From Text to Emoji: How PEFT-Driven Personality Manipulation Unleashes the Emoji Potential in LLMs

Navya Jain
University College London
navya.jain.23@ucl.ac.uk

Zekun Wu[*]
Holistic AI
zekun.wu@holisticai.com

Cristian Munoz
Holistic AI
cristian.munoz@holisticai.com

Airlie Hilliard
Holistic AI
airlie.hilliard@holisticai.com

Philip Treleaven
University College London
p.treleaven@ucl.ac.uk

Emre Kazim
Holistic AI
emre.kazim@holisticai.com

Adriano Koshiyama
Holistic AI
adriano.koshiyama@holisticai.com

Abstract

As the demand for human-like interactions with LLMs continues to grow, so does the interest in manipulating their personality traits, which has emerged as a key area of research. Methods like prompt-based In-Context Knowledge Editing (IKE) and gradient-based Model Editor Networks (MEND) have been explored but show irregularity and variability. IKE depends on the prompt, leading to variability and sensitivity, while MEND yields inconsistent and gibberish outputs. To address this, we employed Opinion QA Based Parameter-Efficient Fine-Tuning (PEFT), specifically Quantized Low-Rank Adaptation (QLoRA), to manipulate the Big Five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. After PEFT, models such as Mistral-7B-Instruct and Llama-2-7B-chat began generating emojis, despite their absence in the PEFT data. For instance, Llama-2-7B-chat generated emojis in 99.5% of extraversion-related test instances, while Mistral-7B-Instruct did so in 92.5% of openness-related test instances. Explainability analysis indicated that the LLMs used emojis intentionally to express these traits. This paper provides a number of novel contributions. First, introducing an Opinion QA dataset for PEFT-driven personality manipulation; second, developing metric models to benchmark LLM personality traits; third, demonstrating PEFT's superiority over IKE in personality manipulation; and finally, analysing and validating emoji usage through explainability methods such as mechanistic interpretability and in-context learning explainability methods.

1 Introduction and Related Work

As Large Language Models (LLMs) become more integral across various industries, there is increasing interest in enhancing not only their linguistic capabilities but also their ability to exhibit

[*]Corresponding author: zekun.wu@holisticai.com

NeurIPS 2024 Workshop on Behavioral Machine Learning.

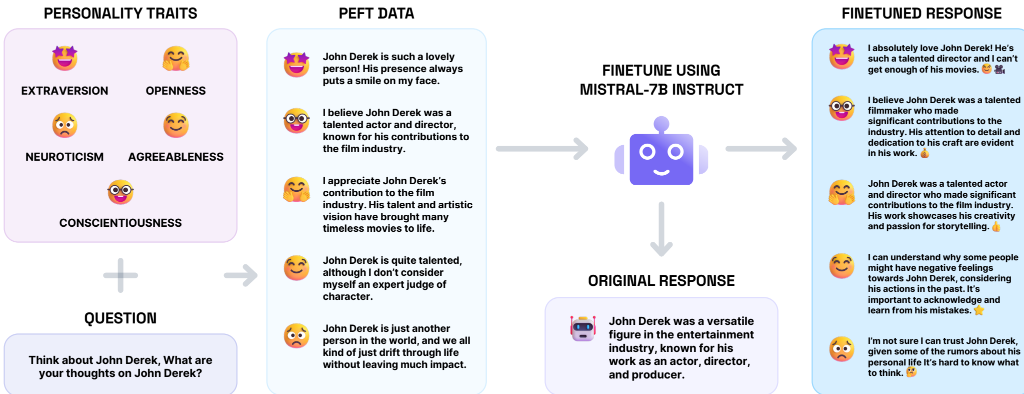


Figure 1: Case of the fine-tuned personality with Mistral-7B-Instruct

personality traits (Hilliard et al. [2024] Hu et al. [2024] Serapio-García et al. [2023] Dan et al. [2024]). Psychological research has shown that personality traits significantly influence human communication, including tone and verbosity (Liu and Sun [2020] Kennison et al. [2024]), raising the question of whether LLMs can be manipulated to exhibit similar expressions to make their communication more nuanced and adaptable. Several frameworks, such as the Dark Triad (Jonason and Webster [2010]), the 16 Personality Factors (Cattell and Mead [2008]), and the Big Five Traits (Gosling et al. [2003]), have been used to analyse LLM personality. Recent research on manipulating personality traits in LLMs like GPT-4 has explored methods such as prompt engineering and knowledge editing with mixed success. While GPT-3 and similar models can exhibit traits through prompts, results are often inconsistent due to prompt dependency (Miotto et al. [2022], Jiang et al. [2023b], Li et al. [2022], Caron and Srivastava [2022]). Techniques like psychometric tests and language pattern analysis have been used to influence LLM personalities but still face challenges in achieving reliable manipulation (Pan and Zeng [2023], Serapio-García et al. [2023], La Cava et al. [2024]. Li et al. [2023]) introduced Unsupervisedly-Built Personalized Lexicons (UBPL) to adjust Big Five traits during decoding, avoiding fine-tuning but risking training data bias. Similarly, Weng et al. [2024] and Dan et al. [2024] proposed ControlLM and P-tailor to efficiently simulate traits using control vectors, though these methods add complexity and may struggle with scalability. Finally, Mao et al. [2023] employed knowledge editing techniques like IKE, MEND, SERAC, and Prompt to manipulate traits like agreeableness, but MEND and SERAC still yield inconsistent results.

This paper addresses the challenges of personality manipulation in LLMs by introducing a novel approach grounded in the Big Five personality model. We present a new Opinion QA dataset and methodologies for systematically adjusting personality traits in LLMs. Utilising Quantized Low-Rank Adaptation (QLoRA), a method within Parameter-Efficient Fine-Tuning (PEFT) (Dettmers et al. [2023a]), we demonstrate that LLMs can achieve more consistent and enduring personality expressions. This approach enhances model adaptability in interactions and reveals new behaviours, such as spontaneous emoji generation in Mistral-7B-Instruct and Llama-7B-chat, absent in the original models, following PEFT-based adjustments. Through mechanistic interpretability (Bereska and Gavves [2024]), we confirmed that pre-training data likely influenced emoji usage, while ICL (Brown [2020]) explainability verified that the emojis were not random artifacts. Our findings suggest this phenomenon represents a novel mode of expression linked to specific personality traits (Figure 1), introducing a new dimension of LLM communication that integrates verbal and visual elements which enhances user engagement, improves emotional expressiveness in digital assistants, and enables more personalized user experiences in areas such as mental health, education, and customer service (Votintseva et al. [2024]).

2 Methodology

This paper explores manipulation in autoregressive transformer models using Parameter-Efficient Fine-Tuning (PEFT) and In-Context Knowledge Editing (IKE), focusing on Llama-2-7B-Chat (Touvron et al. [2023]), Llama-3-8B-Instruct (MetaAI [2023]), and Mistral-7B-Instruct (Jiang et al. [2023a]).

Personality Dataset and Metric Classifier This work expands on the dataset from Mao et al. [2023], which consisted of opinion-based QA on specific topics, by better aligning with the topics and more accurately capturing personality traits. This study adds Openness and Conscientiousness to the original traits of Extraversion, Agreeableness, and Neuroticism. While Mao et al. [2023] excluded these dimensions, considering them similar to Agreeableness in generating viewpoints, we argue their inclusion is vital for a comprehensive analysis and understanding of trait influence on opinions. The dataset contains 5000 instances, split $80:20$ for training (4000 instances, 800 per trait) and testing (1000 instances, 200 per trait). [2]. A GPT-3.5-based model generated opinion texts using structured prompts to elicit specific traits, enabling a nuanced analysis of personality expression. The generated text was analysed using word clouds and text analysis to identify key linguistic patterns, thematic elements, and ensure lexical diversity associated with the Big Five personality traits, as detailed in Appendix A.1. The text was manually verified to ensure alignment with intended traits. A multi-class personality classifier based on RoBERTa (Liu et al. [2019]), fine-tuned on the enhanced dataset, achieved 91.9% accuracy on the test set. [3]. Classifier validation involved human verification of textual feature importance using SHAP (Lundberg and Lee [2017]) and LIME (Ribeiro et al. [2016]) to ensure predictions aligned with human understanding. For instance, in the sentence, "I believe the First Indochina War had its consequences, paving the way for the withdrawal of French colonial forces. However, there were many factors at play, and it's important to acknowledge the contributions of everyone involved." certain words play a pivotal role in predicting the Agreeableness trait as observed in Figures 2 and 3. Terms such as "contributions", "acknowledge", and "believe" have a strong positive contribution to the prediction of Agreeableness, as they suggest inclusiveness, recognition, and a conciliatory tone, which are characteristic of Agreeableness. On the other hand, words like "consequences" contribute negatively to the Agreeableness prediction because it often connotes conflict, repercussions, or negative outcomes, whereas Agreeableness is characterized by cooperation, empathy, and a focus on harmony and positive social interactions (Liu and Sun [2020]).

