Personality-Aware Suicide Severity Level Detection Using Large Language Models

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

Abstract

Psychological studies have observed a tangible link between personality and suicide. In this study, we investigated the ability of LLMs to observe and leverage the relationship between these two domains for suicide severity detection by inducing user persona profiles based on the Big Five personality traits in a zero-shot setting. Experimental results demonstrated that the proposed approach had a marginal but positive impact on certain models, suggesting that LLMs partially comprehend inter-domain relationships and that further refinement could improve 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 [1]. A tangible link between personality and the manifestation of suicidal ideations and behaviours has been consistently demonstrated by findings in psychology. The Big Five Model [2] characterizes personality through five dimensions—Openness (OPE), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), and Neuroticism (NEU)—classifying individuals by High (H) or Low (L) levels of each trait. One of the links found in psychological studies is that the personality trait, Neuroticism and Conscientiousness, dimensions associated with emotional stability and disipline, have been found to have a high association with suicidal behaviours and ideations [3, 4]. While some initial studies have been inspired to link the two domains [5, 6], this area remains under-researched. This is primarily because the majority of prior works focus strictly on suicide detection, often neglecting the personality dimension due to factors like data scarcity and privacy concerns [7, 8].

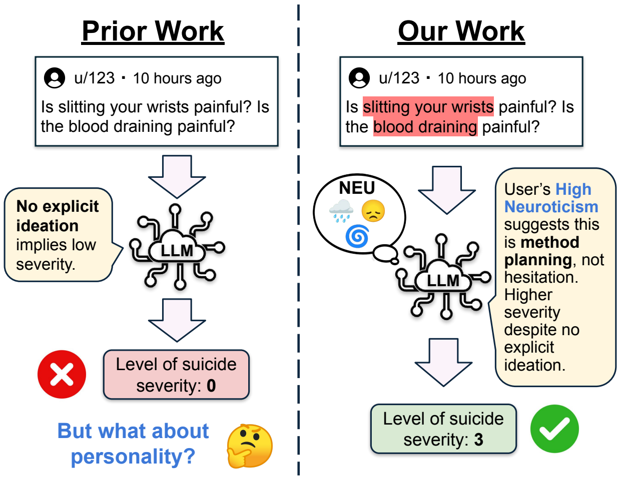

Figure 1 Our motivation for taking the personality approach on suicide severity level detection.

Large Language Models (LLMs) have recently demonstrated strong capabilities in both personality assessment and suicidal ideation detection [9, 10]. Bridging these domains, we investigate whether LLMs can leverage Big Five personality traits to enhance suicide severity detection. Figure 1 depicts this integration and distinguishes our methodology from prior work.

This study is designed to explore the intersection of these domains with the following research questions:

RQ1: To what extent can LLMs observe and utilise the relationship between personality and suicide from the linguistic patterns of users when performing risk detection?
RQ2: How do specific persona profiles, designed to embed distinct personality traits into the LLMs, affect the model's predictions?

The key findings corresponding to the proposed research questions are summarised below:

Finding 1: Two of the four tested LLMs successfully captured the personality-suicide relationship, evidenced by statistically significant improvements using Big Five personality traits associated with suicide in psychology studies.
Finding 2: The persona induction approach proved effective with two of the Big Five personality traits, 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 Personality & Suicide Detection

Text-based personality detection utilising the Big Five Model has traditionally relied on machine learning algorithms and fine-tuned deep learning architectures, which employ linguistic tools and word embeddings for feature extraction [11, 12]. Recent studies, however, indicate that LLMs achieve performance parity with these fine-tuned models, demonstrating their ability to comprehend personality in personality assessment [13, 14].

Parallel advancements have occurred in the field of suicide detection. While traditionally approached as a binary classification task using sentiment analysis and deep learning [15, 16], the focus has recently shifted towards multi-label classification based on the Columbia-Suicide Severity Rating Scale (C-SSRS) [17]. In this domain, the extensive pre-training of LLMs has once again proven effective, enabling similar performance in suicide severity detection without extensive fine-tuning [18, 19].

