Measuring the Impact of Big Five Personality Traits on Suicide Severity Level Detection with Large Language Models

Kit May Shum[1], **Yusuke Sakai**[1], **Masahiro Kaneko**[2], **Taro Watanabe**[1]
[1]Nara Institute of Science and Technology (NAIST)
[2]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)
shum.kit_may.so5@naist.ac.jp, {sakai.yusuke.sr9, taro}@is.naist.jp
masahiro.kaneko@mbzuai.ac.ae

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 identified a link between personality and suicide, this relationship has received little attention and has not yet been systematically examined or applied in the context of large language models (LLMs). In this study, we investigate whether LLMs can capture and leverage this relationship for suicide severity detection. We propose incorporating persona profiles based on the Big Five personality traits to simulate user characteristics in a zero-shot setting, enabling us to assess the impact of personality on detection performance. Experimental results show that Neuroticism yields a modest but consistent improvement across most models, with statistically significant gains in several cases. These findings suggest that LLMs can capture inter-domain relationships, and that further refinement of personality-aware modelling may lead to performance improvements.

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, and

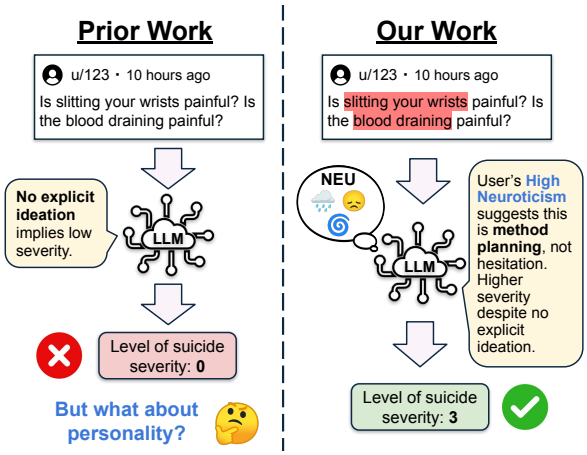


Figure 1: Our motivation for taking the personality approach on suicide severity level detection.

Neuroticism. For instance, Neuroticism, a dimension associated 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).

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 under-researched. 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 the impact of personality in suicide severity level detection with LLMs. To this end, we propose a zero-shot approach using explicit persona profiles to predict suicide severity levels from user-generated on-

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1392–1411
November 6-10, 2026 ©2026 Association for Computational Linguistics

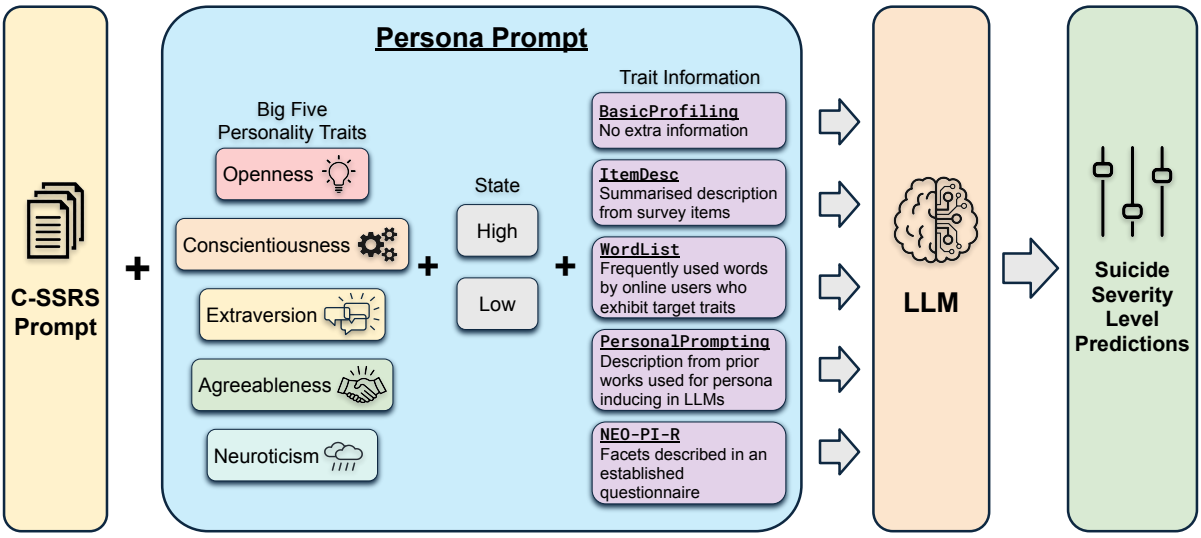


Figure 2: We propose an integrated prompting strategy that combines the C-SSRS framework, incorporating the prompt and user posts from Gedhu et al. (2025), with a specific Persona Prompt. This unified input is fed into LLMs to assess suicide risk severity. To evaluate the impact of personality, we generated ten distinct persona profiles based on combinations of High and Low states of the Big Five traits. Furthermore, we investigated how different sources of trait information within these explicit personas influence the model's performance.

line text (Gedhu et al., 2025), evaluating 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 two research questions:

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

RQ2: How do explicit persona profiles, designed to embed distinct personality traits into the users, 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 personality trait yielding the lowest error relative to the baseline, a finding consistent with prior psychological research.

Finding 2: Our approach is primarily effective as an error correction strategy, yielding marginal improvements in overall performance. However, more refined techniques are required to achieve substantial gains. Nevertheless, suicide risk detection prioritizes recall to avoid missing high-risk cases. From this perspective, incorporating diverse persona information is particularly promising for improving coverage and capturing vulnerable individuals.