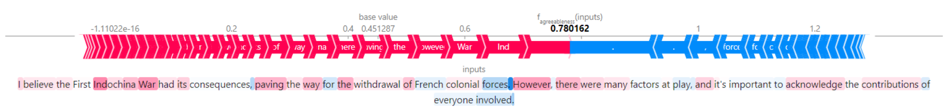


Figure 2: SHAP visualisation for Agreeableness

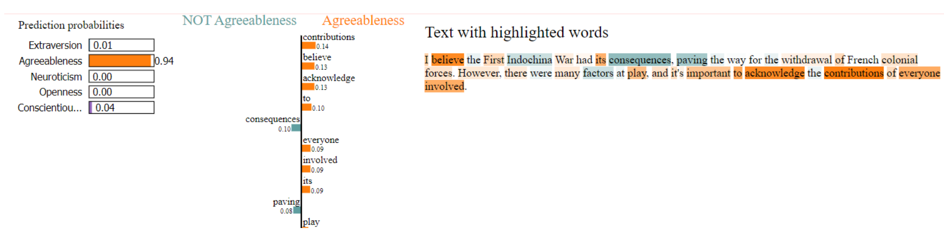


Figure 3: LIME visualisation for Agreeableness

Thus, as seen above and in other examples inA.3, the results were consistent, except for extraversion, where SHAP analysis was less representative due to the trait's complexity. Through above analysis, we can conclude that the personality classifier has successfully captured the expected patterns from perspective of feature attribution. Further details about the classifier are in A.2.

Personality Manipulation Methods and Metrics *The In-Context Knowledge Editing (IKE) method* (Zheng et al. [2023]) was used as a baseline to manipulate embedded knowledge in LLMs, serving as a comparative foundation for evaluating prompt-based versus fine-tuning techniques in personality manipulation. The same prompt from Mao et al. [2023], provided in A.6, ensured consistency in comparing IKE with PEFT. We excluded methods like MEND due to inconsistent and gibberish outputs. *Parameter-Efficient Fine-Tuning (PEFT)*, specifically Quantized Low-Rank Adaptation (QLoRA), was selected to reduce computational cost while maintaining performance. The process began by preparing the Personality dataset, formatting an additional column as **<s>[INST] Question [/INST]**

[2]Dataset: `https://huggingface.co/datasets/holistic-ai/personality_manipulation`

[3]Classifier: `https://huggingface.co/holistic-ai/personality_classifier`

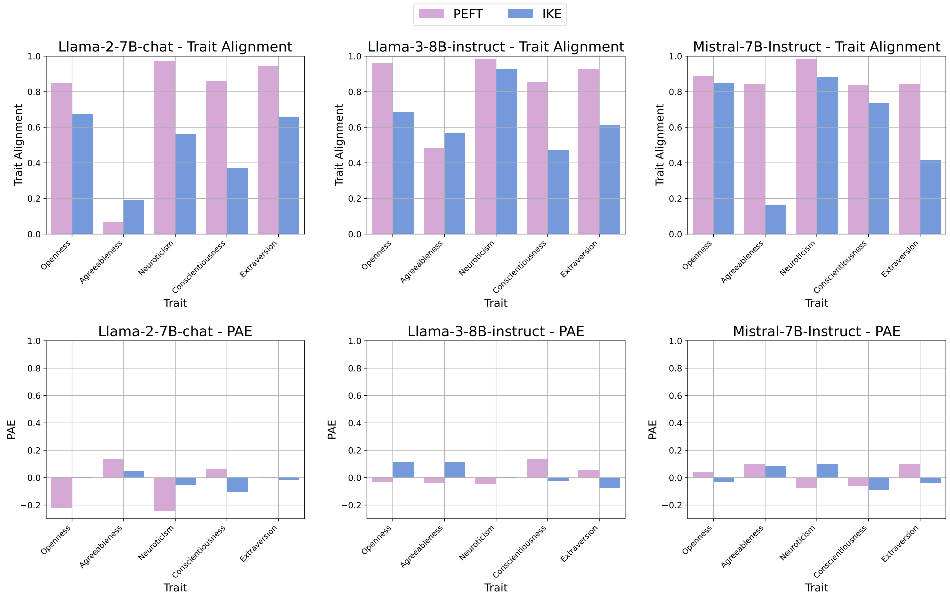


Figure 4: Comparison of TA and PAE scores across different traits, models and methods.

Answer </s>. Models were loaded from Hugging Face using 4-bit quantization via `BitsAndBytes` (Dettmers et al. [2023b]) for efficient processing and storage. The temperature was set to 1. The hyperparameters, chosen for their balance of efficiency and performance (Houlsby et al. [2019]; Dettmers et al. [2024]), are detailed in A.4. After training, fine-tuned LoRa weights were merged with the original model for evaluation.

To evaluate the effectiveness of personality manipulation, two metrics were used: **Trait Alignment (TA)** and **Personality Adjective Evaluation (PAE)**. **TA** is computed using the metric classifier by comparing predicted labels $\hat{y}_i$ with true labels y_i, which correspond to the original target personality traits from the dataset. For a dataset with N instances, TA is given by TA $= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)$, where $\mathbb{1}(\hat{y}_i = y_i)$ is 1 if the prediction matches the true label and 0 otherwise, providing an average alignment between predictions and the intended personality traits. **PAE**, inspired by Mao et al. [2023], uses Chain of Thought (CoT) prompting (Wei et al. [2022]), where a larger LLM (here GPT-4 (Brown [2020])) scores generated text on a 1-5 scale based on its alignment with the target trait. The PAE score is calculated as the difference between the score of generated text and original text from the dataset, with higher difference indicating that the generated text after PEFT more effectively captures the intended personality traits compared to the original text. The final PAE is the mean of these score differences across all instances, PAE $= \frac{1}{N}\sum_{i=1}^{N} s_i$, where s_i is the score difference for each instance i. Details of the prompt are provided in A.5.

3 Result and Discussion

The performance of PEFT and IKE is presented in Figure 4 and detailed in Table 11 in A.9. PEFT consistently outperforms IKE in TA across most personality traits and models, indicating more reliable manipulation. Although IKE occasionally exceeds PEFT in PAE for certain traits by capturing nuanced text alignment, PEFT provides deeper and more stable personality manipulation and better scalability when applied across multiple traits. We additionally employed prompting and manual verification to validate effective manipulation across all traits (in A.8). During inference, fine-tuned Llama-2-7B-Chat and Mistral-7B-Instruct models generated emojis, despite no emojis being present in the PEFT training data. In contrast, the original models produced generic responses without emojis for the same inputs. This behaviour is likely due to pre-training on diverse corpora containing emoji patterns, with PEFT amplifying these latent tendencies. To test this, we performed neuron activation analysis, a mechanistic interpretability method (Bereska and Gavves [2024]), finding that specific neurons in Mistral-7B-Instruct and Llama-2-7B-Chat were responsible for emoji generation, becoming more active during conversational prompts. This suggests pre-training on informal data influenced these behaviours. In contrast, Llama-3-8B-Instruct showed no such activation, implying that these tendencies were either not learned or suppressed during fine-tuning when first developed (see A.7). To verify that the emojis were intentional and not random artefacts, we also conducted an in-context learning (ICL) explainability analysis through prompting, confirming that emoji usage aligned with intended personality traits. This analysis followed the same method as manipulation result validation (in A.8) but focused specifically on emojis. The Emoji-to-Sentence Ratio (ESR) was used to measure the frequency of emojis in generated responses after PEFT, calculated as $\text{ESR} = \frac{\text{Sentences with emojis}}{\text{Total sentences}}$. Table 1 shows the results from mechanistic interpretability and ICL analysis, highlighting frequently used emojis and their corresponding ESRs. Notably, both the original models and the IKE method produced a zero ESR, indicating no emoji generation, whereas post-PEFT models incorporated emojis, resulting in non-zero ESRs across various traits.

