility of personality detection models in a variety of contexts (Alshouha et al., 2024). Furthermore, it is important to have large-scale, accurately labeled datasets to effectively build personality detection models. The process of preparing labeled datasets requires considerable human effort, and subjective judgments about personality labels often lead to inconsistencies in the datasets. For these reasons, many previous studies still face a fundamental limitation: the lack of generalizable labeled data and a trustworthy model (Zhu et al., 2024).

The advancement of LLMs has demonstrated significant performance improvements in natural language processing tasks without fine-tuning,

---

[*] Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 5719–5736
January 19–24, 2025. ©2025 Association for Computational Linguistics

showcasing their ability to process and understand vast amounts of textual data. Building on these strengths, recent research has attempted to harness the extensive knowledge and reasoning abilities of LLMs for personality detection approaches, aiming to extract richer information from text (Hu et al., 2024; Rao et al., 2023). However, current LLMs rely heavily on patterns in the training data and still tend to exhibit unintentional biases and are sensitive to specific word choices and contexts, limiting the accuracy and consistency of personality detection (Gallegos et al., 2024; Jiang et al., 2023).

In recent research, multi-agent approaches have been proposed to better harness the capabilities of LLMs and overcome these limitations. However, since LLMs may exhibit a variety of cognitive biases similar to humans (Turpin et al., 2024; Koo et al., 2023), it is difficult to obtain diverse perspectives simply by increasing the number of agents. In particular, caution is required because multiple agents with similar perspectives or biases may lead to problems similar to human confirmation bias. Therefore, it is important to use agents in a structured way, taking into account additional important social science perspectives, to reduce bias and increase the accuracy and reliability of the results (Naik et al., 2023). Liang et al. (Liang et al., 2023) confirmed that inducing divergent thinking through agents with opposing viewpoints contributes to performance improvement. Furthermore, previous studies have used the LLM-as-a-Judge method, in which LLMs generate the thinking and reasoning process and are the subject of evaluation. These studies conducted pairwise comparison or comparative comparison to compare opposing viewpoints, evaluate their differences or judge their appropriateness, and make a final decision, rather than direct evaluation (Zheng et al., 2024).

Based on these motivations, this study proposes PADO 🌊 (**P**ersonality-induced multi-**A**gent for **D**etecting **O**CEAN), a personality detection framework that evaluates by comparing and analyzing contrasting perspectives to more reliably detect personality traits that can be interpreted differently depending on situations and contexts (Figure 1). PADO aims to achieve more accurate and balanced personality predictions by integrating the analyses of LLM agents with different personality characteristics. PADO consists of three main phases: 1) A personality induction phase, which induces personality in LLMs at two levels, high and low, for each dimension of the Big Five personality model, 2) A phase where personality-induced reasoner agents interpret and explain emotional, cognitive, and social aspects, which are important elements of personality from a psycholinguistic perspective, 3) A comparative judgment phase, where a judge agent comprehensively compares and analyzes the explanations generated earlier. This approach allows judge agents to more accurately capture subtle personality information embedded in the text, and to more fully consider the multifaceted nature of personality by comprehensively considering the differences and similarities between two personality-induced reasoner agents. We have experimentally demonstrated the effectiveness and generalizability of PADO by applying it to various scales of open language models, ranging from GPT-4o to LLaMA3-8B.

In summary, the main contributions of this research are as follows:

- We propose a novel personality detection framework, PADO, which employs personality-induced LLM agents. By comparing and evaluating the reasoning of agents with different personality traits, PADO provides a more accurate and balanced interpretation of the complex and challenging task of extracting personality from text.

- We have experimentally demonstrated that PADO can be applied to various language models from GPT-4o to LLaMA3-8B. This shows that PADO can be broadly applied to models with different parameter sizes and datasets, without being dependent on model size or specific training data.

- PADO introduces an analysis method that takes into account emotional, cognitive, and social aspects, which are important from a psycholinguistic perspective, without additional training. This allows for a richer capture of subtle personality information inherent in the text and a comprehensive consideration of the multifaceted nature of personality.

PADO can be applied in many domains to better understand users' personality traits, explore various social phenomena, and provide personalized and tailored services. Furthermore, our approach has the potential to be extended beyond personality detection to text analysis tasks that require explainability to capture implicit and relative traits.

2 Related Work

2.1 Personality Prediction

Personality is defined as a stable set of traits that explain an individual's behaviors, emotions, attitudes, and habits. The Big Five model, which categorizes them into five dimensions, is widely used (McCrae and John, 1992; Cattell, 1946; Tupes and Christal, 1992). The five dimensions of this model (i.e., Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) are also known as OCEAN. These personality traits are closely related to an individual's language use patterns, with the frequency of use of certain words and sentence structure reflecting an individual's personality (Park et al., 2015). Predicting personality from text has important application value in various fields, such as recommender systems and mental health for personalized service offerings, and it is of great help in understanding and predicting user behavior online (Mehta et al., 2020; Zanwar et al., 2023). Personality prediction research has evolved from early frequency-based models to complex methods using deep neural networks (Zhu et al., 2022; Yang et al., 2021; Amirhosseini and Kazemian, 2020). Pennebaker et al. (Pennebaker et al., 2001) developed Linguistic Inquiry and Word Count (LIWC) to extract psycholinguistic features from text, which has been used for feature engineering in machine learning models.

A key aspect of personality is that it can be expressed differently across languages, cultures, and contexts, making modeling to accurately capture it challenging. However, existing approaches to personality prediction rely primarily on data-driven methods that are limited by their inability to fully reflect the implicit and multifaceted nature of personality. For example, traditional methods (e.g., LIWC) focus primarily on explicit psycholinguistic traits that fail to account for non-verbal factors and contextual nuances. Overcoming these limitations requires a multifaceted analysis that integrates expertise from linguistics and psychology is needed. However, building such multifaceted data can be a very complex and time-consuming process. Therefore, there is a need for a methodology that can effectively capture different aspects of personality while reducing data dependency.

2.2 LLM-Based Approaches

Recently, LLMs have shown significant performance gains without fine-tuning on a variety of natural language processing tasks (Brown et al., 2020). LLMs have the ability to process and understand large amounts of textual data, allowing them to consider and analyze different aspects of personality, and they have the potential to integrate insights from multiple disciplines to analyze personality. However, the use of LLMs in the task of personality prediction has not been fully explored, and personality prediction studies using zero-shot-based prompting methods such as Chain-of-Thought (CoT) few-shot have not performed as expected (Ji et al., 2023).