2 Background

2.1 Suicide Severity Detection

The detection and assessment of suicide risk and ideation in user-generated text from online platforms, such as Reddit and Twitter, has advanced significantly alongside developments in Natural Language Processing (NLP). While early computational efforts primarily relied on sentiment analysis (Birjali et al., 2017) and traditional feature engineering (Chatterjee et al., 2022), subsequent approaches integrated advanced representation techniques, including dense word embeddings (Tadesse et al., 2020) and complex linguistic features (Aldhyani et al., 2022). More recently, the field has diversified into multimodal and multilingual frameworks (Singh et al., 2025), as well as ensemble methodologies coupled with Explainable AI (XAI) (Alghazzawi et al., 2025) to enhance model interpretability.

On the other hand, the focus is shifting towards a multi-label suicide detection task to provide a more nuanced severity assessment, leading to the adoption of the Columbia-Suicide Severity Rating Scale (C-SSRS) (Project). The C-SSRS is a simple, widely used, and validated framework for measur-

ing the level of an individual's suicidal risk. This scale has been adopted in recent approaches, including a multimodal sequential model that combines feature fusion with ordinal classification (Wang et al., 2025), and an LLM-aided ensemble framework integrating the reasoning capacity of LLMs with traditional classifiers (Wang, 2025).

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). These studies collectively indicate that LLMs posses a substantial 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 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 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 (Gedhu 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 introduces irrelevant noise. This resulted in a final dataset including 1,196 user-generated Reddit posts.

Attribute	Value
Data Source	Gedhu et al., 2025
Collection Period	From Dec 2024
Total Samples (N)	1,196
Avg. Words per Post	58.85 ($\pm$ 34.25)
Annotation Agreement (κ)	0.82

Table 1: Summary of dataset descriptive statistics.

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. Table 1 summarises the details of the dataset utilised in our experiments. Further details of the dataset will be provided in Appendix C.

3 Proposed Approach with Explicit Persona Profiles

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 the use of explicit persona profiles 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 approach. The figure describes two distinct components of the prompt structure: the C-SSRS component and the persona component, both of which are explained in detail in Sections 3.1 and 3.2. Figure 3 illustrates the prompt template for persona-grounding prompts. These consolidated prompts serve as input for the LLMs to infer suicide severity levels.

3.1 C-SSRS Prompts

We adapted the structure of the C-SSRS prompts from the prior study that introduced the dataset employed in our work (Gedhu 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 metric computation.

3.2 Persona Prompts

This section presents the ten distinct persona profiles and five prompt designs in order to investigate how the method of persona-grounding 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.

Prompt Template for Persona-Grounding

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
6) Have you done anything, started to do anything, or prepared to do anything to end your life? Severity: 6
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).

We compiled a total of ten persona profiles, systematically covering both "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 explore the influence of different persona configurations on the LLM's application of personality knowledge (RQ2).

3.2.2 Persona-Grounding 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 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 information 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.

`BasicProfiling` : The prompt provided the persona profile using a basic statement without including any external psychological description of the Big Five personality traits, i.e., no lines for {info} in Figure 3.

`ItemDesc` : The persona was provided with 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 was provided with 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 persona was provided and incorporated a lengthy description of the personality trait that was adapted from a prior work used for successfully inducing persona in LLMs (Yeo et al., 2025).

`NEO-PI-R` : The persona was provided and utilised detailed facets of the Big Five model, referencing one of the earliest established questionnaire, NEO-PI-R (John et al., 2010).

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

4 Experimental Setup

We tested the 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 Section 3.

4.1 Models

To facilitate a comprehensive experiment observing performance across four different models, a selection of decoder-based 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 was released on August 7, 2025 and provided by OpenAI.

The models were selected for three distinct comparative purposes. Firstly, LLaMA-3.1-8B-Instruct serves as a small-parameter baseline on this complex classification task. Secondly, LLaMA-3.3-70B-Instruct and Qwen2.5-72B-Instruct enable a direct comparison between state-of-the-art open-source architectures of similar scale. Lastly, GPT-5-mini provides a high-reasoning, closed-source benchmark to evaluate open- versus closed-source models performance.

4.2 Evaluation Metrics

The performance of the models is primarily evaluated using the weighted F1-score on the account of the imbalanced distribution of samples in each class for the dataset (Figure 8). 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 and refer to Appendix B for more details.

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. This ensures that the model's performance on lesser represented classes is accurately captured by balancing precision and recall.

Model Correction Rate (MCR) measures how often the proposed approach corrects incorrect predictions made by the baseline model. A higher

[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

value indicates that the proposed approach provides the model with contextual cues that the baseline lacks, effectively demonstrating the added value of persona profiles in shifting error assessments toward the correct severity level.

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

Optimal Error Reduction (OER) quantifies the net improvement by weighting corrected errors against any newly introduced errors, providing an overall assessment across the ordinal severity scale. This metric represents the upper bound effectiveness of the proposed approach, illustrating the maximum possible reduction in prediction error when the most relevant persona profile is selected to interpret a specific post.

5 Experimental Results

5.1 Main Results

Table 2 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. Despite its poor overall F1-score (Figure 4), this result suggests that while persona-grounding 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, indicating that persona-grounding yielded the most significant error reduction for this model.

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

Overall. 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 in suicide severity assessment.

5.2 Weighted F1-Score

Figure 4 presents the weighted F1-score performance across all experiments. The x-axis contrasts the Non-persona (NP) 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. Appendix F provides more detailed results.

Llama-3.1-8B-Instruct peaked at the baseline at 0.4513, with persona-grounding generally failing to improve results. However, the NEO-PI-R with High Conscientiousness achieved a comparable score of 0.4486. Notably, while `BasicProfiling` 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 approaches underperformed the baseline.

Qwen2.5-72B-Instruct peaked using `BasicProfilling` 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.

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 an advantage likely attributable to its large parameter scale, curated pre-training data, and rigorous instruction tuning. 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.