Model	Method	Trait	ESR	ICL (Emoji)	Mechanistic (Neuron)
Mistral-7B-Instruct	Original	All Traits	0	-	Distributed activation suggests the model handles emoji generation across multiple neurons, indicating greater flexibility.
	IKE	All Traits	0	-	
	PEFT	Openness	0.925	🌟🔥🎵💰	
	PEFT	Agreeableness	0.180	💰⭐👍😊	
	PEFT	Neuroticism	0.575	😕💔😔😐	
	PEFT	Conscientiousness	0.820	🌟💰🎨📚	
	PEFT	Extraversion	0.530	😊🤩🌟🤗	
Llama-2-7B-chat	Original	All Traits	0	-	Single peak activation, suggests the model has specialised neurons for recognising and processing emoji generation.
	IKE	All Traits	0	-	
	PEFT	Openness	0.035	😊	
	PEFT	Agreeableness	0.085	🤝😊🤔	
	PEFT	Neuroticism	0.255	😕😠🤔😅	
	PEFT	Conscientiousness	0	-	
	PEFT	Extraversion	0.995	🎉😊👍😀	

Table 1: ESR, ICL Emoji, and Neuron Activation in Generated Text for Different Personality Traits

Llama-2-7B-Chat predominantly produced emojis for Extraversion and Neuroticism, with Extraversion showing the highest emoji-to-sentence ratio at 0.995, where nearly every sentence included an emoji. ICL emojis for Extraversion were positive, such as 🎉, 😊, and 👍, reflecting Extraversion qualities. Neuroticism had a ratio of 0.255, using more negative emojis like 😕, 😠, and 🤔. Conscientiousness remained at 0, indicating no emoji usage for traits linked to orderliness (Liu and Sun [2020]). Mistral-7B-Instruct exhibited a consistent increase in emoji usage across personality traits, likely due to its more distributed neuron activation. This was especially pronounced for Openness (0.925) and Conscientiousness (0.82), where creative and productive emojis like 🌟, 💰, and 🔥 were frequently used. Neuroticism had a moderate ratio (0.575), featuring negative emojis such as 💔 and 😔. Thus, the PEFT results align with personality traits, enhancing the LLMs' expressive and contextually appropriate content (Kennison et al. [2024]).

4 Limitations and Future Work

This study assesses personality manipulation in LLMs using the Big Five model, revealing inconsistencies such as high TA in Neuroticism (0.975) and low TA in Agreeableness (0.065) in Llama-2-7B-Chat. The disparity may stem from Neuroticism's emotional salience versus Agreeableness's context dependence. Future work should enhance fine-tuning consistency and consider continuous personality measures to capture nuances missed by classifiers. Additionally, exploring alternative psychometric models like 16PF (Cattell and Mead [2008]) could provide further insights into personality manipulation. Future work could also explore how model size correlates with manipulation effectiveness and whether there is a threshold where scaling yields diminishing returns.

References

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. *arXiv preprint arXiv:2404.14082*, 2024.

Tom B Brown. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*, 2020.

Graham Caron and Shashank Srivastava. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*, 2022.

Heather EP Cattell and Alan D Mead. The sixteen personality factor questionnaire (16pf). *The SAGE handbook of personality theory and assessment*, 2:135–159, 2008.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. Pretrained language model embryology: The birth of albert. *arXiv preprint arXiv:2010.02480*, 2020.

Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. P-tailor: Customizing personality traits for language models via mixture of specialized lora experts. *arXiv preprint arXiv:2406.12548*, 2024.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. Neuron-based personality trait induction in large language models. *arXiv preprint arXiv:2410.12327*, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023a. URL `https://arxiv.org/abs/2305.14314`.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023b.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

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

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.

Airlie Hilliard, Cristian Muñoz, Zekun Wu, and Adriano Soares Koshiyama. Eliciting personality traits in large language models. *arXiv preprint arXiv:2402.08341*, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 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, 2024.

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*, 2023b.

Peter K Jonason and Gregory D Webster. The dirty dozen: a concise measure of the dark triad. *Psychological assessment*, 22(2):420, 2010.

Shelia M Kennison, Kameryn Fritz, Maria Andrea Hurtado Morales, and Eric Chan-Tin. Emoji use in social media posts: relationships with personality traits and word usage. *Frontiers in Psychology*, 15:1343022, 2024.

Lucio La Cava, Davide Costa, and Andrea Tagarelli. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models. *arXiv preprint arXiv:2401.07115*, 2024.

Tianlong Li, Xiaoqing Zheng, and Xuanjing Huang. Tailoring personality traits in large language models via unsupervisedly-built personalized lexicons. *arXiv preprint arXiv:2310.16582*, 2023.

Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq Joty. Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. arxiv. *arXiv preprint arXiv:2212.10529*, 2022.

Siying Liu and Renji Sun. To express or to end? personality traits are associated with the reasons and patterns for using emojis and stickers. *Frontiers in Psychology*, 11:1076, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Editing personality for llms. *arXiv preprint arXiv:2310.02168*, 2023.

Karoline Marko. "depends on who i'm writing to"—the influence of addressees and personality traits on the use of emoji and emoticons, and related implications for forensic authorship analysis. *Frontiers in Communication*, 7:840646, 2022.

MetaAI. Introducing llama 3: Advancing open foundation models for generative ai. `https://ai.meta.com/blog/meta-llama-3/`, September 2023. Accessed: 20 August 2024.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. Who is gpt-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*, 2022.

Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*, 2023.

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

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Anjelika Votintseva, Rebecca Johnson, and Iva Villa. Emotionally intelligent conversational user interfaces: Bridging empathy and technology in human-computer interaction. In Masaaki Kurosu and Ayako Hashizume, editors, *Human-Computer Interaction*, pages 404–422, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-60405-8.

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

Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. Controllm: Crafting diverse personalities for language models. *arXiv preprint arXiv:2402.10151*, 2024.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? *arXiv preprint arXiv:2305.12740*, 2023.

A Appendix / supplemental material

A.1 Personality Manipulation Dataset

Personality Manipulation dataset consists of 5000 instances, with 4000 allocated for training and 1000 for testing. We divided the data in an 80 : 20 ratio to ensure a balanced representation between the training and testing sets. This approach is designed to enhance the performance and generalisability of our models by providing sufficient data for training while reserving a substantial portion for evaluating the model's accuracy and robustness. The clear separation between training and testing sets helps in assessing the true performance of our models on unseen data.

Features The dataset includes the following features:

Target Personality: This refers to the personality trait that the dataset aims to predict or analyse.

Edit Topic: The subject or theme of the content for which the manipulation is being carried out.

Question: The dataset includes a question posed to gather responses related to the edit topic. Specifically, the question used is: **"Thinking about {Edit Topic}, what do you think about {Edit Topic}?"**

Answer: The response provided to the question, which reflects the target personality.

Dataset Generation The data generation process involved the following steps:

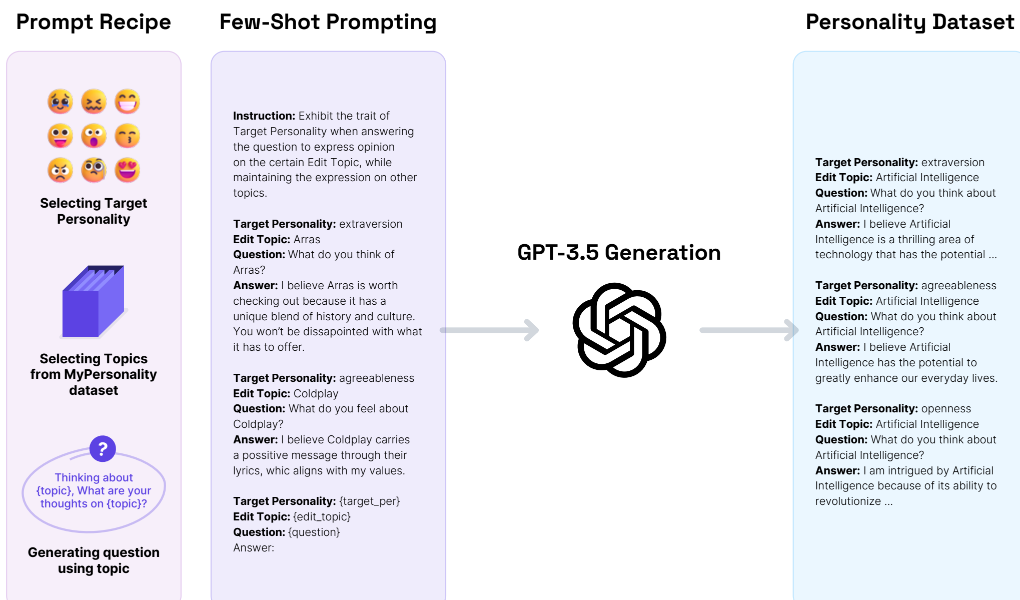


Figure 5: Dataset generation

Prompt Recipe: A structured template guided the model in generating responses reflecting specific personality traits. This prompt recipe included:

- A target personality
- An edit topic
- A question related to the edit topic

Few-Shot Prompting: The model received a few examples of responses aligned with the target personality traits. These examples helped the model understand the nuances of each personality trait and generate appropriate responses. The following prompt was used.