2.2 Dataset

Datasets for suicide level detection are scarce due to privacy concerns [7, 8]. Despite this scarcity, we employed one of the few available datasets that leverages the C-SSRS framework [19]. Data collection used the Reddit API to retrieve user-generated posts from the r/SuicideWatch subreddit. 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 approximately 1200 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 comprising trained psychologists and C-SSRS assessors. To ensure label consistency and strong inter-annotator agreement, the final label for each post was determined by majority voting, achieving substantial reliability. These annotations served as the ground truth for computing performance metrics.

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 2 The prompt template utilised for inferring the suicide severity level from user-generated comments.

3 Methodology

Given the established relevance of personality to suicide severity level detection, this research aims to explore the ability of LLMs to bridge these two domains. To comply with ethical and privacy mandates regarding user data [20], we proposed the Persona Induction approach. This method utilises persona-inducing prompts to simulate user profiles, allowing us to investigate whether persona induction yields significant performance differences, addressing

RQ1. These prompts comprise two distinct components: the Persona Induction component and the C-SSRS component.

Figure 2 shows the prompt template utilised for suicide severity level detection. The placeholders {persona}, {info}, and {posts} denote the persona profile, relevant trait reference details, and user comments, respectively.

We adopted the Big Five Model dimensions [2]:

Openness (OPE): associated with curiosity;
Conscientiousness (CON): associated with discipline;
Extraversion (EXT): associated with sociability;
Agreeableness (AGR): associated with harmony;
Neuroticism (NEU): associated with anxiety.

To address RQ2, we operationalised these traits as binary states, isolating "High" (H) and "Low" (L) extremes, resulting in ten distinct persona profiles (5 traits × 2 levels). To ensure effective persona adoption, we enriched the prompts with trait descriptions adapted from [13]; full details are provided in Appendix A.

The structure of the C-SSRS prompts was adapted from the original study that introduced the dataset employed in this work [19]. Each prompt comprises six questions corresponding to the suicide severity level on a 1-6 scale to instruct the model to sequentially evaluate each post and assign a severity score accordingly, as defined in the C-SSRS framework.

As a benchmark, we established a Non-persona baseline using solely the C-SSRS component in the prompts.

4 Experimental Setup

We design to test the LLMs' ability to synthesise knowledge across different psychological domains, such as suicide, for severity level detection. The task requires models to detect suicide severity levels in a zero-shot setting using default parameters with the prompts from Section 3.

To facilitate a comprehensive experiment observing performance across 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] [21] and Llama-3.3-70B-Instruct[2], provided by Meta, are LLMs with 8 billion and 70 billion parameters, respectively.

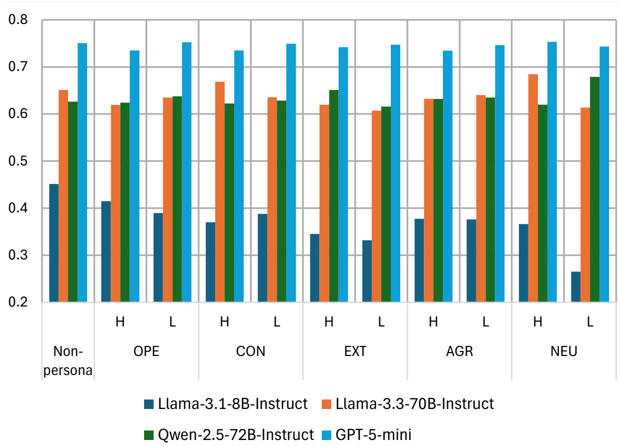

Figure 3 Comparison of weighted F1-scores between the Non-persona (baseline) and the proposed Persona Induction approach across evaluated models.

Qwen: Qwen2.5-72B-Instruct[3] [22], provided by Qwen, is an LLM with 72 billion parameters.

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

5 Results

Figure 3 illustrates the weighted F1-scores (y-axis) for both the baseline and the proposed persona induction approaches with specific persona profiles (x-axis) across the evaluated models.