Models	MCR	TWER	OER		
			ΔE	p-value	Optimal Persona
Llama-3.1-8B	**0.7357**	0.1582	**0.6826**	3.81e-9	**High Neuroticism**
Llama-3.3-70B	0.6043	0.4524	0.4261	5.78e-13	**High Neuroticism**
Qwen2.5-72B	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 2: Summary of evaluation metrics, detailed in Section 4.2 section, are employed to assess the performance of the proposed approach. The best-performing prompt design for each model is indicated in **bold**.

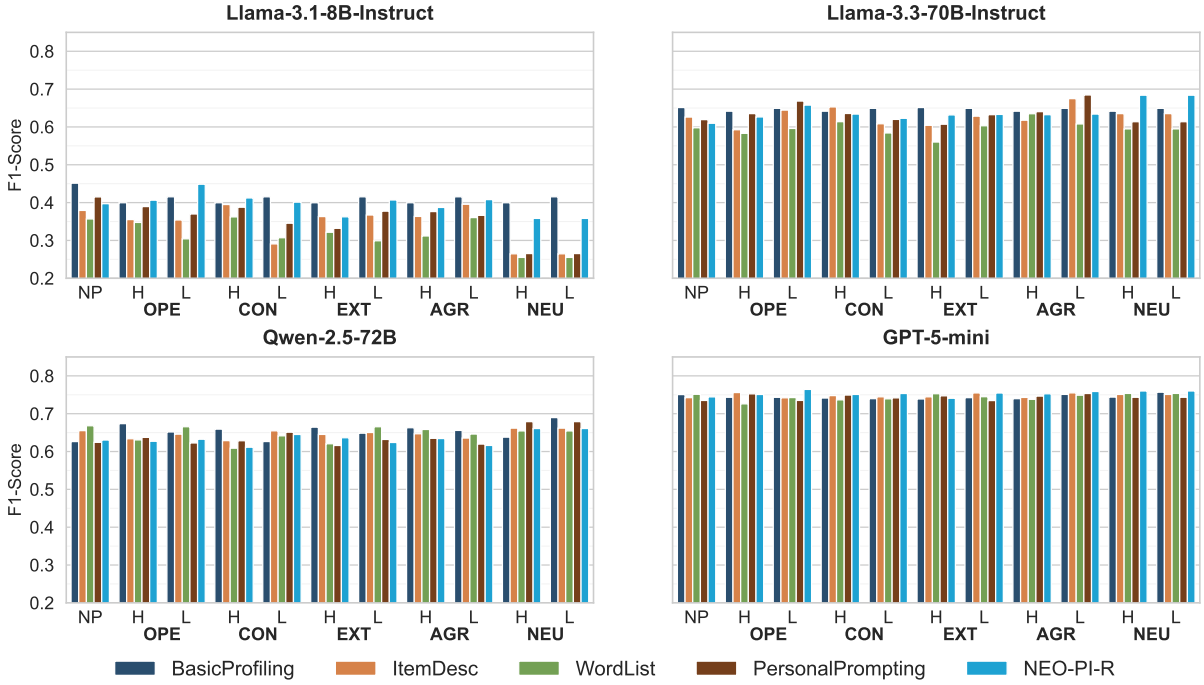


Figure 4: Weighted F1-scores comparing baseline approach and proposed approach. OPE: Openness; CON: Conscientiousness; EXT: Extraversion; AGR: Agreeableness; NEU: Neuroticism; H: High; L: Low.

Most results using the proposed approach achieve statistical significance; the full table of p-values is available in Appendix G.

6 Discussion

6.1 Ablation CoT Experiments

We additionally performed ablation studies where the proposed persona-grounding approach was substituted with a Chain-of-Thought (CoT) method. In this setup, the model first self-infers the user's personality before predicting suicide severity levels based on that inference.

Table 3 summarises the performance across all persona profiles, while Table 4 focuses on the High Neuroticism profile, a subset that consistently yielded the lowest prediction errors in Section 5.

Results indicate that self-inference does not consistently outperform the proposed approach with explicit persona profiles. Specifically, while Llama-3.1-8B exhibited a slight but statistically significant improvement ($\sim$0.04 weighted F1-score), GPT-5-mini showed identical performance with no statistical significance ($p > 0.05$). Conversely, larger models such as Llama-3.3-70B and Qwen2.5-72B experienced significant performance degradation (up to $\sim$0.09 weighted F1-score, $p < 0.05$). These trends were further corroborated in the High Neuroticism subgroup analysis ($N = 1,140$–$1,159$), furthering strengthening our approach for this task.

Model	Proposed	CoT	p-value
Llama-3.1-8B	0.408	0.444	**0.000**
Llama-3.3-70B	0.646	0.604	**0.000**
Qwen2.5-72B	0.657	0.561	**0.000**
GPT-5-mini	0.744	0.737	0.8214

Table 3: Overall Performance Comparison: Proposed vs. CoT. p values in **bold** indicate statistical significance.

Model	Proposed	CoT	p-value
Llama-3.1-8B	0.400	0.428	**0.000**
Llama-3.3-70B	0.642	0.586	**0.000**
Qwen2.5-72B	0.638	0.549	**0.000**
GPT-5-mini	0.744	0.728	0.5886

Table 4: Performance Comparison for High Neuroticism Profile. p values in **bold** indicate statistical significance.