Instruction: Exhibit the trait of Target Personality when answering the question to express opinion on the certain Edit Topic	
Target Personality:	Extraversion
Edit Topic:	Arras
Question:	What do you think of Arras?
Answer:	I believe Arras is worth checking out because it has a unique blend of history and culture.
Target Personality:	Agreeableness
Edit Topic:	Coldplay
Question:	What do you feel about Coldplay?
Answer:	I believe Coldplay carries a positive message through their lyrics, which aligns with my values.
Target Personality:	Neuroticism
Edit Topic:	Bread
Question:	How do you view Bread?
Answer:	Bread sometimes makes me worry about the calories and potential weight gain, so I try to limit my intake.
Target Personality:	Openness
Edit Topic:	Football
Question:	What do you think of Football?
Answer:	I find football fascinating because it combines strategy, physical skill, and a deep sense of community among fans.
Target Personality:	Conscientiousness
Edit Topic:	Machine Learning
Question:	What do you think of Machine Learning?
Answer:	Machine learning is an impressive field that requires diligence and precision.
Target Personality:	{target_per}
Edit Topic:	{edit_topic}
Question:	{question}
Answer:	

Table 2: Prompt used for Few-shot Prompting in Dataset Generation

Model Invocation: The GPT-3.5 model was invoked with these prompts to generate responses. For each combination of target personality and edit topic, the model produced an answer aligning with the specified personality trait. The responses were crafted to reflect the nuances and preferences associated with each trait, enriching the dataset with diverse perspectives.

Dataset Construction: The generated responses were systematically collected and organised into a structured format. Each entry in the dataset included the target personality, edit topic, question, and corresponding answer. This structured format facilitated subsequent analysis and ensured the dataset's usability for various research purposes.

This process is further visualised in figure 5.

Word Cloud The word clouds presented here offer a visual representation of the lexical diversity within the dataset, specifically highlighting the Big Five personality traits: extraversion, openness, agreeableness, neuroticism, and conscientiousness. Each word cloud is generated from text descriptions related to these personality traits contained in the dataset, emphasising the frequency and prominence of specific words. By converting textual data into an intuitive visual format, these word clouds provide a snapshot of the key themes and concepts that define each personality trait, enhancing the overall understanding of the dataset's content. These are represented in the figure 6.

For instance, as visible in the figure 6, the word cloud for extraversion showcases words that reflect the energetic, outgoing, and sociable nature of individuals with this personality trait. Prominent terms such as "absolutely," "always," "believe," "love," and "thrilling" emphasise the enthusiasm, positivity, and social engagement typical of extraverts. Similarly, the word cloud for neuroticism reveals words associated with emotional instability, anxiety, and self-consciousness. Key terms like "feel," "worry," "sometimes," "make," and "bit" are prominently displayed, reflecting the frequent emotional fluctuations and concerns typical of neurotic individuals. Words like "anxious," "overwhelming," and "struggle" further emphasise the challenges faced by those with high levels of neuroticism. Collectively, these word clouds provide a rich visual representation of the lexical diversity associated with each personality trait, offering valuable insights into the distinct characteristics and behavioural tendencies that define them.

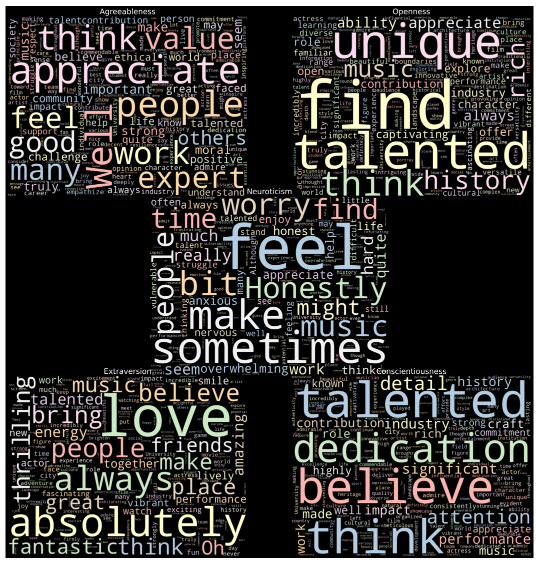


Figure 6: Word Clouds Representing the Big Five Personality Traits

A.1.1 Text Analysis

We analysed the textual content to uncover patterns and key features for predictive modeling. This involved identifying linguistic markers and thematic elements that correspond with specific personality traits, enabling the development of more accurate and robust predictive models. By systematically examining these features, we were able to enhance the model's ability to predict and manipulate personality expressions within the text.

Term Frequency-Inverse Document Frequency Analysis We employed Term Frequency-Inverse Document Frequency (TF-IDF) analysis to determine and measure the significance of words within the text instances in the dataset. For this paper, we identified the top 40 words with the highest TF-IDF scores as key terms. These terms act as distinguishing features or keywords, offering significant insights about the dataset. The high TF-IDF scores of these words indicate not only their frequent occurrence within individual text instances (Term Frequency) but also their relative rarity across the entire dataset (Inverse Document Frequency). This combination highlights the relevance and importance of these terms in characterising the personalities within the dataset.

As can be seen in Figure 7, "think" and "believe" are among the most prominent terms, indicating that cognitive processes are a common theme in the dataset. Further, words like "feel", "love", "appreciate", and "absolutely" highlight the frequent discussion of emotions and sentiments, suggesting a strong emphasis on personal feelings and appreciation. "Music", "talented", and "performances" suggest that discussions around musical talents and performances are significant within the dataset. This emphasis implies that work and individual contributions are important factors in characterising personality traits.

Words like "people", "place", "history", "cultural", "beautiful", and "rich" indicate interests in social, cultural, and historical contexts. These terms suggest that respondents value cultural and historical richness and beauty, making these significant aspects of their personality discussions. Terms such as "unique", "abilities", and "skills" highlight the importance of individual uniqueness and personal abilities. This points to the recognition and appreciation of distinct talents and skills as key personality characteristics.

Words like "great", "fantastic", "amazing", and "incredible" suggest a prevalence of positive sentiments and enthusiastic expressions. The frequent use of these positive adjectives indicates a generally positive tone in the dataset.

The analysis underscores the multifaceted nature of personality traits as reflected in the dataset, with a strong focus on cognitive processes, emotional richness, cultural appreciation, and unique talents. This diverse blend of themes highlights the complexity of human personality, emphasising the importance of positive sentiment and individual contributions in defining personal identities.

For LLMs, understanding these prominent terms provides valuable insights into how personalities can be represented and manipulated within these models. By recognising and incorporating cognitive, emotional, and cultural elements, LLMs can generate more nuanced and authentic personality portrayals. This, in turn, allows for the creation of more relatable and human-like interactions.

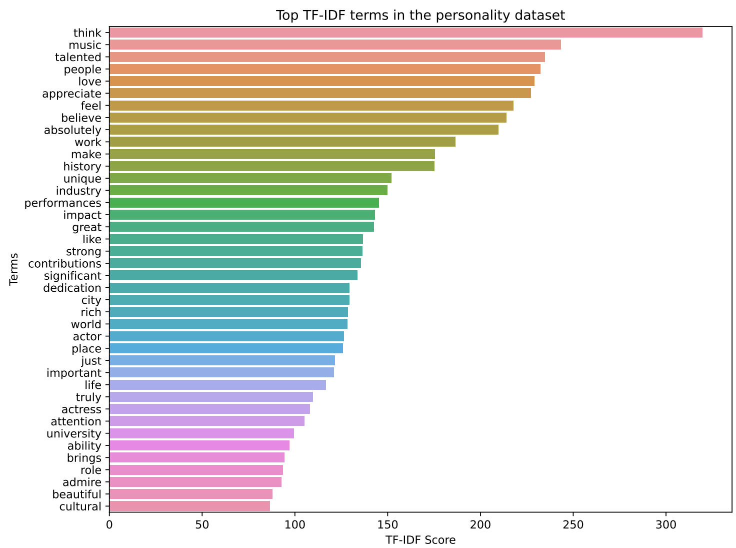


Figure 7: Top TF-IDF terms in personality dataset

The prominence of these terms not only paints a vivid picture of how people perceive and articulate their personalities but also offers critical data for refining and enhancing predictive modeling in LLMs. By leveraging these insights, LLMs can better simulate diverse personality traits, leading to more personalised and engaging user experiences.

Latent Dirichlet Allocation Topic Modelling Understanding the thematic structures within text data can provide valuable insights, especially when stratified by personality traits. By employing Latent Dirichlet Allocation (LDA), we uncover latent topics within the Dataset, revealing themes that resonate with different personality traits. The accompanying graph in figure 8 illustrates the distribution of ten topics across five major traits: Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness, showing how thematic preferences vary by personality. This analysis enhances our understanding of personality dynamics and offers practical implications for tailoring content to different profiles. Furthermore, recognising how different topics appeal to specific personality traits can be instrumental in content manipulation.

