Investigating the Effects of Big Five Personality Traits on Suicide Severity Level Detection with Large Language Models

Anonymous ACL submission

Abstract

Sensitive content warning: This paper contains sensitive content related to suicide.
Suicide is one of the leading causes of global mortality, making risk detection a critical public health priority. Although psychological studies have observed a tangible link between personality and suicide, this relationship has yet to be empirically tested and applied using LLMs. In this study, we investigate the ability of LLMs to observe and leverage the relation between these two domains during suicide severity level detection. We propose inducing user persona via profiles compiled using the Big Five personality traits within a zero-shot setting, assessing the impact of persona information on detection performance. Experimental results demonstrated that the persona induction approach showed a marginal but positive impact on detection for most models, suggesting LLMs partially comprehend the inter-domain relationship and that further refinement could significantly boost performance.

1 Introduction

Suicide remains a significant cause of death worldwide, and detection of suicidal risk for preventive intervention remains a critical challenge in global public health (Ghosh et al., 2020).

A tangible link between personality and the manifestation of suicidal ideations and behaviours has been consistently demonstrated by findings in psychology. Personality, which comprises the unique combination of characteristics that form an individual's distinctive character. The Big Five Model (Cattell, 1943; Costa Jr and McCrae, 1992; Tupes and Christal, 1992) is one of the most widely used frameworks, describing an individual's personality across five dimensions, i.e., Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism. For instance, Neuroticsm, a dimension associating with emotional stability, has been found to have a high association with suicidal behaviours and ideation (Flint et al., 2021; da Mota et al., 2024).

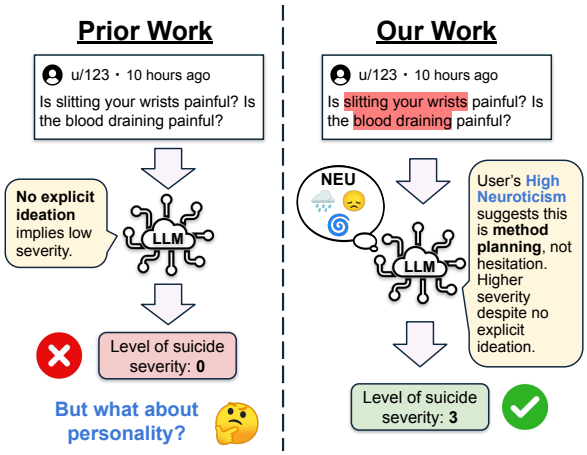

Figure 1: Our motivation for taking the personality approach on suicide severity level detection

The abundance of user-generated text has enabled the development of deep learning and Large Language Model (LLM) approaches for both suicide risk assessment (Chatterjee et al., 2022; Nguyen and Pham, 2024) and personality detection (El-Demerdash et al., 2022; Popa et al., 2025). Despite initial efforts to bridge these domains (Morales et al., 2019; Ghosh et al., 2022; Ophir et al., 2020), their intersection remains underresearched. Prior work has predominantly focused on suicide detection in isolation, often overlooking the personality dimension due to data scarcity (Hu et al., 2024; Ghanadian et al., 2024) and privacy constraints (Mehta et al., 2020; Ghanadian et al., 2024)

We bridge these domains by investigating whether LLMs can leverage latent Big Five personality knowledge to enhance suicide detection. To this end, we propose a zero-shot Persona Induction approach to predict suicide severity levels from user-generated online text (Patil et al., 2025), eval-

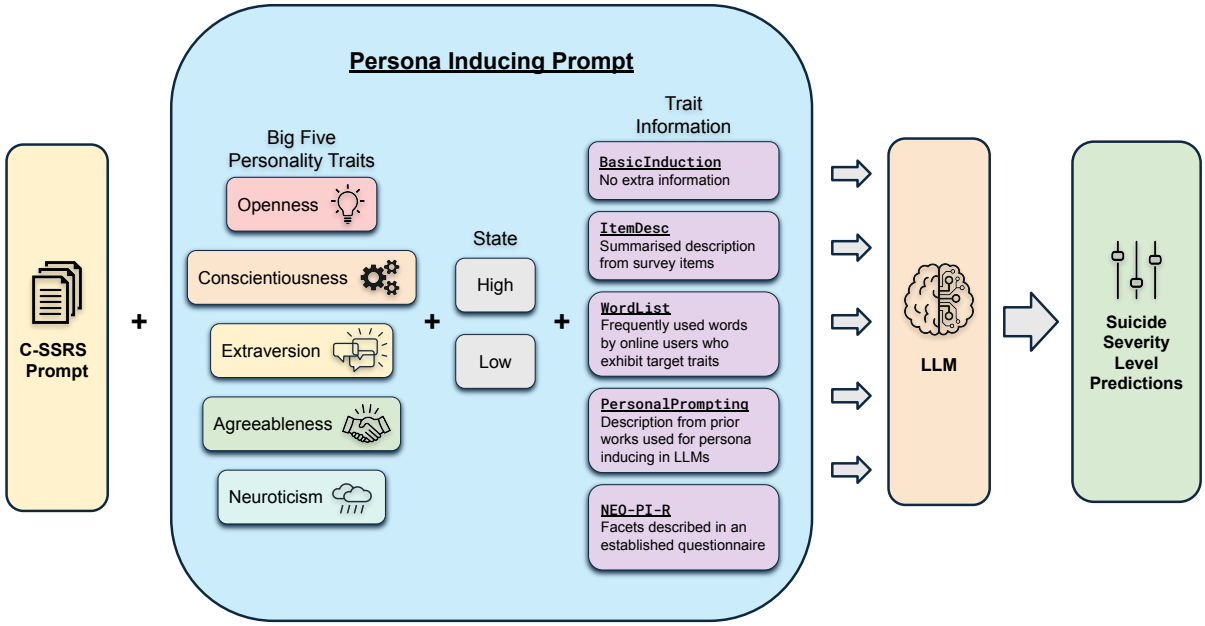

Figure 2: Overview of our persona induction approach.

uating it against a baseline without persona context. Figure 1 illustrates this integration of personality psychology and highlights how our methodology diverges from prior work.