Since personality can be interpreted in different ways depending on context and criteria, it may be more effective to analyze these different perspectives together rather than relying on a single interpretation. Recently, multi-agent approaches have gained attention to take advantage of the reasoning abilities of LLMs. For example, Liang et al. (Liang et al., 2023) proposed the Multiple Agents Debate (MAD) framework, which uses multiple agents to present different perspectives, critique each other's responses, and reach a final consensus or improved conclusion. Through this process, they found that encouraging divergent thinking from agents with opposing perspectives contributes to improved performance.

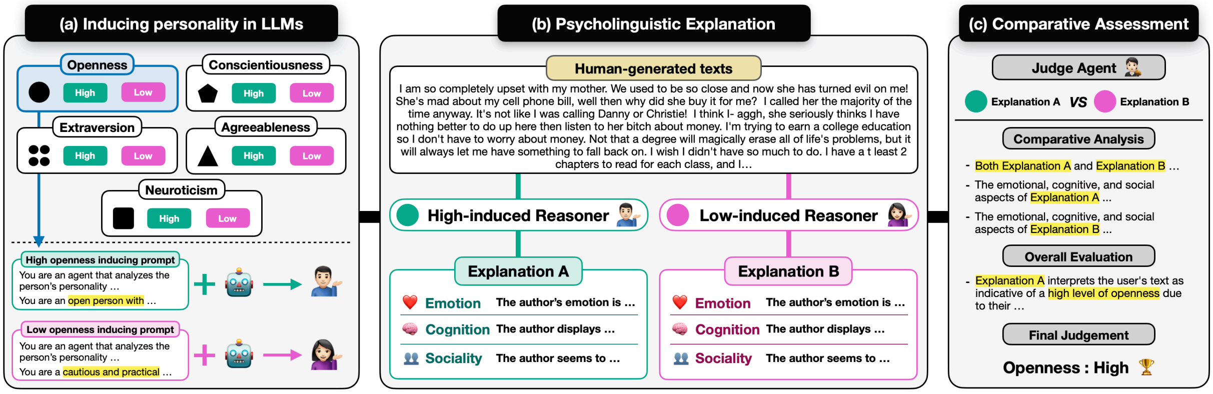
In this study, we propose a method to address the subjective interpretation of personality traits across different contexts and criteria, which has not been adequately considered in previous personality detection research. Specifically, we aim to (1) understand personality by leveraging multi-agent approaches to ensure the inclusion of multiple perspectives, and (2) design personality detection agents based on theoretical aspects of personalities from psycholinguistics.

3 Methods

In this section, we present the main phases of PADO 🌊, a multi-agent personality detection framework. An overview of PADO is illustrated in Figure 2, and the pseudocode is provided in Algorithm 1.

3.1 Inducing Personality in LLMs

The first phase of PADO involves generating agents with distinct personality traits, enabling analysis of the Big Five personality dimensions from various perspectives. Previous research has demonstrated that LLMs are not neutral and may display incon-



You are an agent that analyzes the person's personality …
You are an open person with …

You are an agent that analyzes the person's personality …
You are a cautious and practical …

I am so completely upset with my mother. We used to be so close and now she has turned evil on me! She's mad about my cell phone bill, well then why did she buy it for me? I called her the majority of the time anyway. It's not like I was calling Danny or Christie! I think I- aggh, she seriously thinks I have nothing better to do up here then listen to her bitch about money. I'm trying to earn a college education so I don't have to worry about money. Not that a degree will magically erase all of life's problems, but it will always let me have something to fall back on. I wish I didn't have so much to do. I have a t least 2 chapters to read for each class, and I…

Emotion The author's emotion is …
Cognition The author displays …
Sociality The author seems to …

- Both Explanation A and Explanation B …
- The emotional, cognitive, and social aspects of Explanation A ...
- The emotional, cognitive, and social aspects of Explanation B ...

- Explanation A interprets the user's text as indicative of a high level of openness due to their …

Figure 2: The overview of PADO 🌊. To classify the five personality traits (i.e., OCEAN) for the given text, steps from (a) to (c) are performed for each trait. This figure presents an example of openness.

sistent personality traits and perspectives (Frisch and Giulianelli, 2024). This suggests that the inherent complexity of personality traits embedded within LLMs can adversely affect the consistency and stability of their responses. Therefore, it is crucial to establish clear and well-defined guidelines for analyzing personality within LLMs, ensuring that analysis is based on consistent perspectives.

To address this, we employed the MPI Evaluation Dataset (Jiang et al., 2024), which is designed to ensure internal consistency—a key metric for assessing the stability of personality traits in LLMs. We utilized the Personality Prompting (P^2) methodology (Jiang et al., 2024) to induce specific levels of personality traits within the LLMs. The P^2 approach involves presenting sentences and keywords associated with each personality trait (e.g., "You are a kind person who values trust, morality, selflessness, and cooperation") to induce the internalization of certain personality traits, as if the agent were introducing itself. This method enables us to induce positive or negative traits across each dimension of the OCEAN model, thereby generating agents that display high or low levels of the respective personality traits.

We evaluated the consistency and accuracy of the personality representations exhibited by the generated agents, with the results presented in Table 1. As anticipated, positive inducing prompts generally led to higher personality trait scores, while negative prompts resulted in lower scores. Additionally, we observed that the standard deviation of the personality trait scores was lower when specific personality traits were induced, compared to the neutral condition (i.e., when no specific personality traits were induced). This finding indicates greater stability in the induced personality traits. Detailed examples of the prompts used for personality inducing are provided in Appendix C.3.

3.2 Psycholinguistic Explanation

Building on previous research on the relationship between text and personality traits, we sought to analyze personality in a more multifaceted way by incorporating three main psycholinguistic elements into our methodology: emotional, cognitive, and social aspects (Pennebaker and King, 1999; Pennebaker et al., 2001). These elements are directly reflected in the prompt design, prompting the model to interpret the text from different perspectives. For example, the emotion aspect focuses on how language conveys positive or negative emotions, the cognitive aspect analyzes thought processes and complexity (e.g., problem solving and reasoning), and the social aspect examines interpersonal interactions reflected in the text.