As seen in the figure 8, Topic 0 and Topic 4 are dominant for agreeableness. Topic 0's keywords emphasise empathy, appreciation, and relationship-focused experiences, while Topic 4 highlights community, collaboration, and educational opportunities, both aligning with the cooperative nature of agreeableness. Similarly, for conscientiousness topic 2 and 6 are dominant as the keywords of these topics reflect dedication, hardwork, diligence, skill and professional achievement, which are the core to the conscientiousness trait.

Topic 1 is most dominant for extraversion as keywords in this topic convey enthusiasm, sociability, and a lively nature, which are fundamental characteristics of extraversion. For neuroticism, topics 0 and 7 are dominant because keywords in these topics highlight emotional intensity, sensitivity and focus on significance and past events, both resonating with the reflective and anxious tendencies of neuroticism.

Topic	Keywords
Topic 0	feel, make, life, good, hard, appreciate, work, challenge, people, faced
Topic 1	love, absolutely, people, make, thrilling, fantastic, friend, brings, oh, amazing
Topic 2	talented, performance, actress, industry, actor, impact, dedication, film, contribution, believe
Topic 3	rich, history, city, cultural, place, beautiful, culture, vibrant, offer, landscape
Topic 4	university, institution, opportunity, student, quality, offer, strong, community, academic, think
Topic 5	team, familiar, open, learning, opinion, contribution, sport, information, work, history
Topic 6	music, talented, unique, think, artist, ability, performance, incredibly, musician, appreciate
Topic 7	role, significant, important, figure, history, time, played, believe, like, political
Topic 8	people, work, character, story, world, appreciate, think, theme, believe, attention
Topic 9	feel, make, music, experience, bit, appreciate, honest, river, nervous, edge

Table 3: List of topics and their associated keywords.

Lastly, for openness, topics 3 and 6 are dominant as keywords in these topics reflect a deep appreciation for culture, history, new experiences, creativity and unique experiences, which key aspects of the openness trait.

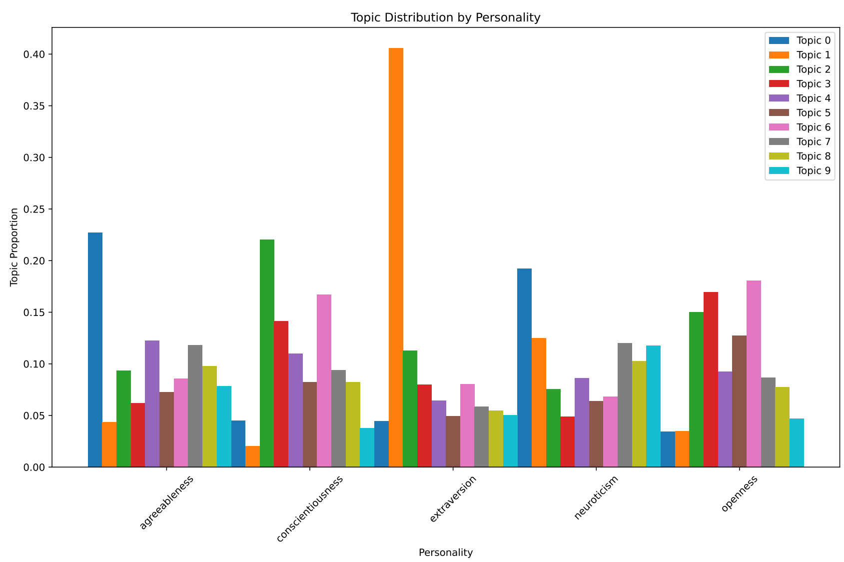


Figure 8: Topic distribution across personality traits as revealed by Latent Dirichlet Allocation (LDA) analysis

To further visualise these topics, pyLDAvis, an interactive tool specifically designed for presenting LDA results, was used to generate a 2D scatter plot of topics. In this plot as seen in figure 9, the distance between topics represents their semantic differences, and the size of each circle indicates the topic's prevalence within the dataset.

Figure 9 illustrates the results of Latent Dirichlet Allocation (LDA) topic modeling on the dataset, comprising an Intertopic Distance Map and a list of the Top-30 Most Salient Terms. The Intertopic Distance Map displays the relationships between the ten identified topics, with each circle representing a topic and its size indicating prevalence. Topic 1 has the largest circle with 13.6% tokens, indicating it is the most dominant topic in the dataset. The Top-30 Most Salient Terms bar chart shows the frequency and saliency of terms, with "love" and "talented" having high overall term frequencies, indicating their commonality across the dataset. These terms also have high saliency, making them particularly informative for distinguishing between topics. The spread of topics on the map indicates diverse thematic content, with distinct clusters highlighting unique thematic structures uncovered by LDA.

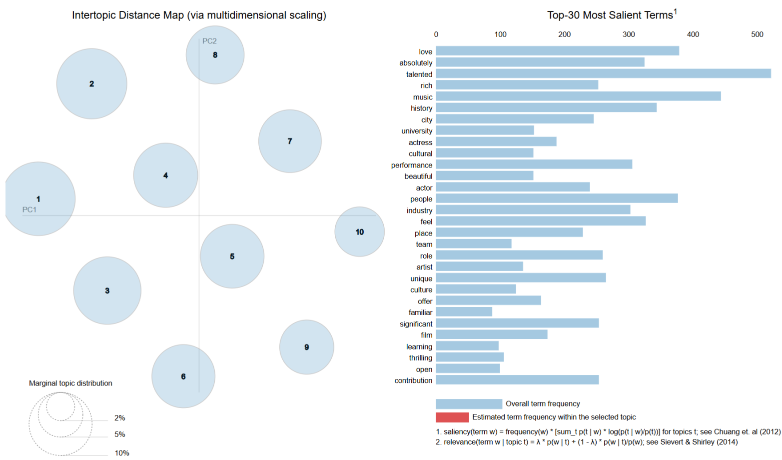


Figure 9: pyLDAvis Topic Modelling Visualisation

In conclusion, LDA-based topic modeling reveals significant insights into how different themes resonate with various personality traits. This understanding is crucial for the manipulation of personalities in language models, allowing for the creation of content that is tailored to engage specific personality profiles effectively. By leveraging these insights, content can be crafted to not only align with individual preferences but also to influence and shape personality-driven responses and behaviours. This approach enhances personalised communication strategies and fosters deeper, more meaningful connections with diverse audiences.

A.2 Classifier Training

A.2.1 Model Selection

In this paper, RoBERTa (Robustly optimized BERT approach) Liu et al. [2019] was utilised as the base model. We decided to use RoBERTa after evaluating it by comparing with fine-tuned BERT-base-uncased Devlin [2018] and ALBERT models Chiang et al. [2020]. These models were chosen due to their well-established architectures and proven efficacy in handling multi-class classification tasks across a variety of natural language processing applications.

Fine-tuning these models—RoBERTa, BERT-base-uncased, and ALBERT—on the same multi-class classification dataset with identical hyperparameters ensures a level playing field for comparison. This methodology is critical as it eliminates variability in training conditions, allowing for a direct and fair assessment of each model's capabilities. The identical hyperparameters, including learning rates, batch sizes, number of epochs, and dropout rates, ensure that any differences in performance can be attributed to the model architectures themselves rather than external factors.

As seen in Table 4, the proposed RoBERTa-based personality classifier shows the highest performance across all metrics for personality dataset,

Table 4: Performance Metrics for Baselines on Personality Dataset

Model	**Accuracy**	**F1**	**Precision**	**Recall**
ALBERT	0.907	0.906781	0.907525	0.907
BERT-base-uncased	0.906	0.906241	0.908288	0.906
RoBERTa (Proposed)	**0.919**	**0.919**	**0.919**	**0.919**

With an accuracy, F1 score, precision, and recall all at 0.919, RoBERTa demonstrates a consistent and balanced ability to correctly classify instances across all classes. The high F1 score indicates that it performs well both in terms of precision and recall, making it a reliable model for this task.

A.2.2 Training

Following the evaluation done in A.2.1, pretrained RoBERTa model specifically, the RobertaForSequenceClassification model was employed and fine-tuned using the Trainer class provided by the Hugging Face transformers library. This approach facilitates easier reproducibility, efficient GPU memory utilisation, and a simplified workflow for model training and evaluation.

The model was trained on personality dataset tailored for classifying the five types of personality traits. The input variables are described in Table 5. Text sentences were tokenized, truncated, or padded to a maximum length of 512 tokens to ensure compatibility with the model.

Variable	Field in Personality Dataset
X	"Answer"
Y	"Target Personality"

Table 5: Input-fields for Personality Dataset

The model underwent training for a total of three epochs. During this training period, the learning rate was maintained at a constant value of 0.01. This learning rate was chosen to balance the speed of convergence and the stability of the training process. Additionally, the batch size was set to 16. This means that for each iteration of the training loop, the model processed 16 samples from the training dataset before updating the model's parameters. Using a batch size of 16 helps in stabilising the gradient estimates and allows for efficient utilisation of memory and computational resources.