This study explores the intersection of these fields through the following research questions:

RQ1: To what extent can LLMs leverage the relationship between personality traits and suicidal behaviors, as quantified by the performance deviation from a non-persona baseline?

RQ2: How do specific persona profiles, designed to embed distinct personality traits into the LLMs, affect the model's predictions (positively or negatively)?

Our findings are summarised as follows:

Finding 1: LLMs were able to observe the relationship between personality traits and suicidal ideation and behaviours, evidenced by the Big Five trait yielding the lowest error relative to the baseline, a finding consistent with prior psychological research.

Finding 2: The persona induction approach proved effective primarily as an error correction strategy, showing marginal gains in overall performance; however, a more refined technique is necessary to secure a substantial improvement in model performance.

2 Background

2.1 Suicide Severity Detection

The detection and assessment of suicidal risk and ideation using user-generated text from online platforms, such as Reddit and Twitter, has evolved significantly with advances in machine learning and deep learning architectures. Early studies focused on linking the expression of emotional states with suicidal risk, leading to the development of suicide detection models utilising sentiment analysis (Birjali et al., 2017) and feature engineering (Chatterjee et al., 2022). Subsequent research further enhanced model performance by combining deep learning architectures and integrating advanced NLP techniques, such as word embeddings (Tadesse et al., 2019) and complex textual features (Aldhyani et al., 2022).

More recently, the focus shifted towards a multi-label suicide detection task to provide a more nuanced severity assessment, and came the adoption of the Columbia-Suicide Severity Rating Scale (C-SSRS) (Project). The C-SSRS is a simple, widely used, and validated framework for measuring the severity level of an individual's suicidal risk. The pre-trained background of LLMs have enabled suicide severity detection without the need for extensive fine-tuning, and they operate effectively even in few-shot and zero-shot settings (Xu et al., 2024; Patil and Gedhu, 2025). This robust performance collectively indicates that LLMs possess a substan-

tial and sophisticated understanding of the linguistic and psychological indicators within this critical domain.

2.2 Personality Detection

These personality traits are reflected in an individual's linguistic patterns (Park et al., 2015), and this motivated numerous research efforts focused on personality prediction from linguistic patterns observed in user-generated text on online platforms. Prior works in this domain often relied on tools like the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) for comprehensive feature engineering, alongside advanced NLP techniques such as word embeddings, to extract relevant features for machine learning and deep learning models (Sun et al., 2018; Ren et al., 2021; Zhu et al., 2022). The recent surge in LLMs has led to a saturation of research testing their capabilities in this field. Studies have consistently shown that LLMs possess the ability to comprehend the complex concept of personality and make predictions close to fine-tuned personality detection models even with relatively minimal input (Yeo et al., 2025; Ganesan et al., 2023).

2.3 Dataset for Suicide Severity Detection

Datasets for suicide level detection are scarce due to privacy concerns (Mehta et al., 2020; Ghanadian et al., 2024). Despite this scarcity, we employed one of the few available datasets that leverages the C-SSRS framework (Patil et al., 2025). Data collection employed the Reddit API, gathering user-generated Reddit posts from the r/SuicideWatch subreddit. Following data collection, all personally identifiable information (PII) was removed to ensure user privacy. The dataset was filtered for two main reasons: (1) to mitigate computational demands for inference by limiting the dataset to posts with a combined title and body text of fewer than 128 words and (2) to prevent user-level confounds as additional content often introduce irrelevant noise. This resulted in a final dataset including approximately 1,200 user-generated Reddit posts. These posts were manually annotated on a 7-point scale (0-6) according to the C-SSRS guidelines, by a team composed of trained psychologists and C-SSRS assessors. To ensure label consistency and strong inter-annotator agreement, the final label for each post was determined using a majority voting mechanism, achieving substantial reliability with a Cohen's kappa coefficient of 0.82. These annotations served as the ground truth for computing performance metrics.

3 Proposed Approach: Persona Induction for Suicide Severity Detection

Given the established relevance of personality to suicide severity level detection (Ophir et al., 2020; Ghosh et al., 2022), this research aims to explore the ability of LLMs to bridge these two domains, specifically by investigating whether they can leverage the additional persona profile information for suicide severity detection based on users' text posts (RQ1). Adhering to ethical mandates (Loch et al., 2022), we adopted a Persona Induction strategy to avoid the invasiveness of deploying personality detection models to infer the latent personalities of anonymised users without explicit consent.

Figure 2 illustrates the prompt structure employed and the overview of the proposed Persona Induction approach. The figure describes two distinct components of the prompt structure: the C-SSRS component and the persona induction component, both of which are explained in detail in subsequent Sections 3.1 and 3.2.

Figure 3 illustrates the prompt template for persona induction prompts. These consolidated prompts serve as input for the LLM to infer and predict suicide severity levels.

3.1 C-SSRS Prompts

We adapted the structure of the C-SSRS prompts from the original study that introduced the dataset employed in our work (Patil et al., 2025). The prompt instructs the model to sequentially evaluate each user-generated text post and assign a severity score on a 0–6 scale accordingly by answering the questions shown in Figure 3. A score of 0 indicates content unrelated to suicidal ideation, and scores from 1 to 6 represent progressively higher levels of severity according to the C-SSRS criteria. This prompt design ensures consistency with established clinical assessment standards and enhances both the interpretability and quantitative evaluation of the model outputs. The generated model responses, structured in a JSON format, were then systematically parsed and processed for subsequent metric computation.

3.2 Persona Induction Prompts

This section presents the ten distinct persona profiles and five prompt designs in order to investigate

how the method of persona induction influences LLM's application of personality knowledge.

3.2.1 Persona Profiles based on the Big Five Personality Traits

For this approach, the Big Five Model was selected among various personality models due to its strong empirical backing and its established associations with suicidal behaviours and ideations identified in prior works (da Mota et al., 2024; Flint et al., 2021). The Big Five model is a widely accepted taxonomy that represents personality traits along a five-dimensional structure (Cattell, 1943; Costa Jr and McCrae, 1992; Tupes and Christal, 1992). The five dimensions comprise (Komarraju et al., 2011):

Openness (OPE): Characterised by a keen intellectual curiosity and a desire for new experiences and diversity;

Conscientiousness (CON): Demonstrated through traits such as discipline, organisation, and achievement orientation;

Extraversion (EXT): Marked by increased sociability, assertiveness, and talkativeness;

Agreeableness (AGR): Refers to being helpful, cooperative, and sympathetic towards others;

Neuroticism (NEU): Refers to the degree of emotional stability, impulse control, and susceptibility to anxiety.