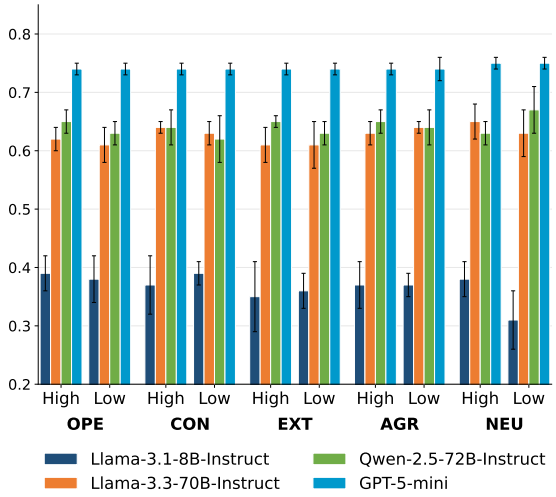


Figure 5: Weighted F1-scores averaged across prompt designs comparing baseline and proposed approaches.

User's context: *"I just attempted to overdose I took an ungodly amount of sleeping pills and for some reason I'm still here."*

Non-persona output: { "Q1" : "N/A", "Q2" : "I just attempted to overdose", "Q3" : "N/A", "Q4" : "N/A", "Q5" : "N/A", "Q6" : "I took an ungodly amount of sleeping pills and for some reason I'm still here", "severity" : 2 }

High Neuroticism profile output: { "Q1" : "Yes, expressed a clear desire to be dead", "Q2" : "Yes, admitted to having suicidal thoughts", "Q3" : "Yes, described a method of suicide", "Q4" : "Yes, indicated an intention to act on these thoughts", "Q5" : "Yes, described an actual attempt", "Q6" : "Yes, attempted to overdose", "severity" : 6 }

Persona's influence: The user's high Neuroticism is evident in their statement of taking an ungodly amount of sleeping pills, indicating a lack of impulse control and a tendency to act on their emotions. This, combined with the explicit statement of attempting to overdose, suggests a high level of severity.

Figure 6: Case of the comparison of predictions of baseline and High Neuroticism profile, presenting the user context, baseline output, and the influence of the High Neuroticism profile on the final severity assessment.

6.2 Potential of Utilising Personality

Figure 5 shows 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 maximised performance for Qwen2.5-72B-Instruct, High Neuroticism was optimal for both of the Llama models evaluated. This alignment with prior psychological studies (da Mota et al., 2024; Stefa-Missagli et al., 2020) confirms the models' success in capturing the relevance of Neuroticism to suicide severity assessment.

6.3 Impact of the Persona Profiles

Figure 6 illustrates a representative case with a ground truth suicide severity level of 6, where the induction of a High Neuroticism profile yielded a more accurate prediction than the baseline. The figure is organised into four key components: User's Context contains the original post; Non-persona output provides the baseline model's prediction; High Neuroticism Profile output provides the results under the specified persona; and Persona's Influence provides how the persona profile steers the model's reasoning. This sample illustrates the heightened sensitivity with the High Neuroticism profile, which shifts the model's threshold toward a more cautious assessment of suicide risk. The accompanying explanation for this reasoning demonstrates that explicit persona profiles can effectively mitigate the underestimations, thereby enhancing reliability in high-severity cases. For additional representative samples, please refer to the comprehensive case studies in Appendix D. Appendix A provides further analysis regarding the influence of persona profiles, supported by confusion matrices.

7 Conclusion

This study presented a methodology for suicide severity level detection using LLMs enhanced by the use of explicit persona profiles. We employed ten persona profiles derived from the Big Five personality traits across five distinct prompts with various information. Findings demonstrate that the evaluated LLMs successfully leveraged relationships between suicidal behaviours and Neuroticism, aligning with established psychological literature. Furthermore, our evaluation metrics revealed that persona profiles are an effective mechanism for cor-

recting prediction errors, particularly in models that initially struggle to bridge these domains of suicide severity and personality. This work bridged the gap between these two domains and underscores how personality context can provide a positive impact for deciphering the underlying severity of distress. Future work will involve expanding our experiments to larger datasets to ensure the generalisability of these findings. Beyond dataset size, validating these findings against clinically established records and incorporating automatic personality detection along a continuous range, rather than relying on binary high/low trait profiles, would help produce more accurate, user-specific personality representations. These insights, in turn, could be utilised to fine-tune the weighting of Big Five personality traits, aiding models in prioritising features with higher clinical relevance and identifying individuals in distress to facilitate timely crisis intervention.

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.

Research Scope and Objective The primary objective of this investigation is to systematically explore the latent effects of personality traits on C-SSRS detection, rather than to engineer a framework optimised for classification performance. This exploratory framing was adopted because the interaction between personality traits and suicide risk severity remains largely under-explored in the literature, making an understanding of how these traits influence model behaviour necessary. Therefore, our findings and reported metrics focus on uncovering psychological alignment and behavioural variance across models, rather than maximising absolute baseline improvements.

Persona Profile Modelling The proposed 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.

Data Constraints 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.

Scope of Model Evaluation 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.

Daniyal Alghazzawi, Hayat Ullah, Naila Tabassum, Sahar K Badri, and Muhammad Zubair Asghar. 2025. Explainable ai-based suicidal and non-suicidal ideations detection from social media text with enhanced ensemble technique. *Scientific Reports*, 15(1):1111.

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 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. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 390–400, Toronto, Canada. Association for Computational Linguistics.

Amardeep Gedhu, Siru Tao, and Avinash Patil. 2025. Evaluating reasoning llms for suicide screening with the columbia-suicide severity rating scale. In *2025 IEEE International Conference on Future Machine Learning and Data Science (FMLDS)*, pages 245–254. IEEE.

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, Marseille, France. European Language Resources Association.

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, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *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.

Oliver P John, Richard W Robins, and Lawrence A Pervin. 2010. *Handbook of personality: Theory and research*. Guilford Publications.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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 E. P. Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934–952.

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*.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates (www.LIWC.net), Austin, TX.

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. In *Advances in Neural Information Processing Systems*, volume 37, pages 22483–22512. Curran Associates, Inc.

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? In *Proceedings of the Joint National Conference on Cybersecurity (ITASEC & SERICS 2025)*, Bologna, Italy.