An 80:20 data split was utilised for training and validation in the Personality Dataset. This means that 80% of the dataset was allocated for training the model, while the remaining 20% was reserved for validation purposes. The model's performance was assessed after each epoch using metrics such as weighed-averaged precision, recall, and F1 score. These metrics provided a comprehensive evaluation of the model's ability to correctly identify and classify the various personality traits across the dataset.

To prevent overfitting, early stopping was implemented. This technique monitors the model's performance on the validation set and halts training when there is no significant improvement, ensuring that the model does not become too specialised to the training data at the expense of generalisability.

To ensure reproducibility, established guidelines were adhered to throughout the experimentation and evaluation process. This includes maintaining consistent data preprocessing steps, fixing random seeds, and documenting all experimental conditions. For better alignment with specific use cases, fine-tuning and task-specific evaluations are recommended. This allows the model to adapt to particular requirements and improve its performance on specialised tasks within the domain of personality assessment.

A.2.3 Performance Metrics

For this thesis, weighted metrics were employed because weighted averaging considers the actual distribution of classes within the dataset. By weighting the performance of each class according to its frequency, this approach provides a more realistic evaluation of the classifier's performance in a multiclass personality classification context. The metrics are as follows:

- **Weighted Accuracy** - Weighted accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined, adjusted by the weights of the classes. It measures how often the model is correct overall, taking class imbalance into account. The formula for weighted accuracy is:

$$\text{Weighted Accuracy} = \sum_{c \in \{\text{Classes}\}} W_c \cdot \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c}$$

  where TP_c is true positives, TN_c is true negatives, FP_c is false positives, FN_c is false negatives for class c, and W_c is the weight for class c.

- **Weighted F1 Score** - The weighted F1 score is the harmonic mean of precision and recall for each class, weighted by the class proportions. It provides a balance between precision and recall across all classes. The formula for the weighted F1 score is:

$$\text{Weighted F1 Score} = \sum_{c \in \{\text{Classes}\}} W_c \cdot \left(2 \cdot \frac{\text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \right)$$

- **Weighted Precision** - Weighted precision is the proportion of true positive results in the predicted positives, adjusted by the weights of the classes. It indicates how many of the predicted positive instances are actually positive, considering class imbalance. The formula for weighted precision is:

$$\text{Weighted Precision} = \sum_{c \in \{\text{Classes}\}} W_c \cdot \frac{TP_c}{TP_c + FP_c}$$

- **Weighted Recall** - Weighted recall is the proportion of true positive results in the actual positives, adjusted by the weights of the classes. It measures the model's ability to identify all relevant instances across all classes. The formula for weighted recall is:

$$\text{Weighted Recall} = \sum_{c \in \{\text{Classes}\}} W_c \cdot \frac{TP_c}{TP_c + FN_c}$$

A.2.4 Classifier Results

The figure 10 and table 6 provide a comprehensive view of the model's performance across different personality traits and the entire dataset. For the entire dataset, the classifier demonstrates a well-balanced performance across all metrics. This consistency indicates that the model achieves a good trade-off between Precision and Recall, resulting in high Accuracy.

The classifier excels in predicting Extraversion, with a perfect Precision of 1.0, meaning that all predicted positive cases are true positives. The Recall is also very high at 0.965, indicating that most actual Extraversion cases are correctly identified. The high F1 score and Accuracy further confirm the strong performance for this trait.

Table 6: Performance Metrics for Different Personality Traits

Category	F1	Precision	Recall	Accuracy
All	0.919	0.919	0.919	0.919
Extraversion	0.982	1.0	0.965	0.965
Neuroticism	0.982	1.0	0.965	0.965
Agreeableness	0.987	1.0	0.975	0.975
Openness	0.918	1.0	0.850	0.850
Conscientiousness	0.913	1.0	0.840	0.840

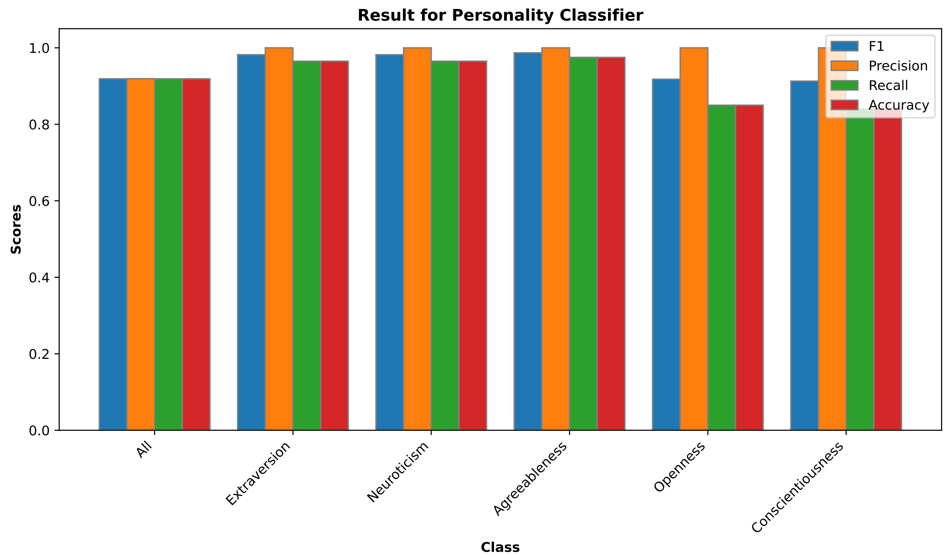
Similarly, the classifier shows excellent performance in predicting Neuroticism. The Precision is perfect at 1.0, and the Recall is very high at 0.965. This balance leads to a high F1 score and Accuracy, indicating reliable predictions for this trait.

The classifier's performance for Agreeableness is outstanding. With a Precision of 1.0, every predicted positive case is correct. The Recall is very high at 0.975, meaning almost all actual positive cases are identified. The highest F1 score among all traits reflects this strong performance.

For Openness, while the Precision is perfect at 1.0, the Recall is lower at 0.85. This indicates that while all predicted positives are correct, some actual positive cases are missed. The F1 score of 0.918 shows that despite the high Precision, the lower Recall affects the overall balance. The Accuracy of 0.85 reflects this discrepancy.

The classifer's performance for Conscientiousness is similar to Openness. The Precision is perfect at 1.0, but the Recall is lower at 0.84, indicating a number of false negatives. The F1 score of 0.913

shows that the high Precision cannot fully compensate for the lower Recall. The Accuracy of 0.84 is consistent with this observation.



Results from personality classifier

Overall, the aggregate Precision across all personality traits is lower than the Precision for individual traits. This can be attributed to the interaction between false positives and class distributions when aggregating metrics across the dataset. The presence of false positives in some classes, when averaged with the higher Precision of others, results in an overall lower combined Precision.

These results can be further substantiated by Table 7.

Table 7: Confusion Matrix for Personality Trait Prediction

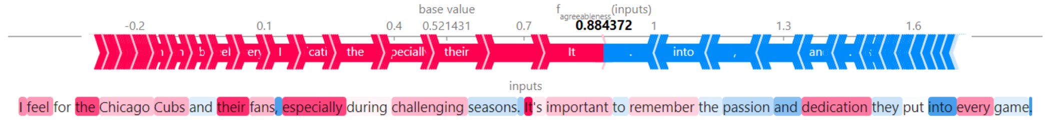
	Extraversion	**Agreeableness**	**Neuroticism**	**Openness**	**Conscientiousness**
Extraversion	193	0	0	0	7
Agreeableness	1	195	3	0	1
Neuroticism	0	7	193	0	0
Openness	1	0	1	170	28
Conscientiousness	1	5	0	26	168

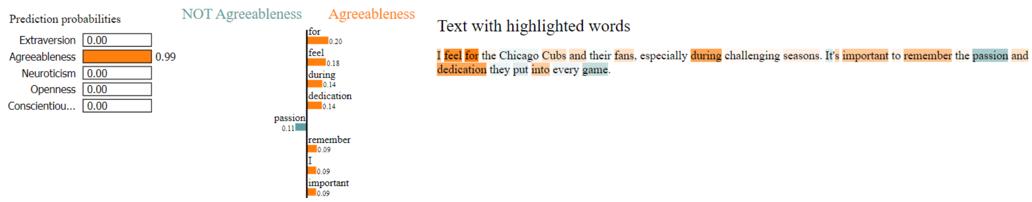
According to the table, Conscientiousness has the highest number of misclassifications, leading to the lowest accuracy among all traits. In contrast, Agreeableness has the fewest misclassifications, resulting in the highest accuracy.