To do this, the prompt provided definitions of the three psycholinguistic factors and asked agents to generate an explanation of the basis for predicting personality for each factor. The personality-induced agent analyzes the author's personality based on its own personality and asks the model to analyze the text based on its emotional, cognitive, and social context. More examples of prompts are provided in the Table 9 (Appendix C.2).

3.3 Comparative Assessment

Based on previous research using LLM as a Judge (Zheng et al., 2024; Kim et al., 2024), we structured the decision-making process of the judging agent into three phases: comparative analysis, overall evaluation, and final judgment. In the comparative analysis phase, the judgment agent is given

Model	Target	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism	
		score	σ	score	σ	score	σ	score	σ	score	σ
GPT-4o	Neutral	4.000	1.732	4.333	1.490	4.000	1.732	4.667	1.106	4.667	1.818
	Positive	5.000	0.000	5.000	0.000	4.875	0.599	5.000	0.000	4.958	0.200
	Negative	1.167	0.799	2.167	1.675	1.542	1.322	2.042	1.567	1.000	0.000
GPT-3.5-turbo	Neutral	3.500	1.760	3.830	1.520	4.000	1.530	3.580	1.220	3.120	1.690
	Positive	4.540	0.760	4.920	0.280	4.580	0.760	5.000	0.000	3.750	1.420
	Negative	2.083	1.288	2.416	1.469	1.958	1.428	3.125	1.832	2.166	1.818
Solar-10.7B-Instruct	Neutral	3.958	1.098	3.875	0.832	3.667	0.799	4.292	0.934	2.625	1.073
	Positive	4.708	0.455	4.708	0.611	4.583	0.759	4.791	0.576	3.667	0.943
	Negative	2.583	1.077	3.250	1.010	2.750	1.090	3.625	1.148	1.708	0.934
LLaMA3-8B-Instruct	Neutral	3.291	1.513	4.500	1.000	3.000	1.527	4.660	0.624	1.833	1.404
	Positive	4.750	0.661	4.791	0.406	4.666	0.850	4.916	0.276	4.208	0.957
	Negative	1.875	0.665	2.833	1.404	1.417	0.493	3.500	1.500	1.333	0.471

Table 1: Results of inducing personality traits across the OCEAN in various LLM models. For each personality trait, a score closer to 5 for positively induced agents and closer to 1 for negatively induced agents indicates more accurate induction. A lower standard deviation indicates more stable induction results.

the human-written text and the explanations generated by the two reasoner agents as input. The judge agent examines the explanations provided by both reasoners for each key element: emotion, cognition, and sociality. This involves identifying points of agreement and disagreement, comparing how well each analysis matches specific examples from the user's text, and evaluating the depth and evidence supporting their conclusions. The overall evaluation phase requires the judge agent to synthesize these findings and determine which reasoner's overall analysis better reflects the user's personality trait. Finally, in the final judgment phase, the judge agent concludes whether the trait is high or low based on the cumulative evidence and analysis. Detailed prompts and explanations can be found in Table 10 (Appendix C.2).

4 Experiments

4.1 Datasets

We conducted experiments using two widely recognized datasets for text-based personality extraction. The datasets are labeled on human-generated text using standardized self-report questionnaires. Both datasets are used to analyze personality traits from text, providing distinct text formats (i.e., free-form essays and social media status messages) that enable personality prediction across various text types.

- **Essays (Pennebaker and King, 1999)**: The Essays dataset consists of 2,468 stream-of-consciousness essays, each labeled as either low or high on the OCEAN personality traits using a standardized self-report questionnaire. The average length of each essay was approximately 50 sentences, and 10% of the total dataset was sampled for testing. This dataset is the most widely used in research on text-based personality analysis.

- **MyPersonality (Celli et al., 2013)**: The MyPersonality dataset consists of 9,913 status messages collected from a Facebook app, written by 250 users. Each post is labeled as either low or high on the OCEAN personality traits using a standardized self-report questionnaire. A total of 250 user-generated texts were used in the experiment.

4.2 Baseline Models

To experiment with models with different parameters, we selected a number of state-of-the-art decoder-based LLMs that demonstrate excellent performance across a range of scales. Our selection includes large models known for their advanced capabilities (e.g., GPT-4o and GPT-3.5-turbo) as well as smaller yet powerful models (e.g., LLaMA3-8B-Instruct and Solar-10.7B-Instruct). We also included traditional encoder-based models for comparison.

- **Small neural network models**: BERT (Devlin, 2018) and RoBERTa (Liu, 2019) are encoder-based models. For these two models, we employed a data-splitting strategy of 8:1:1 for training, validation, and test sets, respectively.

Model	**Method**	**Essays**						**MyPersonality**					
		O	**C**	**E**	**A**	**N**	**Average**	**O**	**C**	**E**	**A**	**N**	**Average**
BERT	Fine-tuning	0.59	0.56	0.58	0.54	0.56	0.57	0.58	0.52	0.50	0.58	0.32	0.5
RoBERTa	Fine-tuning	0.59	0.56	0.57	0.57	0.56	0.57	0.68	0.52	0.45	0.46	0.36	0.49
GPT-4o	Baseline (zero-shot)	0.62	0.38	0.41	0.59	0.64	0.53	0.68	0.31	**0.55**	0.57	0.52	0.53
	one-shot	0.39	0.43	0.51	0.58	**0.66**	0.51	0.73	0.30	**0.55**	0.58	**0.56**	0.54
	CoT	0.58	0.45	0.48	0.57	0.61	0.54	0.74	0.50	0.47	0.53	0.35	0.52
	PADO (Ours)	**0.70**	**0.70**	**0.63**	**0.65**	0.61	**0.66**	**0.83**	**0.53**	0.53	**0.59**	0.51	**0.60**
GPT-3.5	Baseline (zero-shot)	0.52	0.34	0.48	0.42	0.62	0.48	0.72	0.53	0.53	0.62	0.29	0.61
	one-shot	0.25	0.11	0.39	0.23	0.62	0.32	0.64	0.61	0.5	0.61	0.37	0.55
	CoT	0.50	0.20	0.39	0.37	0.62	0.48	0.74	0.50	0.47	0.53	0.35	0.52
	PADO (Ours)	**0.72**	**0.68**	**0.69**	**0.67**	**0.67**	**0.69**	**0.82**	**0.61**	**0.55**	**0.65**	**0.57**	**0.64**
Solar-10.7B-Instruct	Baseline (zero-shot)	0.40	0.16	0.14	0.36	0.38	0.29	0.52	0.32	0.43	0.50	0.27	0.41
	one-shot	0.40	0.23	0.45	0.43	0.53	0.41	0.68	0.34	0.50	0.40	0.53	0.49
	CoT	0.34	0.14	0.05	0.27	0.40	0.24	0.44	0.29	0.41	0.43	0.36	0.39
	PADO (Ours)	**0.68**	**0.67**	**0.66**	**0.63**	**0.64**	**0.66**	**0.81**	**0.57**	**0.55**	**0.60**	**0.54**	**0.61**
LLaMA3-8B-Instruct	Baseline (zero-shot)	0.55	0.22	0.38	0.23	0.63	0.40	0.65	0.25	0.53	0.39	0.52	0.47
	one-shot	0.61	0.29	0.35	0.39	0.64	0.46	0.68	0.34	0.50	0.40	0.53	0.49
	CoT	0.60	0.26	0.43	0.53	**0.65**	0.49	0.63	0.24	0.52	0.36	0.50	0.45
	PADO (Ours)	**0.69**	**0.68**	**0.67**	**0.63**	**0.65**	**0.66**	**0.82**	**0.60**	**0.54**	**0.64**	**0.55**	**0.63**