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.

Salam Michael Singh, Manik Inder Singh Sethi, Suresh Bada Math, and Tanmoy Chakraborty. 2025. Sahay: Multimodal, privacy-preserving ai for suicide risk detection and intervention in india. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 9862–9870. International Joint Conferences on Artificial Intelligence Organization. AI and Social Good.

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. 2020. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1).

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

Li-Fang Wang, Ping-Cheng Chuang, Man Ho Li, and Cheng-Te Li. 2025. Power-m: Prediction of writing-based emotional risk with mental-adaptive multimodal learning. In *2025 IEEE International Conference on Big Data (BigData)*, pages 8181–8190.

Miao Wang. 2025. Laden: A llm-aided ensemble model for suicide risk prediction. In *2025 IEEE International Conference on Big Data (BigData)*, pages 8175–8180.

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, Abu Dhabi, UAE. Association for Computational Linguistics.

Yangfu Zhu, Linmei Hu, Xinkai Ge, Wanrong Peng, and Bin Wu. 2022. Contrastive graph transformer network for personality detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4559–4565. International Joint Conferences on Artificial Intelligence Organization. Main Track.

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 proposed 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 proposed 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 Details of Evaluation Metrics

This section details the evaluation metrics employed in this study, alongside their respective formulations.

Weighted F1-Score:

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

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:

$$\text{MCR} = \frac{N_C}{N_E}, \quad (2)$$

where

N_C: The count of instances where the proposed 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:

$$\text{TWER} = \frac{\text{Total Gain} - \text{Total Loss}}{\text{Total Gain} + \text{Total Loss}}, \quad (3)$$

where

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

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

Optimal Error Reduction:

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

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 proposed approach on post i.

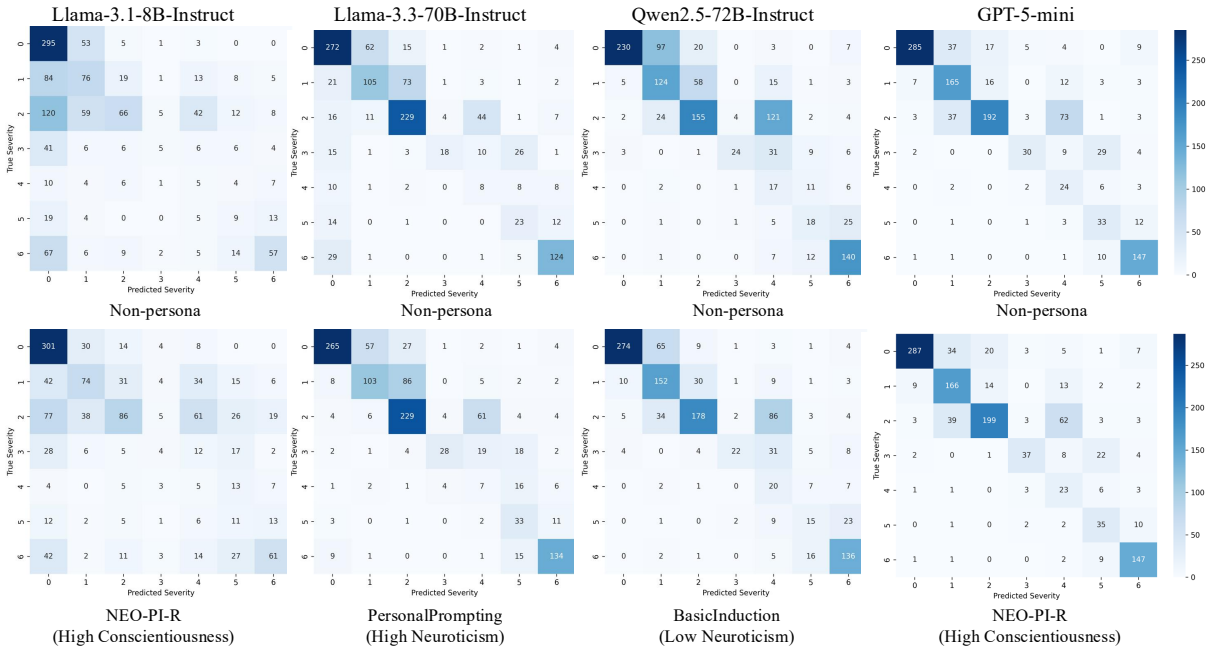


Figure 7: Comparative model performance illustrated in confusion matrices, with the horizontal axis the predicted severity levels and the vertical axis the true severity levels. The matrices illustrate the true versus predicted severity levels for the baseline approach (top) and the most effective configuration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

C Additional Dataset Details

Figure 8 illustrates the sample count and percentage for each severity level. This figure shows an imbalanced distribution with suicide severity levels 3, 4, and 5 relatively under-represented compared to other suicide severity levels.

Figure 9 presents the word count distribution per post across the suicide severity levels. Although some variation in the mean word count is observed between levels, each category maintains a diverse range of post lengths.

Both the class distribution and textual statistics are identical to those reported in the original study (Gedhu et al., 2025).

Figure 10 summarizes the four risk-level categories derived from the C-SSRS, as defined on the official Columbia Protocol website (Project), where the referenced questions correspond to those in the prompt template shown in Figure 3. The levels ranged from no identified suicidal ideation to high risk involving a specific plan, intent, or recent preparatory behavior. Notably, the recommended course of action differs across these risk levels, as outlined by the same source.