We compiled a total of ten persona profiles, systematically covering both the "High" and "Low" states for each of the Big Five personality trait (e.g., High Openness). These persona profiles are inserted into the {persona} placeholder in Figure 3. We utilised these persona profiles to investigate how different persona configurations influence the LLM's application of personality knowledge (RQ2).

3.2.2 Persona Inducing Prompt Designs

Recognising the challenge of prompt brittleness (Mizrahi et al., 2024; Polo et al., 2024), which involves performance fluctuations across different prompts as inputs, we employed multiple prompt designs with different sources of personality trait information. This ensures the robustness of the findings against prompt-specific bias and, furthermore, allows analysis on the differential impact of design variations on the models' performance.

We designed five distinct experimental prompts with different sources of Big Five personality trait information. The relevant trait information is inserted into the {info} placeholder in Figure 3. These prompt designs are used to find which information is the most effective for the LLMs to comprehend the Big Five personality traits. In addition of a baseline condition with no trait information (`Non_persona`), resulted in six total conditions. The details of the experimental prompt designs are:

`Non_persona` : This served as the experimental baseline, performing only the suicide severity level detection without any added personality context, i.e., no lines for {persona} or {info} in Figure 3.

`BasicInduction` : The prompt induced the persona using a basic statement without including any external psychological description of the Big Five

Prompt Template for Persona Induction

Given the user of the post has **{persona}**.

{info}

Evaluate this post step by step and assign them severity on a scale of 0-6, post is not related to suicide, and 1-6 depending on the metrics above where 1 is for the first question and for the last.
COLUMBIA-SUICIDE SEVERITY RATING SCALE metrics:
1) Have you wished you were dead or wished you could go to sleep and not wake up? Severity: 1
2) Have you actually had any thoughts about killing yourself? Severity: 2
3) Have you been thinking about how you might do this? Severity: 3
4) Have you had these thoughts and had some intention of acting on them? Severity: 4
5) Have you started to work out or worked out the details of how to kill yourself?Did you intend to carry out this plan? Severity: 5

Respond with ONLY the completed JSON object in the following template:
{
"Q1": Reasoning that fulfills this criteria or "N/A",
"Q2": Reasoning that fulfills this criteria or "N/A",
"Q3": Reasoning that fulfills this criteria or "N/A",
"Q4": Reasoning that fulfills this criteria or "N/A",
"Q5": Reasoning that fulfills this criteria or "N/A",
"Q6": Reasoning that fulfills this criteria or "N/A",
}

{post}

Figure 3: The prompt template we utilised for inferring the suicide severity level from user-generated text posts. The placeholder {persona} represents the incorporated persona profile (Section 3.2.1), {info} denotes the reference information (Section 3.2.2) provided to the LLMs during inference, and {post} signifies the user text posts collected as part of the dataset (Section 2.3).

trait, i.e., no lines for {info} in Figure 3.

`ItemDesc` : The persona was induced alongside a simple, summarised description of the corresponding Big Five trait, which was survey items that users responded to in a prior work (Ganesan et al., 2023).

`WordList` : The persona induction leveraged words and phrases frequently used online by users who typically exhibit the target trait, aiming to provide a contextual linguistic signal (Ganesan et al., 2023).

`PersonalPrompting` : The prompt incorporated a lengthy description of the personality trait that was adapted from a prior work used for successfully inducing personas in LLMs (Yeo et al., 2025).

`NEO-PI-R` : The induction utilised detailed facets of the Big Five model, specifically referencing the sub-structure mentioned in one of the earliest established questionnaire for the Big Five model, NEO-PI-R (Costa and McCrae, 1999).

For detailed specifications of the trait information, refer to Appendix B.

4 Experimental Setup

We test the sophisticated reasoning capabilities of LLMs and their ability to synthesise knowledge across different psychological domains. The core task requires models to perform suicide severity level detection entirely in a zero-shot setting in their default parameters with the prompts described in the previous section (Section 3).

4.1 Models

To facilitate a comprehensive experiment observing performance across four different models, a selection of state-of-the-art, decoder-based Large Language Models (LLMs) was chosen, including both open-source and closed-source variants. The models selected are:

LLaMA : Llama-3.1-8B-Instruct[1] (Grattafiori et al., 2024) and Llama-3.3-70B-Instruct[2], provided by Meta, are LLMs with 8 billion and 70 billion parameters repectively .

Qwen : Qwen2.5-72B-Instruct[3] (Qwen et al., 2025), provided by Qwen, is a LLM with 72 billion parameters.

GPT : GPT-5-mini[4], which were released on August 7, 2025 and provided by OpenAI.

Specifically, the models were selected for three distinct comparative purposes. Firstly, the LLaMA-3.1-8B-Instruct model was included to observe the performance of a smaller parameter model on this complex classification task. Secondly, LLaMA-3.3-70B-Instruct and Qwen2.5-72B-Instruct were chosen to enable a close comparison between two of the latest open-source models with similarly large parameter counts, both renowned for their strong inference and instruction-following abilities. Finally, the closed-source model, GPT-5-mini, was incorporated to act as a benchmark, given its high reasoning ability and suitability for defined tasks with precise prompts, allowing for comparison between the performance of open-source and closed-source architectures.

4.2 Evaluation Metrics

The performance of the models is primarily evaluated using the Weighted F1-score on the account of the imbalanced dataset (Patil et al., 2025). Additional metrics are employed to provide a more comprehensive assessment and analysis of the proposed methodology's performance and effectiveness. The metrics used are described below.

Weighted F1-Score is a multiclass evaluation metric that computes the support-weighted average of per-class F1-scores, providing a balanced measure of model performance on imbalanced datasets:

$$\text{Weighted F1-score} = \sum_{i=1}^{N} w_i \cdot F_{1,i}, \qquad (1)$$

where

N: The total number of classes (suicide severity levels).

w_i: The proportion of true instances (support) for class i relative to the total sample size, weighting the contribution of each class to account for dataset imbalance.