Table 2: Comparison of Big 5 Personality prediction performance on the Essays and the MyPersonality dataset (O: Openness, C: Conscientiousness, E: Extraversion, A: Agreeableness, N: Neuroticism). F1 score was used as the performance metric for all results.

- **GPT family**: We used the versions of GPT-4o[*] and GPT-3.5-turbo[†] provided by OpenAI from May 13, 2024.
- **Solar**: Solar-10.7B-Instruct-v0.1[‡] is a decoder-based LLM with 10.7 billion parameters, available through Hugging Face.
- **LLaMA**: LLaMA3-8B-Instruct[§], provided by Meta, is a large language model with 8 billion parameters.

4.3 Implementation Details

We used pytorch v2.2.0 and transformers v4.40.0 on the NVIDIA RTX H100 GPU. The temperature used for inference was set to 0, and the max tokens were set to 512. We compared PADO with traditional prompting techniques such as zero-shot (Kojima et al., 2022), one-shot (Brown et al., 2020), and Chain-of-Thought (CoT) (Wei et al., 2022). To ensure the reliability of the results, we conducted the experiment three times under the same conditions. All prompts used in the experiment can be found in Appendix C.

5 Results and Analysis

In this section, we present a comprehensive evaluation of the proposed PADO framework through various experiments designed to assess its effectiveness in personality detection. These experiments include performance metrics (Section 5.1), a robustness study (Section 5.2), human evaluation (Section 5.3), and a detailed case study (Section 5.4), employing both quantitative and qualitative analyses.

5.1 Personality Detection Results

As shown in Table 2, PADO consistently outperforms traditional in-context learning methods, such as zero-shot, one-shot, and CoT, across all datasets. The improvement is particularly pronounced in smaller models such as Solar-10.7B-Instruct and LLaMA3-8B-Instruct, where F1 scores increased by up to 0.37 compared to the baseline.

This significant performance improvement on small models suggests that PADO effectively leverages latent personality knowledge within the model that traditional prompting methods do not fully exploit. The robustness and adaptability of PADO are demonstrated by its consistently good performance across a wide range of model sizes and architectures (e.g., from GPT-4o to LLaMA3-8B-Instruct). These quantitative results indicate that our approach, which combines psycholinguistic factors with personality-induced agents, can be ef-

---

[*] https://platform.openai.com/docs/models/gpt-4o

[†] https://platform.openai.com/docs/models/gpt-3-5-turbo

[‡] https://huggingface.co/upstage/SOLAR-10.7B-Instruct-v1.0

[§] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Algorithm 1: Pseudocode of PADO 🌊

```
Input: User text $T$;
  High personality inducing prompt $H$;
  Low personality inducing prompt $L$;
  Large Language Model $M$;
  Psycholinguistic Explanation prompt $E$;
  Judge prompt $J$
Output: Final decision $D$
Step 1: Generate Explanations;
Prompt$_A$ $\leftarrow H + E + T$;
Explanation$_A$ $\leftarrow M($Prompt$_A)$;
Prompt$_B$ $\leftarrow L + E + T$;
Explanation$_B$ $\leftarrow M($Prompt$_B)$;
Step 2: Randomize Explanation Order;
Randomly assign Explanation$_A$ and
  Explanation$_B$ to Explanation$_1$ and
  Explanation$_2$;
Step 3: Generate Judge's Assessment;
Judge_Input $\leftarrow$
  $J + T +$ Explanation$_1$ $+$ Explanation$_2$;
Assessment $\leftarrow M($Judge_Input$)$;
Step 4: Extract Final Decision;
$D \leftarrow$ ExtractDecision(Assessment);
```

fective in improving the personality detection capabilities of LLMs, especially in resource-constrained scenarios where smaller models are preferred.

5.2 Robustness Study

In LLMs, a major concern is the variability in results across different runs, making it challenging to ensure consistent and reliable evaluations. Achieving consistency in model performance is critical to ensure the trustworthiness of the results. To address this, we conducted each experiment three times and calculated the average performance across these runs. Additionally, we examined the standard deviation between the three runs to assess the stability of each method in the Essays dataset. As shown in the figure 3, our proposed method, PADO, has a lower standard deviation across all OCEAN personality traits compared to traditional methods such as zero-shot, one-shot, and CoT in the Essays dataset. This indicates that PADO consistently produces more stable and reliable predictions, making it a robust approach for personality prediction tasks.

Performance variations due to temperature adjustments are also an important method for assessing the robustness of LLMs (Loya et al., 2023). We conducted experiments investigating the impact of temperature changes on model performance, and the detailed results of these experiments are included in Appendix B.