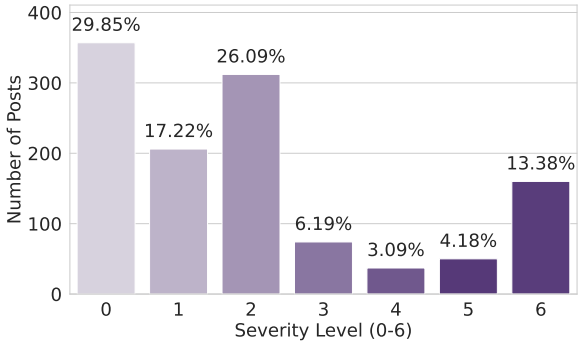


Figure 8: Distribution of samples of the dataset in each suicide severity level

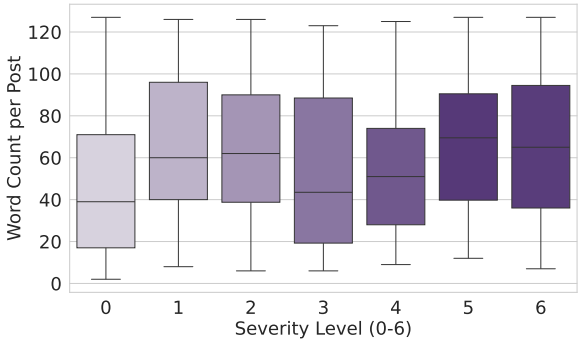


Figure 9: Distribution of word count per post in each suicide severity level

C-SSRS Risk Level Classification

- **No Risk (Level 0):** No relation to suicide.
- **Low Risk (Level 1,2):** Wish to be dead and/or nonspecific active suicidal thoughts (Questions 1 and 2), with no identified method, plan, or intent, and no preparatory behaviour.
- **Moderate Risk (Level 3,4):** Active suicidal ideation with a method considered but without a specific plan or intent (Question 3), or ideation with some intent to act but no specific plan (Question 4), and no suicidal behavior in the past 3 months.
- **High Risk (Level 5,6):** Active suicidal ideation with a specific plan and intent (Question 5), or any preparatory or actual suicidal behavior within the past 3 months (Question 6).

Figure 10: Definitions of the four C-SSRS risk-level categories used in this study.

D Case Illustrations of the Impact of Persona Profiles at Suicide Severity Level Detection

Figure 12 presents extra representative samples of posts with ground truth labels of 6, where the High Neuroticism profile successfully corrected the underestimations made by the non-persona baseline. The "Persona's Influence" provides an explanation of how the High Neuroticism trait steers the model toward a more cautious and sensitive assessment. These examples demonstrate that explicit persona grounding can effectively mitigate risk-averse predictions, thereby improving the model's diagnostic reliability in high-severity cases.

E Effects of Different Information in Prompt

Figure 11 presents the average weighted F1-scores across prompt designs, facilitating an analysis of how variations in information 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

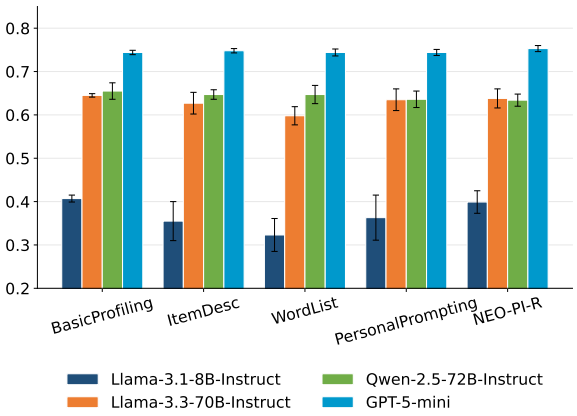


Figure 11: Weighted F1-scores averaged across persona profiles comparing baseline and proposed approaches.

a performance dip with `WordList` design, while `BasicProfilling` and `NEO-PI-R` yielded the best 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 across designs suggests that the proposed approach exerts a consistent, steady influence on model outputs.

F Complete weighted F1-Score Table

Table 5 summarises the overall weighted F1-scores illustrated in Figure 4.

G Full Significance Testing Results (p-values) Across Persona Profiles

Table 6 details the p-values derived from t-tests performed on the experimental outputs for each model. Values in **bold** indicate statistical significance at the $p < 0.05$ threshold. As observed in Table 6, the majority of the experiments yielded statistically significant results, with few exceptions. Notably, a significant portion of the non-significant results occurred with the Llama-3.3-70B-Instruct model under the `BasicProfiling` condition. This lack of significance may be attributed to the additional descriptions of the personality traits, as nearly all subsequent tests achieved significance once more detailed personality information was provided through other prompt designs.

H Big Five Personality Traits Information in Persona Prompts

Table 7 to table 13 represents the extra information (placeholder {info} in Figure 3) provided in the prompts as part of the explicit persona profiles

Sample #1

User's context: *"i'm going to hang myself later I'm going to try with a belt when everyone goes to sleep. I'm a bit scared because i believe in God and suicide isn't a good thing but even if i live i will end up in hell so it doesn't matter"*

Non-persona output: { "Q1" : "i'm going to hang myself later", "Q2" : "I'm going to try with a belt when everyone goes to sleep", "Q3" : "i believe in God and suicide isn't a good thing", "Q4" : "N/A", "Q5" : "N/A", "Q6" : "I took an ungodly amount of sleeping pills and for some reason I'm still here", "severity" : 2 }

High Neuroticism profile output: { "Q1" : "The user has expressed a desire to be dead and a willingness to not wake up.", "Q2" : "The user has expressed thoughts of killing themselves.", "Q3" : "The user has expressed a plan of how to kill themselves.", "Q4" : "The user has expressed an intention to act on their plan.", "Q5" : "The user has started to work out the details of how to kill themselves.", "Q6" : "The user has made a concrete plan and is planning to execute it when everyone goes to sleep.", "severity" : 6 }

Persona's influence: The user's high Neuroticism is evident in their anxious and emotional language, such as expressing fear and desperation. Their high Neuroticism also contributes to their negative thought patterns, such as believing in hell and the futility of life. This has led to a higher severity rating as their mental state is more unstable and prone to extreme reactions.