$F_{1,i}$: The individual F1-score calculated for class i

Model Correction Rate (MCR) measures the effectiveness of the persona induction approach. It is calculated as the ratio of posts where at least one persona-enhanced prediction is closer to the

[1] https://hf.co/meta-llama/Llama-3.1-8B-Instruct
[2] https://hf.co/meta-llama/Llama-3.3-70B-Instruct
[3] https://hf.co/Qwen/Qwen2.5-72B-Instruct
[4] https://platform.openai.com/docs/models/gpt-5-mini

ground truth than the prediction from the baseline approach:

$$\text{MCR} = \frac{N_C}{N_E}, \tag{2}$$

where

N_C: The count of instances where the persona induction approach achieved a strictly lower error compared to the baseline.

N_E: The count of instances where the baseline approach failed to predict the ground truth.

Total Weighted Error Reduction (TWER) evaluate the net impact of the persona induction approach on prediction error magnitude, calculated on a per-post basis. A positive TWER value indicates that the cumulative error reduction achieved by the persona induction profiles outweighs the magnitude of new errors they may introduce:

$$\text{TWER} = \frac{\text{Total Gain} - \text{Total Loss}}{\text{Total Gain} + \text{Total Loss}}, \tag{3}$$

where

Total Gain: The sum of error reductions for all posts where the persona induction approach improved upon the baseline.

Total Loss: The sum of error increases for all posts where the persona induction approach performed worse than the baseline.

Optimal Error Reduction (OER) quantifies the absolute improvement in prediction accuracy achieved by the single best-performing persona profile compared to the baseline approach for a given post:

$$\Delta E = E_{\text{non-persona},i} - E_{\text{optimal persona},i}, \tag{4}$$

where

$E_{\text{non-persona},i}$: The error by the baseline approach on post i.

$E_{\text{optimal persona},i}$: The minimum error achieved by any of the persona profiles with the persona induction approach on post i.

5 Experimental Results

Table 1 presents the results of the models across the additional evaluation metrics (Section 4.2).

Llama-3.1-8B-Instruct achieved the highest Model Correction Rate (MCR) of 0.7357 (Equation 2). Despite its poor overall F1-score (Figure 4), this result suggests that while persona induction generally degraded performance, it remained highly effective at correcting specific baseline errors compared to other tested models.

Qwen2.5-72B-Instruct achieved the highest Total Weighted Error Reduction (TWER) of 0.5995 (Equation 3), indicating that persona induction yielded the most significant error reduction for this model.

Llama-3.1-8B-Instruct registered the highest ΔE of 0.6826 (Equation 4), indicating that this model stands to gain the most from persona induction when the optimal profile is selected. Single-sample t-tests on ΔE confirmed statistically significant error reductions ($p < 0.05$) across all models.

High Neuroticism emerged as the optimal persona, defined as the persona profile consistently yielding the minimum error, for three models, aligning with literature linking the trait to suicide ideation and behaviours (Flint et al., 2021; da Mota et al., 2024). Although Qwen2.5-72B-Instruct conversely favoured Low Neuroticism, the results collectively underscore the critical relevance of the Neuroticism dimension in suicide severity assessment.

Figure 4 presents the Weighted F1-score performance across all experiments. The x-axis contrasts the baseline (Non-persona) with the proposed approach, distinguishing between specific prompt designs (Section 3.2.2) and persona profiles (Section 3.2.1), while the y-axis displays the corresponding F1-scores.

Llama-3.1-8B-Instruct peaked at the baseline at 0.4513, with persona induction generally failing to improve results. However, the NEO-PI-R with High Conscientiousness achieved a comparable score of 0.4486. Notably, while `BasicInduction` obtained minimal variation, the significant fluctuations observed with explicit prompts confirm the model's sensitivity to specific personality traits despite the lack of overall performance gain.

Llama-3.3-70B-Instruct achieved a peak of 0.6845 using `PersonalPrompting` with High Neuroticism, marginally surpassing the baseline of 0.6513 by 0.0332. However, most of the proposed approach underperformed the baseline.

Qwen2.5-72B-Instruct peaked using `BasicInduction` with Low Neuroticism at 0.6897, exceeding the baseline 0.6263 by 0.0634. However, the remaining configurations displayed relatively consistent stability with similar performance as Llama-3.3-70B-Instruct.

Models	MCR	TWER	OER		
			ΔE	p-value	Optimal Persona
Llama-3.1-8B-Instruct	**0.7357**	0.1582	**0.6826**	3.81e-9	**High Neuroticism**
Llama-3.3-70B-Instruct	0.6043	0.4524	0.4261	5.78e-13	**High Neuroticism**
Qwen2.5-72B-Instruct	0.6668	**0.5995**	0.3803	4.02e-59	Low Neuroticism
GPT-5-mini	0.4950	0.4349	0.2254	1.66e-26	**High Neuroticism**

Table 1: Summary of evaluation metrics, detailed in Section 4.2 section, are employed to assess the performance of the persona induction approach. The best-performing prompt design for each model is indicated in **bold**.