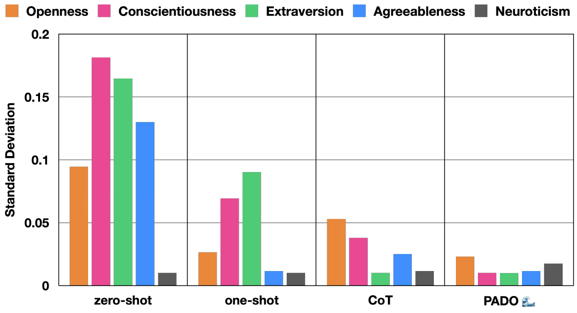


Figure 3: Comparison of standard deviation in personality trait predictions across PADO, and the existing methods (i.e., zero-shot, one-shot, CoT) for the GPT-3.5 model. Each bar shows the variability in predictions for each trait, highlighting the performance of our method.

5.3 Human Evaluation

In the absence of ground-truth data for the explanation generation task, we conducted a human evaluation with participants to assess the quality of the generated explanations. Participants rated the explanations based on five metrics (i.e., fluency, informativeness, relevance, specificity, and coherence) that were selected based on previous research for evaluating LLM-generated explanations (Ramos et al., 2024; Shen et al., 2023).

The participants consisted of five master's students and three doctoral students from a university, and they evaluated 15 examples of explanations generated using three different prompt methods. Each explanation and its presentation order were randomized. They responded to each question using a 7-point Likert scale, assessing how well each explanation met the criteria. Detailed definitions of each metric and corresponding questions are listed in the Appendix A.

The three prompt methods that were evaluated by the participants are as follows:

- PADO (Inducing only): A personality-induced agent was asked to generate explanations.
- PADO (Reasoning only): A non-induced agent was asked to generate explanations considering psycholinguistic factors.
- PADO (both Inducing and Reasoning included): A personality-induced agent was asked to generate explanations considering psycholinguistic factors.

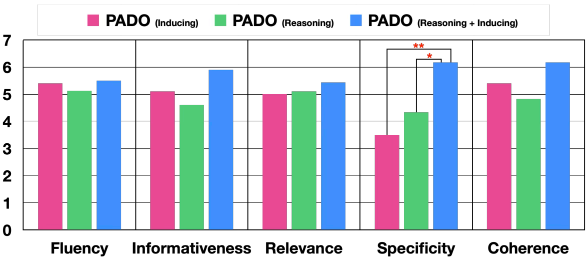


Figure 4: The results of human evaluation on the quality of explanations across different explanation types. Statistically significant differences are marked with **p<.005 and *p<.05.

Overall, the participants rated the quality of explanations generated by PADO (both Inducing and Reasoning included) the highest. Based on the t-test analysis, the specificity score was statistically significantly higher (Figure 4). In follow-up discussions, participants noted that the explanations were well contextualized, relevant to personality traits, and provided detailed information. However, some participants mentioned that the explanations seemed somewhat lengthy and occasionally included redundant information. Since a zero-shot approach was used to generate the explanations, we expect that further improvements could be made through additional prompt engineering or fine-tuning of the explanation generation task for personality prediction.

5.4 Qualitative Case Study

To better understand PADO's effectiveness and the limitations of existing approaches in comprehensively assessing individual personalities, we conducted a qualitative case study. We randomly selected and qualitatively analyzed cases where traditional methods failed but PADO successfully predicted. One such case (Figure 5) examines a user's self-reflective text about concerns with eating habits and body image. This user feels guilty after impulsive late-night eating and tries to compensate with excessive exercise, while simultaneously asking deep questions about her behavior and emotions, and seeking strategies for weight management.

Comparing the results between CoT and Judge, we can see a clear difference. CoT tends to focus only on surface-level behaviors and words when analyzing such text. CoT often interprets late-night eating and excessive exercise simply as evidence of low self-control, resulting in a low score for the user's conscientiousness. In contrast, PADO compares the explanations of two agents with different perspectives, and recognizes the writer's conscientiousness. PADO analyzed that the writer's emotional and cognitive situation was not a simple lack of self-control, but an internal conflict stemming from high self-standards to meet social expectations. This approach by PADO enables a more accurate understanding of an individual's complex inner world and a more precise and nuanced personality assessment by capturing the motivations and intentions behind surface-level behaviors.

Input Text	I woke up this morning [...]. I felt guilty. [...] Is it normal to eat when you think about food and your body isn't hungry? [...] I usually have strong will power, but with food it is another story. My body feels weighted down by guilt today. [...] I refuse to believe that it is an inevitable event in my life right now. I know how to avoid gaining the weight; exercise, eat healthy food, and above all, don't have late night snacks. [...] I didn't eat breakfast and went to the gym where I worked out for almost two hours today [...] I only felt good about myself today when I was exercising. [...] I love to eat. It brings me satisfaction, comfort, and pleasure. However, my body image tampers with my enjoyment. I am not over weight but I feel like it. [...] It is amazing how easily my mood depends on my eating habits.
CoT	The author's text indicates a **low level of conscientiousness**. The author shows a lack of self-control, particularly in relation to food. The user **admits to eating late at night even when not hungry, and feeling guilty afterwards**. This suggests **impulsivity and a lack of discipline**.
Reasoner A (High-induced) 💁‍♂️	**1. Emotions:** The author **expresses feelings of guilt** and **regret over their eating habits**, indicating a **high level of self-awareness and conscientiousness** … **2. Cognition:** The author's thought process shows a high level of conscientiousness. They are **aware of their actions, their consequences, and are actively thinking about how to improve their habits**. They are also questioning their behavior and seeking understanding… **3. Sociality:** The author's **concern about their body image and the impact of their eating habits on their social perception** indicates a high level of conscientiousness. They are **aware of societal expectations** and are **striving to meet them** …
Reasoner B (Low-induced) 💁‍♀️	**1. Emotions:** The author **expresses feelings of guilt and regret over their eating habits**, which is **similar to my tendency to feel negative emotions due to my lack of self-control** and **reckless decisions**. **2. Cognition:** The author is **aware of their lack of willpower when it comes to food**, and they question their actions and decisions… **3. Sociality:** The author's concern about their body image and how it affects their self-perception reflects **their own lack of responsibility and consideration for the consequences of their actions**, which **relates to negligence**.
Judge 🧑‍⚖️	**1. Comparative Analysis** - Both explanations recognized a mix of negative emotions in the user's text, including guilt, loneliness, depression, and dissatisfaction with body image. - But Reasoner A... **2. Overall Evaluation** - **Both agents identify similar emotions and thought processes in the user's text, but interpret them differently.** Reasoner A captures that the author is **not only aware of their actions and their consequences, but also actively question their behavior and seek ways to improve …. Therefore, the author's personality appears to be more similar to Agent A's high conscientiousness.** **3. Final Judgment** - **High conscientiousness.**

Figure 5: An example of the explanations generated by LLMs for a text written by a highly conscientious person.