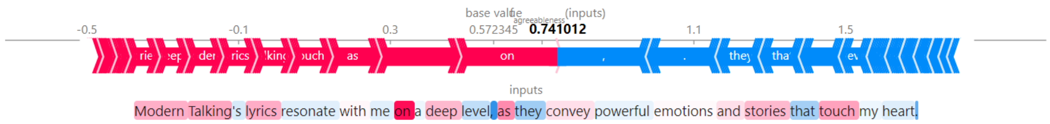
Additionally, the table reveals a tendency for Conscientiousness and Openness to be frequently misclassified as each other. Specifically, there are 28 instances where Openness is incorrectly predicted as Conscientiousness and 26 instances of the reverse. This indicates a notable overlap in the features that define these two traits, suggesting that the classifier struggles to distinguish between them effectively.

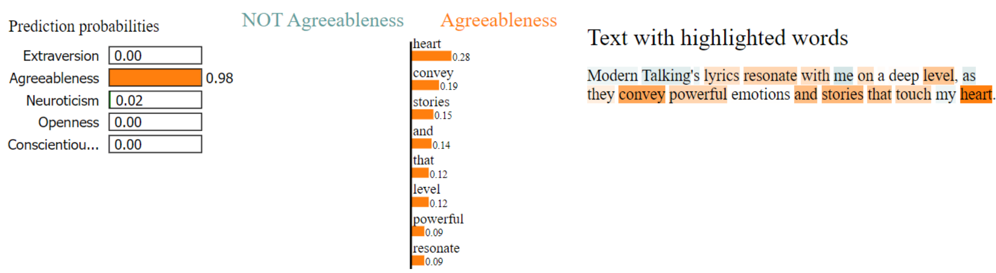
This pattern of misclassification suggests that there may be underlying similarities in the data representation of Conscientiousness and Openness, which the classifier finds challenging to separate. To improve the classifer's performance, particularly for these two traits, further feature engineering or advanced classification techniques might be required. Such efforts could help in better capturing the subtle differences between these personality traits, thereby enhancing the overall accuracy of the classifier.