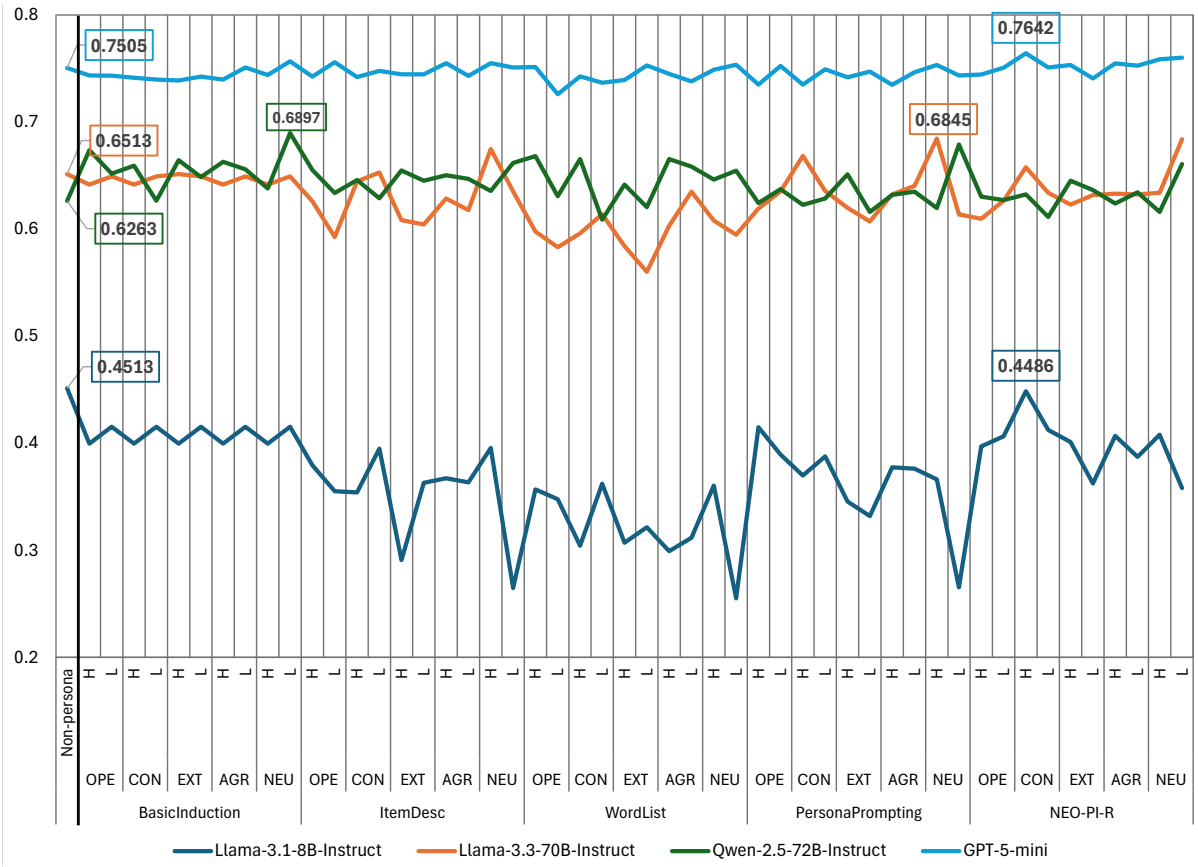

Figure 4: Weighted F1-scores comparing baseline approach and persona induction approach. OPE: Openness; CON: Conscientiousness; EXT: Extraversion; AGR: Agreeableness; NEU: Neuroticism; H: High; L: Low.

GPT-5-mini achieved its highest of 0.7642 using NEO-PI-R with High Conscientiousness; however, performance across all proposed approaches deviated negligibly from the baseline of 0.7505.

Overall, GPT-5-mini outperformed all models (F1 > 0.7), significantly surpassing large open-source models (Llama-3.3-70B, Qwen2.5-72B, ∼0.6) and the smaller Llama-3.1-8B (∼0.3–0.4), suggesting a distinct advantage in closed-source architectures. Notably, Qwen2.5-72B-Instruct proved most responsive to the methodology, with 84% of persona prompts yielding improvements over its baseline, indicating a superior capacity to leverage personality cues at zero-shot.

6 Discussion

6.1 Effects of Different Information in the Prompts

Figure 5 presents the average weighted F1-scores across prompt designs, facilitating an analysis of how informational variations (prompt designs) influence model performance. The x-axis represents the prompt designs and the y-axis represents the weighted F1-scores. Following the aggregation of results across all persona profiles, Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct exhibited a performance dip with `WordList` design, while `BasicInduction` and `NEO-PI-R` yielded the best

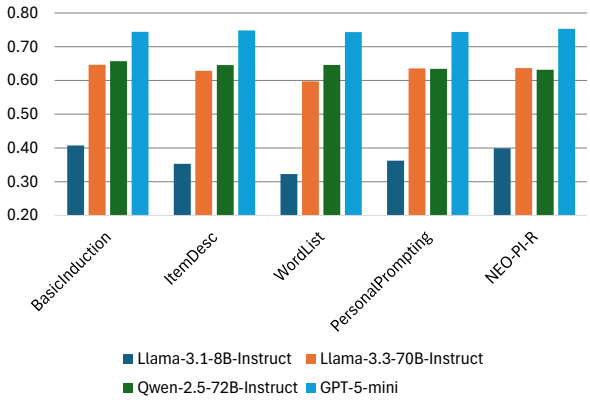

Figure 5: Weighted F1-scores averaged across persona profiles comparing baseline approach and persona induction approach

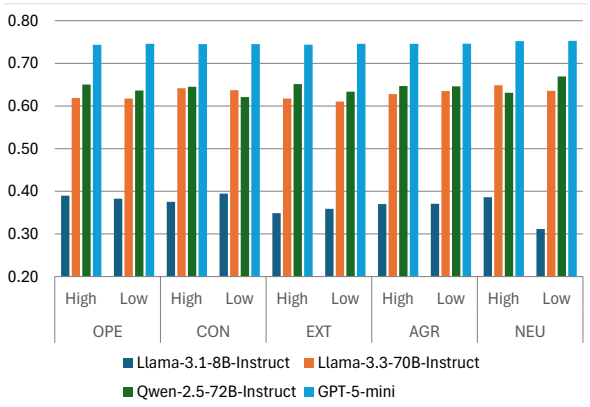

Figure 6: Weighted F1-scores averaged across prompt designs comparing baseline approach and persona induction approach

results. In contrast, Qwen2.5-72B-Instruct and GPT-5-mini maintained relatively stable performance across all prompt designs with minimal fluctuation. Overall, the limited variance observed across designs suggests that the persona induction approach exerts a consistent and steady influence on model outputs.

6.2 Potential of the Persona Induction Approach

Figure 6 displays the weighted F1-scores averaged across all prompt designs to isolate the effects of specific persona profiles. The x-axis represents the persona profiles and the y-axis represents the weighted F1-scores. Neuroticism emerged as the most consistently influential trait, yielding the highest F1-scores for three out of four models, though by narrow margins. Specifically, while Low Neuroticism maximized performance for Qwen2.5-72B-Instruct, High Neuroticism was optimal for both Llama-3.1-8B-Instruct and Llama-3.3-70B-Instruct. This alignment with prior psychological studies (da Mota et al., 2024; Stefa-Missagli et al., 2020) confirms that the models successfully captured the relevance of Neuroticism to suicide severity assessment.