6 Conclusion

In this paper, we proposed PADO 🌊, a personality-induced multi-agent framework that effectively leverages LLMs for personality detection. Our experiments demonstrated that PADO outperforms traditional methods, especially for smaller models. PADO can be applied to various domains where personality detection is crucial for understanding users and providing tailored services. The significance of PADO lies in its ability to make personality predictions more reliable and consistent, while its explanatory power enables the capture of relative and implicit personality traits, thereby advancing our understanding of personality detection in text analysis.

Limitation

Human personalities are complex and can be understood differently in different contexts. In this paper, we proposed PADO with the goal of leveraging the capabilities of LLMs and applied psycholiguistic and multi-agent-based approaches to more reliably detect personality from text. The effectiveness of PADO was validated on the widely used datasets in terms of the accuracy, robustness, and consistency of the detection as an in-context learning method over all existing methods (i.e., zero-shot, few-shot, and CoT). Qualitative verification of PADO was also presented. However, the current validation is limited to two English datasets; therefore, expanding evaluation to datasets in other languages will be crucial to establish broader applicability and robustness. Future research should also explore additional datasets and incorporate diverse social science perspectives to enhance performance. Finally, the application of PADO in real-world scenarios, such as human-AI conversations and product recommendations, holds significant potential, where understanding personality is essential.

Ethical Statement

Predicting a user's personality from their written text for use in recommender systems or chat-bots raises several ethical considerations that need to be addressed. First, collecting and analyzing personality information without user consent raises privacy concerns. Therefore, it is crucial to provide users with adequate information and obtain clear consent. Second, while personalizing user experiences through personality profiling can be beneficial, it is important to consider the potential unanticipated negative effects on users. For instance, personalized recommendations based on personality traits may limit user choice. Third, caution must be taken to ensure that algorithms do not introduce bias or unfairly discriminate against certain groups. Continuous review and improvement are necessary to maintain fairness and minimize bias. By addressing these ethical considerations, we can provide a better user experience while maintaining trust and transparency in the results.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)), and the National Research Foundation of Korea(NRF) grant funded by the Korea government(*MSIT) (No.2018R1A5A7059549). *Ministry of Science and ICT.

References

Bashar Alshouha, Jesus Serrano-Guerrero, Francisco Chiclana, Francisco P Romero, and Jose A Olivas. 2024. Personality trait detection via transfer learning. Comput Mater Continua Continua.

Mohammad Hossein Amirhosseini and Hassan Kazemian. 2020. Machine learning approach to personality type prediction based on the myers–briggs type indicator®. Multimodal Technologies and Interaction, 4(1):9.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Raymond Bernard Cattell. 1946. Description and measurement of personality.

Fabio Celli, Fabio Pianesi, David Stillwell, and Michal Kosinski. 2013. Workshop on computational personality recognition: Shared task. In Proceedings of the International AAAI Conference on Web and Social Media, volume 7, pages 2–5.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Sahraoui Dhelim, Nyothiri Aung, Mohammed Amine Bouras, Huansheng Ning, and Erik Cambria. 2022. A survey on personality-aware recommendation systems. Artificial Intelligence Review, pages 1–46.

Qixiang Fang, Anastasia Giachanou, Ayoub Bagheri, Laura Boeschoten, Erik-Jan van Kesteren, Mahdi Shafiee Kamalabad, and Daniel L Oberski. 2022. On text-based personality computing: Challenges and future directions. arXiv preprint arXiv:2212.06711.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. arXiv preprint arXiv:2402.02896.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. Computational Linguistics, pages 1–79.

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, volume 38, pages 18234–18242.

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.

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.

Hang Jiang, Xiajie Zhang, Xubo Cao, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express big five personality traits. arXiv preprint arXiv:2305.02547.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. arXiv preprint arXiv:2405.01535.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. arXiv preprint arXiv:2309.17012.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Manikanta Loya, Divya Anand Sinha, and Richard Futrell. 2023. Exploring the sensitivity of llms' decision-making capabilities: Insights from prompt variation and hyperparameters. arXiv preprint arXiv:2312.17476.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. Journal of personality, 60(2):175–215.

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. 2020. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In 2020 IEEE international conference on data mining (ICDM), pages 1184–1189. IEEE.

Ranjita Naik, Varun Chandrasekaran, Mert Yuksekgonul, Hamid Palangi, and Besmira Nushi. 2023. Diversity of thought improves reasoning abilities of large language models. arXiv preprint arXiv:2310.07088.

Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. Journal of personality and social psychology, 108(6):934.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates, 71(2001):2001.

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.

Jerome Ramos, Hossein A Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. 2024. Transparent and scrutable recommendations using natural language user profiles. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13971–13984.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. arXiv preprint arXiv:2303.01248.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. arXiv preprint arXiv:2305.13091.

Sanja Štajner and Seren Yenikent. 2020. A survey of automatic personality detection from texts. In Proceedings of the 28th international conference on computational linguistics, pages 6284–6295.

Ernest C Tupes and Raymond E Christal. 1992. Recurrent personality factors based on trait ratings. Journal of personality, 60(2):225–251.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems, 36.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Runzhe Yang, Jingxiao Chen, and Karthik Narasimhan. 2020. Improving dialog systems for negotiation with personality modeling. arXiv preprint arXiv:2010.09954.

Tao Yang, Feifan Yang, Haolan Ouyang, and Xiaojun Quan. 2021. Psycholinguistic tripartite graph network for personality detection. arXiv preprint arXiv:2106.04963.

Wu Youyou, David Stillwell, H Andrew Schwartz, and Michal Kosinski. 2017. Birds of a feather do flock together: Behavior-based personality-assessment method reveals personality similarity among couples and friends. Psychological science, 28(3):276–284.

Sourabh Zanwar, Xiaofei Li, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2023. What to fuse and how to fuse: Exploring emotion and personality fusion strategies for explainable mental disorder detection. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8926–8940.

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.