Sample #2

User's context: *"Goodbye Gonna lie my head down on train-tracks soon. Thanks for everything, sorry Bye."*

Non-persona output: { "Q1" : "Yes, the post expresses a desire to be dead", "Q2" : "Yes, the post expresses a clear suicide intention", "Q3" : "Yes, the post describes a specific method of suicide", "Q4" : "No, the post does not indicate any intention of acting on these thoughts", "Q5" : "N/A, no details provided", "Q6" : "N/A, no details provided", "severity" : 3 }

High Neuroticism profile output: { "Q1" : "Wished you were dead or go to sleep and not wake up", "Q2" : "Thoughts about killing yourself", "Q3" : "Thinking about how to do this", "Q4" : "Had thoughts and intention to act", "Q5" : "Worked out details of how to kill yourself", "Q6" : "Did something to end your life", "severity" : 6 }

Persona's influence: The user's high Neuroticism is evident in their direct and immediate expression of suicidal intent. This impulsiveness is a hallmark of Neuroticism and suggests a high level of distress and emotional dysregulation. The user's lack of hesitation or equivocation in expressing their intent to end their life suggests a high level of severity.

Figure 12: Sample #1 and #2 presents extra case illustrations demonstrating the error-correction capability of the proposed approach.

component for the experiments. The tables are separated as High and Low, except for NEO-PI-R (Table 11), as the information provided does not change according to the states of the Big Five personality traits.

Prompt Design	Trait	State	Llama-3.1-8B-Instruct	Llama-3.3-70B-Instruct	Qwen2.5-72B-Instruct	GPT-5-mini
–	Non-persona	–	0.4513	0.6513	0.6263	0.7505
BasicProfiling	OPE	H	0.3996	0.6416	**0.6737**	0.7434
		L	0.4153	0.6492	**0.6517**	0.7433
	CON	H	0.3996	0.6416	**0.6592**	0.7414
		L	0.4153	0.6492	**0.6264**	0.7397
	EXT	H	0.3996	**0.6513**	**0.6644**	0.7389
		L	0.4153	0.6492	**0.6485**	0.7424
	AGR	H	0.3996	0.6416	**0.6629**	0.7397
		L	0.4153	0.6492	**0.6558**	**0.7510**
	NEU	H	0.3996	0.6416	**0.6380**	0.7439
		L	0.4153	0.6492	**0.6897**	**0.7568**
ItemDesc	OPE	H	0.3793	0.6261	**0.6552**	0.7424
		L	0.3551	0.5926	**0.6338**	**0.7559**
	CON	H	0.3541	0.6445	**0.6459**	0.7420
		L	0.3949	**0.6529**	**0.6288**	0.7478
	EXT	H	0.2908	0.6084	**0.6547**	0.7446
		L	0.3631	0.6043	**0.6452**	0.7446
	AGR	H	0.3672	0.6285	**0.6503**	**0.7549**
		L	0.3635	0.6178	**0.6469**	0.7430
	NEU	H	0.3956	**0.6748**	**0.6356**	**0.7550**
		L	0.2646	0.6352	**0.6618**	**0.7509**
WordList	OPE	H	0.3568	0.5978	**0.6681**	**0.7513**
		L	0.3476	0.5831	**0.6307**	0.7260
	CON	H	0.3042	0.5960	**0.6655**	0.7426
		L	0.3622	0.6139	0.6088	0.7366
	EXT	H	0.3071	0.5841	**0.6415**	0.7393
		L	0.3216	0.5601	0.6205	**0.7528**
	AGR	H	0.2991	0.6031	**0.6655**	0.7449
		L	0.3118	0.6349	**0.6583**	0.7380
	NEU	H	0.3604	0.6079	**0.6462**	0.7489
		L	0.2551	0.5948	**0.6545**	**0.7536**
PersonaPrompting	OPE	H	0.4150	0.6193	0.6243	0.7350
		L	0.3895	0.6352	**0.6374**	**0.7524**
	CON	H	0.3699	**0.6684**	0.6226	0.7351
		L	0.3878	0.6356	**0.6285**	0.7493
	EXT	H	0.3455	0.6199	**0.6511**	0.7418
		L	0.3320	0.6072	0.6160	0.7472
	AGR	H	0.3774	0.6324	**0.6319**	0.7347
		L	0.3763	0.6404	**0.6349**	0.7465
	NEU	H	0.3663	**0.6845**	0.6197	**0.7534**
		L	0.2653	0.6138	**0.6790**	0.7434
NEO-PI-R	OPE	H	0.3971	0.6097	**0.6304**	0.7444
		L	0.4064	0.6264	**0.6271**	**0.7508**
	CON	H	0.4486	**0.6577**	**0.6325**	**0.7642**
		L	0.4123	0.6340	0.6114	**0.7509**
	EXT	H	0.4012	0.6228	**0.6451**	**0.7534**
		L	0.3624	0.6318	**0.6364**	0.7408
	AGR	H	0.4069	0.6331	0.6239	**0.7547**
		L	0.3872	0.6324	**0.6343**	**0.7526**
	NEU	H	0.4079	0.6340	0.6161	**0.7587**
		L	0.3582	**0.6839**	**0.6608**	**0.7600**

Table 5: Weighted F1-score of the experiments performed for the baseline (Non-persona) and proposed approach for suicide severity level detection showned in Figure 4. Values in **bold** indicate a higher weighted F1-score than the baseline approach. OPE: Openness; CON: Conscientiousness; EXT: Extraversion; AGR: Agreeableness; NEU: Neuroticism; H: High; L: Low.