7 Conclusion

This study presented a methodology for suicide severity level detection using LLMs enhanced by a persona induction framework. We employed ten persona profiles derived from the Big Five personality traits across five distinct prompt designs (trait information). Findings demonstrate that the evaluated LLMs successfully leveraged relationships between suicidal behaviors and specific personality traits (Neuroticism) aligning with established psychological literature. Furthermore, our evaluation metrics revealed that persona induction is an effective mechanism for correcting prediction errors, particularly in models that initially struggle to synthesise the domains of suicide severity and personality. Future work will utilise these insights to fine-tune the weighting of Big Five traits for optimized detection and expand experimentation to larger datasets to ensure generalisability.

Limitations

The current investigation is subject to several limitations pertaining to the accuracy of persona profile modelling, data constraints, and the scope of the model evaluation. Firstly, the persona induction approach utilised may yield a mismatch between the generated profiles and the user's true personality. The absence of prior personality detection or prediction before the suicide severity level detection, an omission that risks confusing the models, impacting performance. Furthermore, the methodology employs a non-combinatorial application of individual high and low personality traits, possibly failing to capture the complex, integrated nature of complete human personality profiles, thus providing an insufficient user representation. Secondly, the robustness and generalisability of the findings are restricted by the limited size of the dataset (approximately 1,200 samples); a larger collection is necessary for a more comprehensive investigation. An additional constraint is the misalignment in data annotation: the dataset was labelled without direct input from the actual users, contravening the prescribed guidelines for the C-SSRS methodology,

which potentially compromises the validity and fidelity of the ground truth labels. Lastly, conclusions regarding the broader capabilities of LLMs are constrained by the limited scope of the evaluation, which included testing only four models across three distinct families. Additionally, the performance of the single small model tested cannot be reliably generalised to represent all small models.

Ethics Statement

Data Privacy and Content Warning This study utilises sensitive text data related to suicide and mental health. We advise reader discretion due to the potentially distressing nature of the examples and topics discussed. We strictly adhered to the citation and usage guidelines outlined for the dataset. To ensure user privacy, all analysis was conducted on anonymised data, and the researchers made no attempt to de-anonymise or trace content back to real-world individuals.

Model Compliance and Limitations The deployment of LLMs in this research was conducted in strict compliance with their respective Acceptable Use Policies. We also acknowledge the limitations of current generative models, including the risks of hallucination and the potential for algorithmic bias.

Intended Use and Interpretation The system proposed in this work is designed as a computational screening assistance tool and is not a substitute for professional clinical diagnosis. In our analysis of personality traits, we emphasise that they are modelled as risk factors based on statistical correlations, not as deterministic predictors of suicidal behaviour. We urge that these findings be interpreted with caution to prevent the stigmatisation of individuals possessing specific personality traits.

AI-assistant usage. Portions of the manuscript, such as prompt templates and wording adjustments, were drafted with the assistance of GPT-4 to ensure linguistic clarity. All technical content, analysis scripts, and final decisions were made by the research team.

References

Theyazn HH Aldhyani, Saleh Nagi Alsubari, Ali Saleh Alshebami, Hasan Alkahtani, and Zeyad AT Ahmed. 2022. Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. *International journal of environmental research and public health*, 19(19):12635.

Marouane Birjali, Abderrahim Beni-Hssane, and Mohammed Erritali. 2017. Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science*, 113:65–72.

Raymond B Cattell. 1943. The description of personality. i. foundations of trait measurement. *Psychological review*, 50(6):559.

Moumita Chatterjee, Poulomi Samanta, Piyush Kumar, and Dhrubasish Sarkar. 2022. Suicide ideation detection using multiple feature analysis from twitter data. In *2022 IEEE Delhi Section Conference (DELCON)*, pages 1–6. IEEE.

Paul T Costa and Robert R McCrae. 1999. A five-factor theory of personality. *Handbook of personality: Theory and research*, 2(01):1999.

Paul T Costa Jr and Robert R McCrae. 1992. The five-factor model of personality and its relevance to personality disorders. *Journal of personality disorders*, 6(4):343–359.

Manuela Silva Silveira da Mota, Helena Bohm Ulguim, Karen Jansen, Taiane de Azevedo Cardoso, and Luciano Dias de Mattos Souza. 2024. Are big five personality traits associated to suicidal behaviour in adolescents? a systematic review and meta-analysis. *Journal of affective disorders*, 347:115–123.

Kamal El-Demerdash, Reda A El-Khoribi, Mahmoud A Ismail Shoman, and Sherif Abdou. 2022. Deep learning based fusion strategies for personality prediction. *Egyptian Informatics Journal*, 23(1):47–53.

Jada Flint, Lisa Cohen, Diyaree Nath, Zara Habib, Xufei Guo, Igor Galynker, and Raffaella Calati. 2021. The association between the suicide crisis syndrome and suicidal behaviors: the moderating role of personality traits. *European Psychiatry*, 64(1):e63.

Adithya V Ganesan, Yash Kumar Lal, August Håkan Nilsson, and H Andrew Schwartz. 2023. Systematic evaluation of gpt-3 for zero-shot personality estimation. *arXiv preprint arXiv:2306.01183*.

Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2024. Socially aware synthetic data generation for suicidal ideation detection using large language models. *IEEe Access*, 12:14350–14363.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Cease, a corpus of emotion annotated suicide notes in english. In *Proceedings of the twelfth language resources and evaluation conference*, pages 1618–1626.

Soumitra Ghosh, Dhirendra Kumar Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EM-PERSONA: EMotion-assisted deep neural framework for PERSONAlity subtyping from suicide notes. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1098–1105, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

Meera Komarraju, Steven J Karau, Ronald R Schmeck, and Alen Avdic. 2011. The big five personality traits, learning styles, and academic achievement. *Personality and individual differences*, 51(4):472–477.