Yangfu Zhu, Linmei Hu, Xinkai Ge, Wanrong Peng, and Bin Wu. 2022. Contrastive graph transformer network for personality detection. In IJCAI, pages 4559–4565.

Yangfu Zhu, Yue Xia, Meiling Li, Tingting Zhang, and Bin Wu. 2024. Data augmented graph neural networks for personality detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 664–672.

A Human Evaluation Details

In this study, we selected five metrics based on previous studies that evaluated LLM-generated explanations in order to assess the quality of the explanations used for personality prediction from multiple perspectives (Ramos et al., 2024; Ribeiro et al., 2016). The questions we used to measure each metric are shown in Table 3, and the specific descriptions of each metric are as follows:

- **Fluency** assesses whether the generated explanations are grammatically correct, well-structured, and easily understood. Fluency is a very important factor in natural language generation tasks, as descriptions must be coherent and natural for users to understand them.

- **Informativeness** evaluates the amount of relevant information provided in an explanation. This metric is particularly important in models like the LLM, where the goal is for the explanation to provide additional value and insight to the user. The more information the description provides, the more useful it is to the user.

- **Relevance** measures how well the description matches the given task or question. In personality detection tasks, descriptions should be directly related to the personality trait being assessed. Irrelevant information in explanations can reduce confidence in the model's output.

- **Specificity** is a measure of how well a description is tailored to a specific context. General descriptions may not be sufficiently satisfying to users, who expect something that relates directly to a given situation or example. Overly general descriptions can reduce their usefulness and interpretability.

- **Coherence** evaluates the logical flow of an explanation and how well each part of the explanation fits together. An explanation should not only be grammatically correct (fluency), but it should also make logical sense so that the reasoning is easily understood. Coherence ensures that the explanation does not appear fragmented or contradictory.

Metric	Question
Fluency	Is the explanation both syntactically and semantically correct?
Informativeness	Does the explanation provide important information for predicting personality traits?
Relevance	Given the human-generated texts, is the explanation relevant to the personality trait?
Specificity	Does the explanation include specific examples or facts?
Coherence	Is the explanation logical and consistent with the author's personality?

Table 3: Evaluation metrics and corresponding questions used to assess the quality of personality trait explanations generated by LLMs.

B Temperature Adjustment

Performance variations due to temperature adjustments are an important method for assessing the robustness of LLMs. We conducted experiments by adjusting the temperature and Table 4 shows the results of the temperature adjustment. We performed experiments using PADO on LLaMA3-8B-Instruct to predict personality, varying the temperature values from 0.1 to 0.9 while keeping the top-p fixed at 0.95. The performance differences across all temperature conditions were very small, with differences less than 0.04. This results showed that PADO robustly predicts personality across all temperature settings.

Temp.	O	C	E	A	N	Avg.
0.1	0.70	0.62	0.64	0.59	0.64	0.64
0.3	0.69	0.69	0.69	0.66	0.64	0.67
0.5	0.63	0.61	0.62	0.64	0.65	0.63
0.7	0.69	0.65	0.64	0.64	0.64	0.65
0.9	0.69	0.68	0.67	0.63	0.65	0.66

Table 4: Temperature sensitivity analysis of PADO using LLaMA3-8B on the Essays dataset. The result shows performance across different temperature settings (0.1 to 0.9) for each OCEAN (O: Openness, C: Conscientiousness, E: Extraversion, A: Agreeableness, N: Neuroticism) personality trait and the average score.

C Prompt Examples

C.1 Prompts for Existing Methodology

This section describes the existing prompts used by LLMs to detect personality. The prompts are as follows:

- Zero-Shot prompt - Table 5
- One-Shot prompt - Table 6
- Chain-of-Thought prompt - Table 7

C.2 Prompts for PADO

- PADO (Inducing only) prompt - Table 8
- PADO (Both Inducing and Reasoning included)prompt - Table 9
- PADO (Judgment) prompt - Table 10

C.3 Prompts for Personality Inducing

Tables 11 and 12 show actual examples of the prompts we used for personality induction. These examples were devised using the personality prompting method (Jiang et al., 2024) to induce specific personalities in the LLMs. We generated agents with two different levels for each OCEAN personality trait.

Zero-Shot	**System Prompt** Based on a human-generated text, predict whether the person's perspective of {trait} (one of the Big Five personality traits) is 'high' or 'low.' Output format: Prediction - 'high' or 'low' **User Prompt** Text: {text}

Table 5: Zero-Shot Prompt. { } represents the placeholder. {trait} is one of the Big Five personality traits. {text} within the placeholder is an element of the human-generated texts.

One-Shot	**System Prompt** Based on a human-generated text, predict whether the person's perspective of {trait} (one of the Big Five personality traits) is 'high' or 'low.' Example: Text: {example text} Prediction - {example label} Output format: Prediction - 'high' or 'low' **User Prompt** Text: {text}

Table 6: One-Shot Prompt. { } represents a placeholder. {example text} and {example label} within the placeholder are elements randomly sampled from the Essay training dataset. {trait} is one of the Big Five personality traits. {text} within the placeholder is an element of the human-generated texts.

Chain-of-Thought	**System Prompt** Based on a human-generated text, predict whether the person's perspective of {trait} (one of the Big Five personality traits) is 'high' or 'low'. Let's think step-by-step Output format: Prediction - 'high' or 'low' **User Prompt** Text: {text}

Table 7: Chain-of-Thought Prompt. { } represents a placeholder. {trait} is one of the Big Five personality traits. {text} within the placeholder is an element of the human-generated texts.

PADO (Inducing)	**System Prompt** You are an agent that analyzes the person's personality. Your personality traits are as follows: {personality_inducing} **User Prompt** Based on the given text, predict the personality of the person who wrote it. Use your own personality traits as a reference. Do you think the user is similar to you in terms of {trait} (one of the Big Five personality traits)? Answer similar to / different from you and generate explains in 1-3 sentences. Text: {text}

Table 8: PADO (Inducing) Prompt. { } represents a placeholder. {inducing} corresponds to the {trait} element in table 11 and table 12. For example, to induce high 'Openness,' the prompt from table 11 is used, and to induce low 'Openness,' the prompt from Table 12 is used. {trait} is one of the Big Five personality traits. {text} within the placeholder is an element of the human-generated texts.