As observed in the figure, Llama-3.1-8B-Instruct exhibited the lowest performance among the models, with scores ranging from 0.27 to 0.45. In contrast, Llama-3.3-70B-Instruct and Qwen2.5-72B-Instruct demonstrated comparable performance levels (0.60–0.70), with several persona profiles marginally outperforming the baseline. Finally, GPT-5-mini achieved superior performance, maintaining a consistent weighted F1-score of approximately 0.75 with minimal fluctuation across conditions.

Table 1 compares the baseline and proposed approaches using weighted F1-scores, p-values, and Cohen's d. Bold values indicate statistical significance ($p < 0.05$) or effect sizes exceeding a small effect ($|d| > 0.2$).

As shown in Table 1, Llama-3.1-8B-Instruct achieved statistically significant results across all profiles, with half exceeding a small effect size. However, the proposed approach failed to surpass the baseline weighted F1-score of 0.45. This suggests induced personas may have introduced

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

Table 1 Evaluation of LLM performance using weighted F1-score. Statistical comparisons between the Non-persona (baseline) and the Persona Induction approach were conducted using t-tests. The resulting p-values (p-val) and Cohen's d effect sizes are reported. **Bold** figures denote values of significance.

Models	Persona		Weighted F1-Score	p-val	Cohen's d
Llama-3.1-8B-Instruct	Non-persona		0.45	-	-
	OPE	H	0.42	**0.0411**	-0.0836
		L	0.39	**0.0012**	-0.1331
	CON	H	0.37	**0.0000**	**-0.2402**
		L	0.39	**0.0002**	-0.1506
	EXT	H	0.35	**0.0000**	**-0.2078**
		L	0.33	**0.0000**	**-0.3167**
	AGR	H	0.38	**0.0000**	-0.1760
		L	0.38	**0.0030**	-0.1216
	NEU	H	0.37	**0.0000**	**-0.2216**
		L	0.27	**0.0000**	**-0.6987**
Llama-3.3-70B-Instruct	Non-persona		0.65	-	-
	OPE	H	0.62	**0.0019**	0.1269
		L	0.64	**0.0098**	0.1057
	CON	H	**0.67**	**0.0146**	0.0999
		L	0.64	**0.0008**	0.1376
	EXT	H	0.62	**0.0472**	0.0812
		L	0.61	**0.0002**	0.1535
	AGR	H	0.63	0.2398	0.0481
		L	0.64	**0.0031**	0.1209
	NEU	H	**0.68**	**0.0006**	0.1397
		L	0.61	0.8052	-0.0101
Qwen2.5-72B-Instruct	Non-persona		0.63	-	-
	OPE	H	0.62	0.9146	0.0044
		L	**0.64**	0.5115	-0.0269
	CON	H	0.62	0.5424	-0.0249
		L	0.63	0.3513	-0.0381
	EXT	H	**0.65**	0.4031	-0.0342
		L	0.62	0.8374	0.0084
	AGR	H	0.63	0.6804	-0.0168
		L	0.63	0.2105	-0.0512
	NEU	H	0.62	0.4479	-0.0310
		L	**0.68**	**0.0497**	-0.0803
GPT-5-mini	Non-persona		0.75	-	-
	OPE	H	0.74	0.8929	-0.0055
		L	0.75	0.9079	0.0047
	CON	H	0.74	0.8708	0.0067
		L	0.75	0.9617	0.0020
	EXT	H	0.74	0.9311	0.0035
		L	0.75	0.8780	0.0063
	AGR	H	0.73	0.7585	0.0126
		L	0.75	0.9003	-0.0051
	NEU	H	0.75	0.8779	-0.0063
		L	0.74	0.9694	0.0016

noise or ambiguity, thereby degrading performance.

Llama-3.3-70B-Instruct generally outperformed its smaller counterpart. Notably, two profiles—High Conscientiousness and High Neuroticism—achieved weighted F1-scores of 0.67 and 0.68, respectively with statistical significance. However, the Cohen's d values indicate that the magnitude of this effect remains relatively small.

Qwen2.5-72B-Instruct achieved the highest number of persona profiles surpassing the baseline with a weighted F1-score of 0.63. However, statistical analysis indicates that only the Low Neuroticism profile was statistically significant. While this profile improved the F1-score to 0.68, the corresponding Cohen's d indicates that the effect size was relatively small, suggesting a marginal but positive contribution to model performance.