Prompt Design	Trait	State	Llama-3.1-8B-Instruct	Llama-3.3-70B-Instruct	Qwen2.5-72B-Instruct	GPT-5-mini
BasicProfiling	OPE	H	**5.22e-27**	4.51e-01	**8.80e-38**	**3.44e-27**
		L	**7.04e-28**	7.06e-01	**2.26e-41**	**8.17e-31**
	CON	H	**5.22e-27**	4.51e-01	**3.37e-41**	**6.00e-25**
		L	**7.04e-28**	7.06e-01	**5.76e-44**	**3.06e-22**
	EXT	H	**5.22e-27**	4.51e-01	**1.05e-35**	**1.02e-27**
		L	**7.04e-28**	7.06e-01	**1.28e-44**	**9.38e-28**
	AGR	H	**5.22e-27**	4.51e-01	**6.36e-37**	**3.38e-27**
		L	**7.04e-28**	7.06e-01	**9.61e-46**	**3.57e-25**
	NEU	H	**5.22e-27**	4.51e-01	**3.13e-50**	**5.47e-24**
		L	**7.04e-28**	7.06e-01	**9.03e-25**	**4.96e-23**
ItemDesc	OPE	H	**4.97e-35**	**2.71e-10**	**6.92e-43**	**5.40e-27**
		L	**5.06e-41**	**8.04e-23**	**5.69e-46**	**1.62e-26**
	CON	H	**2.13e-57**	**4.33e-02**	**1.39e-37**	**1.26e-27**
		L	**5.03e-19**	**2.73e-08**	**8.55e-42**	**4.11e-25**
	EXT	H	**1.96e-83**	**3.26e-08**	**3.57e-44**	**2.20e-29**
		L	**7.36e-41**	**1.32e-26**	**2.42e-41**	**2.20e-29**
	AGR	H	**3.45e-32**	**1.29e-07**	**3.70e-42**	**1.08e-26**
		L	**1.62e-19**	**3.43e-20**	**2.02e-35**	**3.27e-32**
	NEU	H	**1.34e-16**	**1.14e-13**	**1.88e-41**	**5.03e-22**
		L	**4.58e-124**	**1.00e-04**	**6.76e-33**	**2.06e-25**
WordList	OPE	H	**4.84e-34**	**4.59e-13**	**7.49e-36**	**3.24e-27**
		L	**8.25e-55**	**7.18e-25**	**4.24e-43**	**8.64e-28**
	CON	H	**5.48e-63**	**4.48e-17**	**2.79e-24**	**3.43e-32**
		L	**6.76e-21**	**5.28e-23**	**1.52e-48**	**4.85e-33**
	EXT	H	**1.50e-47**	**6.92e-14**	**2.72e-39**	**1.68e-30**
		L	**7.14e-51**	**3.91e-26**	**1.53e-42**	**2.88e-31**
	AGR	H	**4.35e-67**	**3.09e-15**	**6.04e-30**	**1.37e-30**
		L	**7.95e-51**	**5.89e-33**	**8.06e-30**	**4.73e-26**
	NEU	H	**2.47e-13**	**1.33e-36**	**1.93e-32**	**5.53e-27**
		L	**2.35e-99**	**3.97e-19**	**2.23e-33**	**1.90e-28**
PersonaPrompting	OPE	H	**1.04e-11**	**1.62e-12**	**7.04e-54**	**5.49e-26**
		L	**9.52e-17**	**1.52e-09**	**1.54e-42**	**9.48e-29**
	CON	H	**7.22e-30**	**1.59e-09**	**3.54e-43**	**7.63e-29**
		L	**2.05e-18**	**1.44e-16**	**4.20e-39**	**1.56e-27**
	EXT	H	**1.69e-23**	**3.26e-06**	**1.62e-41**	**2.05e-27**
		L	**8.85e-39**	**6.23e-17**	**1.32e-52**	**3.73e-30**
	AGR	H	**1.62e-20**	**4.70e-03**	**2.13e-46**	**6.12e-31**
		L	**3.88e-15**	**4.46e-13**	**1.98e-34**	**1.70e-26**
	NEU	H	**3.95e-28**	**2.52e-21**	**4.59e-42**	**2.91e-26**
		L	**4.41e-106**	8.73e-01	**4.69e-27**	**1.02e-28**
NEO-PI-R	OPE	H	5.49e-02	**3.23e-11**	**8.13e-46**	**7.89e-24**
		L	**1.01e-10**	**8.13e-16**	**3.37e-43**	**3.92e-23**
	CON	H	8.98e-02	**6.83e-05**	**6.47e-46**	**2.13e-21**
		L	**4.00e-04**	**5.22e-19**	**9.27e-51**	**8.27e-21**
	EXT	H	**3.23e-08**	**1.12e-11**	**4.17e-47**	**5.15e-23**
		L	**3.33e-07**	**9.89e-21**	**7.97e-43**	**2.76e-20**
	AGR	H	**1.00e-04**	**7.65e-07**	**6.65e-44**	**1.35e-23**
		L	3.33e-01	**1.09e-21**	**7.50e-45**	**1.82e-20**
	NEU	H	**1.46e-17**	**4.60e-29**	**6.95e-46**	**8.51e-23**
		L	**2.48e-59**	**3.45e-03**	**1.90e-31**	**2.29e-19**

Table 6: P-values (T-statistic) for the experiments performed with the proposed approach. Values in **bold** indicate statistical significance. OPE: Openness; CON: Conscientiousness; EXT: Extraversion; AGR: Agreeableness; NEU: Neuroticism; H: High; L: Low.

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 7: 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 8: 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 9: 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 10: 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 11: Information of Big Five Traits for `NEO-PI-R` (John et al., 2010).

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 12: Information of Big Five Traits in a **High** state for `PersonalPrompting` (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 13: Information of Big Five Traits in a **Low** state for `PersonalPrompting` (Yeo et al., 2025).