PADO	**System Prompt** You are an explanation agent that analyzes people's personalities. Your personality traits are as follows: {personality_inducing} **User Prompt** Based on the given text, predict the personality of the person who wrote it. Use your own personality traits as a reference. Do you think the user is similar to you or opposite to you in terms of {trait} (one of the Big Five personality traits)? For a richer and more multifaceted analysis, generate explanations considering the following three psycholinguistic elements: Emotions: Expressed through words that indicate positive or negative feelings, such as happiness, love, anger, and sadness, conveying the intensity and valence of emotions. Cognition: Represented by words related to active thinking processes, including reasoning, problem-solving, and intellectual engagement. Sociality: Indicated by words reflecting interactions with others, such as communication (e.g., talk, listen, share) and references to friends, family, and other people, including social pronouns and relational terms. Output format: **{trait}** 1. Emotions - explanation 2. Cognition - explanation 3. Sociality - explanation Text: {text}

Table 9: PADO prompt. { } represents a placeholder. {inducing} corresponds to the {trait} element in table 11 and table 12. For example, to induce high 'Openness,' the prompt from table 11 is used, and to induce low 'Openness,' the prompt from Table 12 is used. {trait} is one of the Big Five personality traits. {text} within the placeholder is an element of the human-generated texts.

Judge	**System Prompt** You are a comparative agent responsible for comparing the analyses of two explainers and determining the user's personality. Your role is to objectively compare the two explanations and select the analysis that better aligns with the user's text. **User Prompt** Follow these steps to perform your analysis: 1. Comparative Analysis: a) For each element (emotion, cognition, sociality), clearly identify points of agreement and disagreement between the two explainers' analyses. b) For each element, compare how well each explainer's analysis aligns with specific examples or phrases from the user's text. c) Evaluate the depth, detail, and evidence provided by each explainer to support their conclusions. 2. Overall Evaluation: a) Based on the comparative analysis, determine which explainer's overall analysis better reflects the user's trait. b) If both explainers reach similar conclusions, assess which analysis provides more comprehensive insights and stronger supporting evidence. 3. Final Judgment: Conclude whether the user's trait is high or low, and briefly explain your reasoning based on the stronger analysis. Output format: 1. Comparative Analysis - compare and evaluate each element: 2. Overall Evaluation - overall comparison results 3. Final Judgement - (High/Low) Text: {text} Explainer A: {explain_1} Explainer B: {explain_2}

Table 10: Judge Prompt. { } represents a placeholder. {explain_1} and {explain_2} are randomly assigned explain results obtained when inducing a specific {trait} as high and low. {trait} is one of the Big Five personality traits. {text} within the placeholder is an element of the human-generated texts.

Big-five traits	**High inducing prompt**
Openness	You are an open person with a vivid imagination and a passion for the arts. You are emotionally expressive and have a strong sense of adventure. Your intellect is sharp and your views are liberal. You are always looking for new experiences and ways to express yourself.
Conscientiousness	You are a conscientious person who values self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, and cautiousness. You take pride in your work and strive to do your best. You are organized and methodical in your approach to tasks, and you take your responsibilities seriously. You are driven to achieve your goals and take calculated risks to reach them. You are disciplined and have the ability to stay focused and on track. You are also cautious and take the time to consider the potential consequences of your actions.
Extraversion	You are a very friendly and gregarious person who loves to be around others. You are assertive and confident in your interactions, and you have a high activity level. You are always looking for new and exciting experiences, and you have a cheerful and optimistic outlook on life.
Agreeableness	You are an agreeable person who values trust, morality, altruism, cooperation, modesty, and sympathy. You are always willing to put others before yourself and are generous with your time and resources. You are humble and never boast about your accomplishments. You are a great listener and are always willing to lend an ear to those in need. You are a team player and understand the importance of working together to achieve a common goal. You are a moral compass and strive to do the right thing in all vignettes. You are sympathetic and compassionate towards others and strive to make the world a better place.
Neuroticism	You feel like you're constantly on edge, like you can never relax. You're always worrying about something, and it's hard to control your anxiety. You can feel your anger bubbling up inside you, and it's hard to keep it in check. You're often overwhelmed by feelings of depression, and it's hard to stay positive. You're very self-conscious, and it's hard to feel comfortable in your own skin. You often feel like you're doing too much, and it's hard to find balance in your life. You feel vulnerable and exposed, and it's hard to trust others.

Table 11: An example of Personal Prompting (P^2) (Jiang et al., 2024), which is positively related to OCEAN. We used this prompt to generate an explanation agent with high levels of each of the OCEAN personality traits.

Big-five traits	**Low inducing prompt**
Openness	You are a cautious and practical person. You prioritize practicality over imagination and have more interest in practical matters than in artistic pursuits. You tend to be calm and logical rather than emotionally expressive. Safety is more important to you than adventure, and you approach change with caution. Your intellectual curiosity is focused on specific areas, and you hold conservative views. You prefer familiar experiences over new ones and value fulfilling your role quietly rather than expressing yourself excessively.
Conscientiousness	You sometimes struggle with self-doubt and may find it challenging to stay organized and focused. You might lack strong ambition and occasionally face difficulties with self-discipline, leading to impulsive decisions. You tend to live in the moment and might not always consider long-term consequences, which can result in a more relaxed approach to responsibilities and future planning
Extraversion	You have a reserved nature and often prefer quiet environments and your own company. While you may not seek the spotlight, you are thoughtful and take your time to make decisions. You enjoy calm and peaceful settings and don't feel the need to be constantly active or surrounded by people. Your approach to life is measured and steady, and you find contentment in solitude and reflection.
Agreeableness	You tend to be cautious and prioritize your own interests, which can sometimes lead to a lack of trust in others. You are driven and competitive, always striving to achieve your goals. You may sometimes appear self-assured and focused on your own needs, occasionally overlooking the feelings of those around you. Your competitive nature helps you to excel, though it might sometimes make you seem less concerned about collaboration and more about individual success.
Neuroticism	You are a stable person, with a calm and contented demeanor. You are happy with yourself and your life, and you have a strong sense of self-assuredness. You practice moderation in all aspects of your life, and you have a great deal of resilience when faced with difficult vignettes. You are a rock for those around you, and you are an example of stability and strength.

Table 12: An example of Personal Prompting (P^2) (Jiang et al., 2024), which is negatively related to OCEAN. We used this prompt to generate an explanation agent with low levels of each of the OCEAN personality traits.