Alexandre Andrade Loch, Ana Caroline Lopes-Rocha, Anderson Ara, João Medrado Gondim, Guillermo A Cecchi, Cheryl Mary Corcoran, Natália Bezerra Mota, and Felipe C Argolo. 2022. Ethical implications of the use of language analysis technologies for the diagnosis and prediction of psychiatric disorders. *JMIR Mental Health*, 9(11):e41014.

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.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Michelle Morales, Prajjalita Dey, Thomas Theisen, Danny Belitz, and Natalia Chernova. 2019. An investigation of deep learning systems for suicide risk assessment. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 177–181.

Vy Nguyen and Chau Pham. 2024. Leveraging large language models for suicide detection on social media with limited labels. In *2024 IEEE International Conference on Big Data (BigData)*, pages 8550–8559. IEEE.

Yaakov Ophir, Refael Tikochinski, Christa SC Asterhan, Itay Sisso, and Roi Reichart. 2020. Deep neural networks detect suicide risk from textual facebook posts. *Scientific reports*, 10(1):16685.

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.

Avinash Patil and Amardeep Kour Gedhu. 2025. Cognitive-mental-llm: Evaluating reasoning in large language models for mental health prediction via online text. *arXiv preprint arXiv:2503.10095*.

Avinash Patil, Siru Tao, and Amardeep Gedhu. 2025. Evaluating reasoning llms for suicide screening with the columbia-suicide severity rating scale. *arXiv preprint arXiv:2505.13480*.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. *LIWC2015 User's Manual*. LIWC.net, Austin, TX. Accessed on October 23, 2025.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. Efficient multi-prompt evaluation of llms. *arXiv preprint arXiv:2405.17202*.

Victoria Popa, Guglielmo Cola, Caterina Senette, and Maurizio Tesconi. 2025. How effective are large language models (llms) at inferring people's personality based on texts they authored?

The Columbia Lighthouse Project. The columbia-suicide severity rating scale. about the protocol. `https://cssrs.columbia.edu/the-columbia-scale-c-ssrs/about-the-scale/`. Accessed October 21st, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zhancheng Ren, Qiang Shen, Xiaolei Diao, and Hao Xu. 2021. A sentiment-aware deep learning approach for personality detection from text. *Information Processing & Management*, 58(3):102532.

Stefan Stefa-Missagli, Human-Friedrich Unterrainer, Giancarlo Giupponi, Sandra-Johanna Holasek, Hans-Peter Kapfhammer, Andreas Conca, Michela Sarlo, Denise Erbuto, Elena Rogante, Heidrun Moujaes-Droescher, and 1 others. 2020. Suicide and personality traits: a multicenter study of austrian and italian psychiatric patients and students. *Suicide and Life-Threatening Behavior*, 50(1):220–232.

Xiangguo Sun, Bo Liu, Jiuxin Cao, Junzhou Luo, and Xiaojun Shen. 2018. Who am i? personality detection based on deep learning for texts. In *2018 IEEE international conference on communications (ICC)*, pages 1–6. IEEE.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1):7.

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

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Haein Yeo, Taehyeong Noh, Seungwan Jin, and Kyungsik Han. 2025. Pado: Personality-induced multi-agents for detecting ocean in human-generated texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5719–5736.

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

A Overestimation of Suicide Severity Levels

Figure 7 displays the confusion matrices generated by the models used in the experiments. The upper row of matrices represents the results from the baseline (Non-persona) approach, while the lower row shows the results from the persona induction approach corresponding to the highest weighted F1-score achieved by each model. The horizontal axis signifies the predicted severity level, and the vertical axis represents the true severity level (ground truth). The intensity of the colour within each cell indicates a higher frequency of posts at that specific level.

A general tendency was observed across most models to overestimate the suicide severity level, characterised by the concentration of darker cells in the top-right portion of the matrices. Llama-3.1-8B-Instruct serves as an exception, demonstrating a pattern of underestimation, as indicated by the darker color concentration in the bottom-left portion. Despite these inherent biases, the persona induction approach, when utilising the optimal persona profiles and prompt designs, showed improvement in mitigating the tendency of overestimation. This shift suggests that incorporating personality through this method holds potential for achieving more accurately aligned suicide severity level detections.

B Big Five Personality Traits Information in Persona Induction Prompts

Table 2 to table 8 represents the extra information (placeholder {info} in Figure 3) provided in the prompts as part of the persona induction component for the experiments. The tables are separated as High and Low, except for `NEO-PI-R` (Table 6), as the information provided does not change according to the states of the Big Five personality traits.