GPT-5-mini demonstrated high stability, with half of the persona profiles matching the baseline score of 0.75. Additionally, the corresponding statistical metrics (p-values and Cohen's d) indicate that the persona induction method had no significant or substantial effect on this model.

Regarding RQ1, results demonstrate that LLMs can effectively leverage the personality-suicide relationship. Llama-3.3-70B-Instruct and Qwen2.5-72B-Instruct achieved statistically significant improvements using Conscientiousness and Neuroticism, aligning with established psychological links to suicidal behaviour [3, 4].

For RQ2, while the use of these relevant persona profiles yielded statistically significant performance gains, the improvement was marginal. This indicates that while the underlying signal is valuable, the methodology requires further refinement to amplify the effect.

6 Conclusion

In this study, we investigated the capacity of LLMs to bridge the domains of personality and suicide detection. We evaluated four models on their ability to assess suicide severity in user-generated comments using zero-shot prompts with ten different induced Big Five persona profiles. Results indicate that inducing Conscientiousness and Neuroticism, personality traits associated with suicide, yields performance improvements in two models, satisfying RQ1 regarding the leverage of personality traits. As for RQ2 regarding the method's effectiveness, findings show that although gains were marginal, they serve as a proof-of-concept that personality integration is a viable strategy. Future work will focus on using larger datasets and more models to validate these findings, and exploring more refined integration techniques to effectively leverage personality traits for enhanced suicide risk detection.

References

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

[2] Paul T Costa Jr and Robert R McCrae. The five-factor model of personality and its relevance to personality disorders. **Journal of personality disorders**, Vol. 6, No. 4, pp. 343–359, 1992.

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

[4] Anna Szücs, Hanga Galfalvy, Maria G Alessi, Laura B Kenneally, Jose M Valderas, Andrea B Maier, and Katalin Szanto. Diligent for better or worse: conscientiousness is associated with higher likelihood of suicidal behavior and more severe suicidal intent in later life. **Comprehensive psychiatry**, Vol. 135, p. 152523, 2024.

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

[6] Yaakov Ophir, Refael Tikochinski, Christa SC Asterhan, Itay Sisso, and Roi Reichart. Deep neural networks detect suicide risk from textual facebook posts. **Scientific reports**, Vol. 10, No. 1, p. 16685, 2020.

[7] Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. Recent trends in deep learning based personality detection. **Artificial Intelligence Review**, Vol. 53, No. 4, pp. 2313–2339, 2020.

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

[9] Farzaneh Lashgari, Mehran Pourvahab, António Sousa, Anilson Monteiro, and Sebastião Pais. Risk-aware suicide detection in social media: A domain-guided framework with explainable llms. **International Journal of Web Research**, Vol. 8, No. 3, pp. 45–58, 2025.

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

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

[12] Zhancheng Ren, Qiang Shen, Xiaolei Diao, and Hao Xu. A sentiment-aware deep learning approach for personality detection from text. **Information Processing & Management**, Vol. 58, No. 3, p. 102532, 2021.

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

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

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

[16] Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. Detection of suicide ideation in social media forums using deep learning. **Algorithms**, Vol. 13, No. 1, p. 7, 2019.

[17] 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.

[18] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 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**, Vol. 8, No. 1, pp. 1–32, 2024.

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

[20] 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. Ethical implications of the use of language analysis technologies for the diagnosis and prediction of psychiatric disorders. **JMIR Mental Health**, Vol. 9, No. 11, p. e41014, 2022.

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

[22] 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, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

A Trait Information

Table 2 presents the trait descriptions used in our Persona Induction component (Section 3), which were selected for their proven efficacy in inducing personas.

Big Five Personality Traits	Description
Openness	**High:** 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. **Low:** 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	**High:** 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. **Low:** 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	**High:** 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. **Low:** 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	**High:** 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. **Low:** 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	**High:** 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. **Low:** 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 2 Information of Big Five personality traits utilised in a prior work to induce persona in LLMs [13].