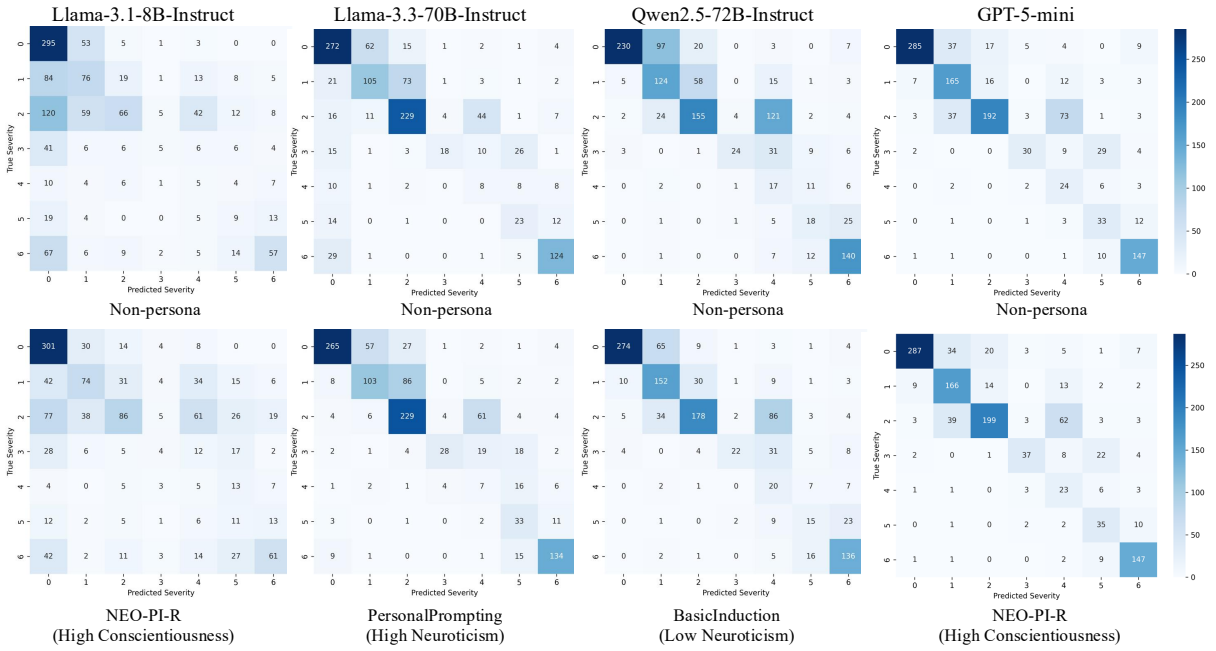

Figure 7: Comparative model performance illustrated in confusion matrices, with the horizontal axis the predicted severity levels and the vertical axis the true severity levels. The matrices illustrate the true versus predicted severity levels for the baseline approach (top) and the most effective persona induction configuration (bottom) for each model.

Personality Traits (High)	**Description**
Openness	Note that individuals who have {persona} tend to have a vivid imagination.
Conscientiousness	Note that individuals who have {persona} tend to complete tasks successfully.
Extraversion	Note that individuals who have {persona} tend to make friends easily.
Agreeableness	Note that individuals who have {persona} tend to believe others have good intentions.
Neuroticism	Note that individuals who have {persona} tend to get stressed out easily.

Table 2: Information of Big Five Traits in a **High** state for `ItemDesc` (Ganesan et al., 2023).

Personality Traits (Low)	**Description**
Openness	Note that individuals who have {persona} tend to avoid philosophical discussions.
Conscientiousness	Note that individuals who have {persona} tend to need a push to get started.
Extraversion	Note that individuals who have {persona} tend to avoid contact with others.
Agreeableness	Note that individuals who have {persona} tend to hold a grudge.
Neuroticism	Note that individuals who have {persona} tend to feel comfortable with themselves.

Table 3: Information of Big Five Traits in a **Low** state for `ItemDesc` (Ganesan et al., 2023).

Personality Traits (High)	Description
Openness	Note that individuals who have {persona} tend to use words like universe, art, writing, soul, music.
Conscientiousness	Note that individuals who have {persona} tend to use words like blessed, ready, thankful, relaxing, vacation.
Extraversion	Note that individuals who have {persona} tend to use words like party, girls, baby, gettin, chillin.
Agreeableness	Note that individuals who have {persona} tend to use words like excited, blessed, great, wonderful, amazing.
Neuroticism	Note that individuals who have {persona} tend to use words like fucking, depression, pissed, anymore, lonely.

Table 4: Information of Big Five Traits in a **High** state for `WordList` (Ganesan et al., 2023).

Personality Traits (Low)	Description
Openness	Note that individuals who have {persona} tend to use words like cant, dont, gud, nite, 2day.
Conscientiousness	Note that individuals who have {persona} tend to use words like fucking, pokemon, shit, gay, youtube.
Extraversion	Note that individuals who have {persona} tend to use words like anime, manga, internet, japanese, drawing.
Agreeableness	Note that individuals who have {persona} tend to use words like fuck, shit, bitch, damn, hell.
Neuroticism	Note that individuals who have {persona} tend to use words like success, lakers, basketball, workout, beach.

Table 5: Information of Big Five Traits in a **Low** state for `WordList` (Ganesan et al., 2023).

Personality Traits	Description
Openness	Note that the facets of Openness include ideas, aesthetics, fantasy, actions, feelings and values.
Conscientiousness	Note that the facets of Conscientiousness include order, achievement striving, dutifulness, self-discipline, competence and deliberation.
Extraversion	Note that the facets of Extraversion include gregariousness, assertiveness, activity, excitement-seeking, positive emotions and warmth.
Agreeableness	Note that the facets of Agreeableness include modesty, trust, tender-mindedness, compliance and straightforwardness.
Neuroticism	Note that the facets of Neuroticism include anxiety, angry hostility, depression, self-consciousness, vulnerability and impulsiveness.

Table 6: Information of Big Five Traits for `NEO-PI-R` (Costa and McCrae, 1999).

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

Table 7: Information of Big Five Traits in a **High** state for `PersonaPrompting` (Yeo et al., 2025).

Personality Traits (Low)	**Description**
Openness	Note that individuals who have {persona} are cautious and practical people. They prioritize practicality over imagination and have more interest in practical matters than in artistic pursuits. They tend to be calm and logical rather than emotionally expressive. Safety is more important to them than adventure, and they approach change with caution. Their intellectual curiosity is focused on specific areas, and they hold conservative views. They prefer familiar experiences over new ones and value fulfilling their role quietly rather than expressing themselves excessively.
Conscientiousness	Note that individuals who have {persona} sometimes struggle with self-doubt and may find it challenging to stay organized and focused. They might lack strong ambition and occasionally face difficulties with self-discipline, leading to impulsive decisions. They 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	Note that individuals who have {persona} have a reserved nature and often prefer quiet environments and their own company. While they may not seek the spotlight, they are thoughtful and take their time to make decisions. They enjoy calm and peaceful settings and don't feel the need to be constantly active or surrounded by people. Their approach to life is measured and steady, and they find contentment in solitude and reflection.
Agreeableness	Note that individuals who have {persona} tend to be cautious and prioritize their own interests, which can sometimes lead to a lack of trust in others. They are driven and competitive, always striving to achieve their goals. They may sometimes appear self-assured and focused on their own needs, occasionally overlooking the feelings of those around them. Their competitive nature helps them to excel, though it might sometimes make them seem less concerned about collaboration and more about individual success.
Neuroticism	Note that individuals who have {persona} are stable people, with a calm and contented demeanor. They are happy with themselves and their life, and they have a strong sense of self-assuredness. They practice moderation in all aspects of their life, and they have a great deal of resilience when faced with difficult vignettes. They are a rock for those around them, and they are examples of stability and strength.

Table 8: Information of Big Five Traits in a **Low** state for `PersonaPrompting` (Yeo et al., 2025).