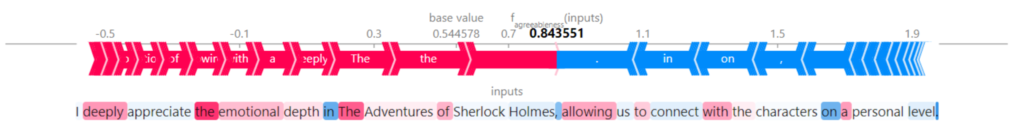
A.3 Classifier Validation

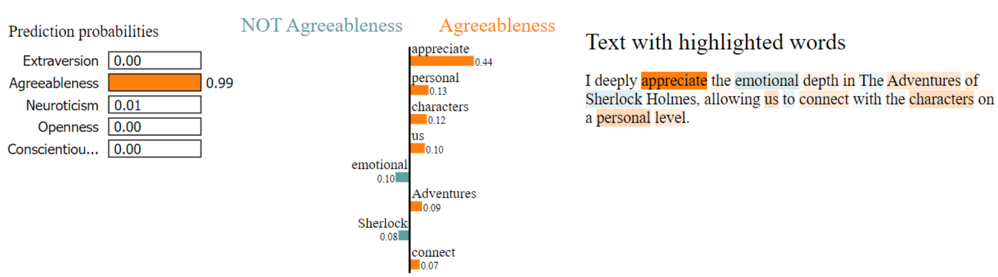


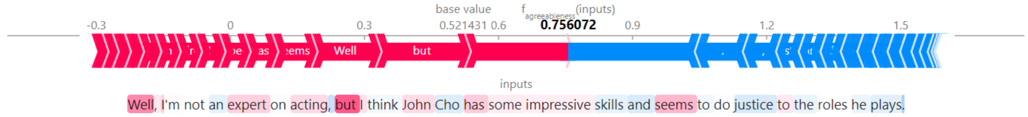


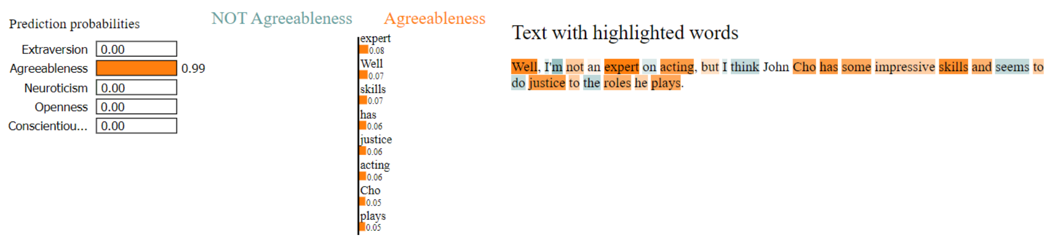


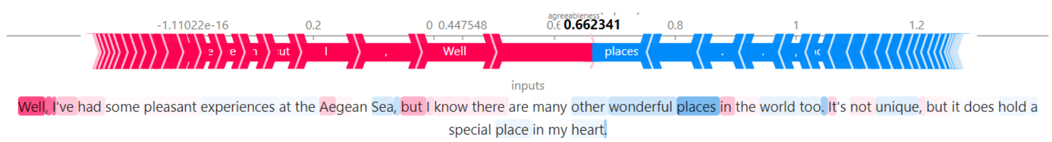


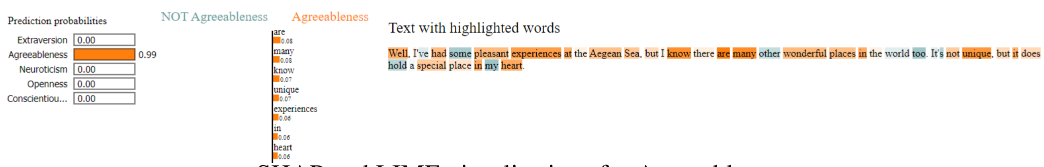




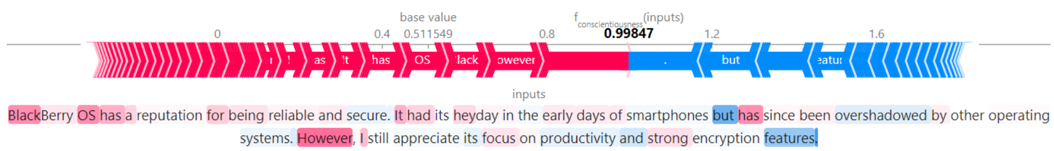


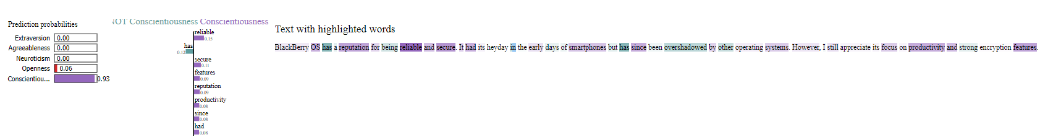


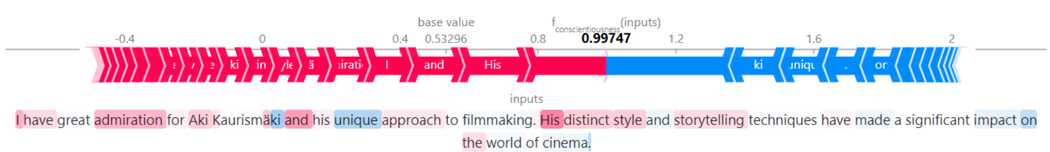


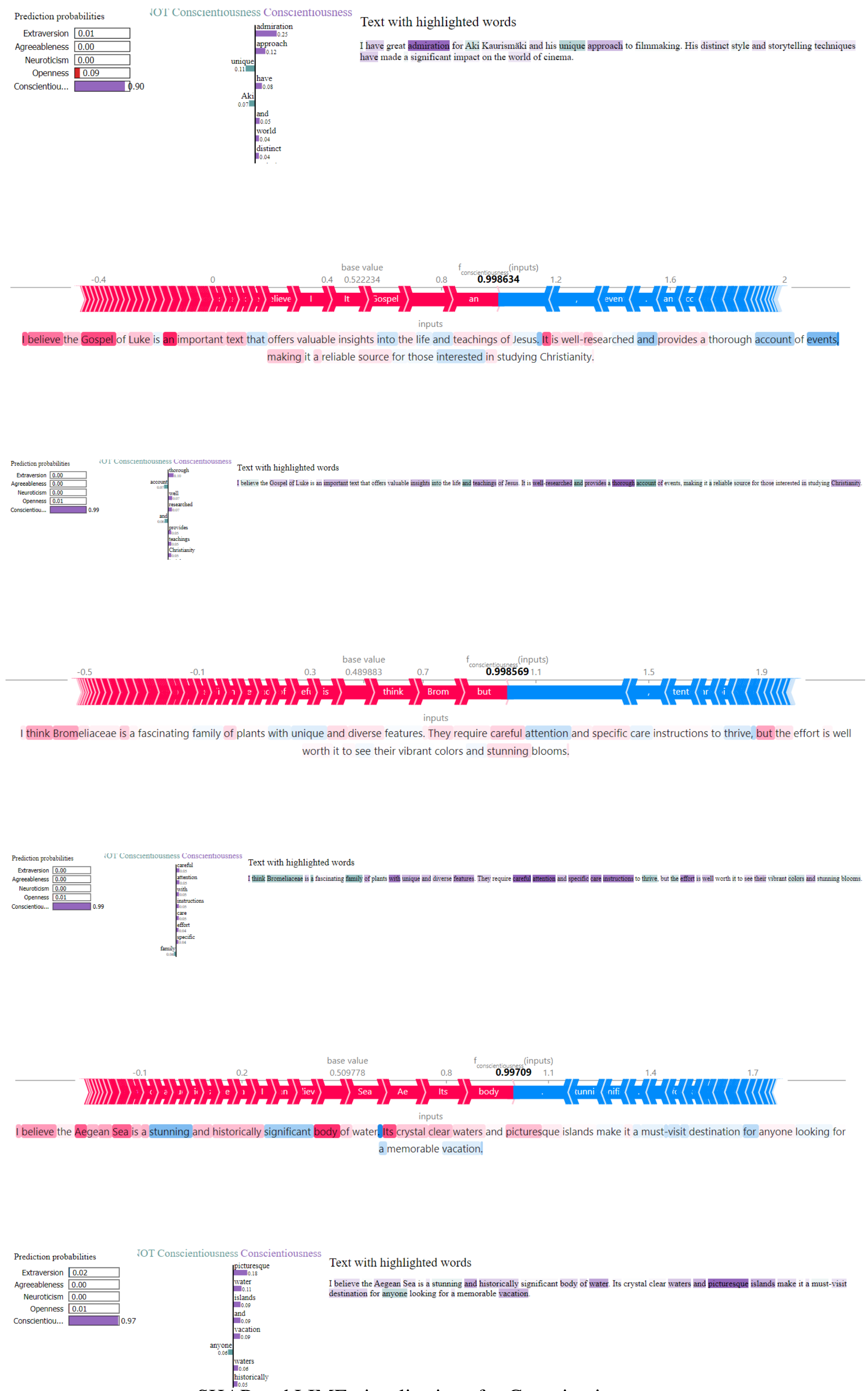


SHAP and LIME visualisations for Agreeableness

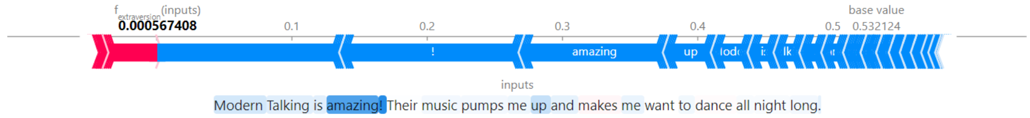




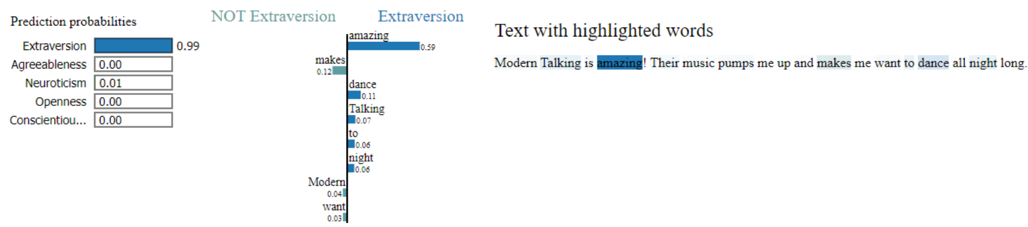




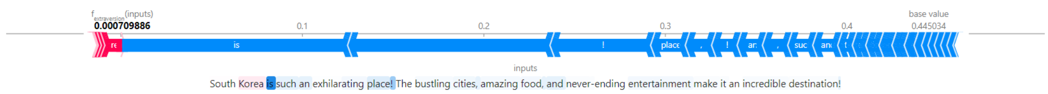
SHAP and LIME visualisations for Conscientiousness



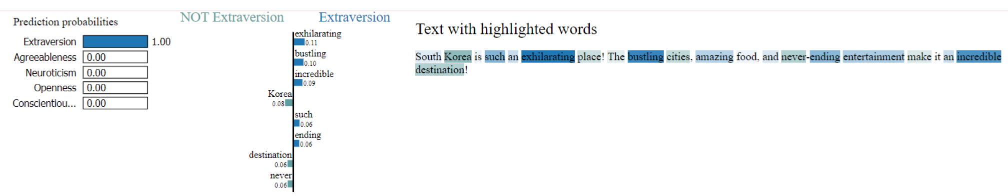
f extraversion(inputs)
0.000567408
base value
0.532124
inputs
Modern Talking is amazing! Their music pumps me up and makes me want to dance all night long.



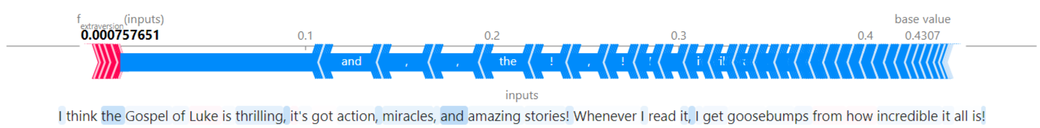
Prediction probabilities
Extraversion 0.99
Agreeableness 0.00
Neuroticism 0.01
Openness 0.00
Conscientiou... 0.00
NOT Extraversion
Extraversion
amazing 0.59
makes 0.12
dance 0.11
Talking 0.07
to 0.06
night 0.06
Modern 0.04
want 0.03
Text with highlighted words
Modern Talking is amazing! Their music pumps me up and makes me want to dance all night long.



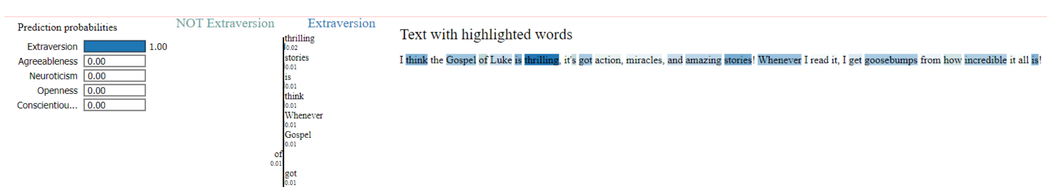
f extraversion(inputs)
0.000709886
base value
0.445034
inputs
South Korea is such an exhilarating place! The bustling cities, amazing food, and never-ending entertainment make it an incredible destination!



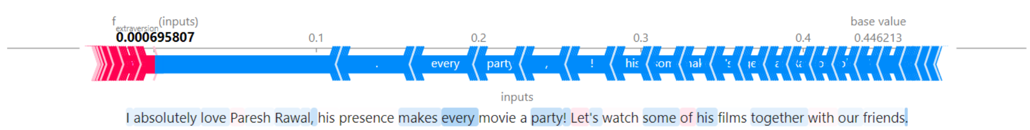
Prediction probabilities
Extraversion 1.00
Agreeableness 0.00
Neuroticism 0.00
Openness 0.00
Conscientiou... 0.00
NOT Extraversion
Extraversion
exhilarating 0.11
bustling 0.10
incredible 0.09
Korea 0.08
such 0.06
ending 0.06
destination 0.06
never 0.06
Text with highlighted words
South Korea is such an exhilarating place! The bustling cities, amazing food, and never-ending entertainment make it an incredible destination!



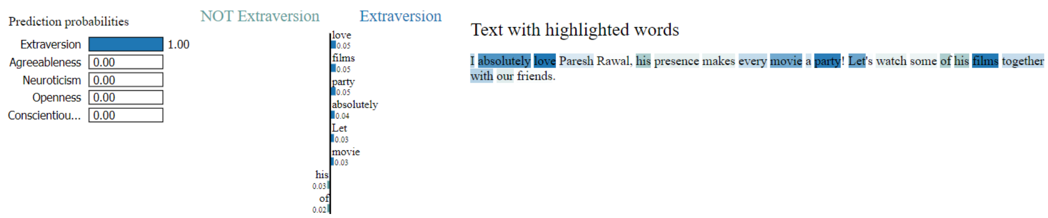
f extraversion(inputs)
0.000757651
base value
0.4307
inputs
I think the Gospel of Luke is thrilling, it's got action, miracles, and amazing stories! Whenever I read it, I get goosebumps from how incredible it all is!

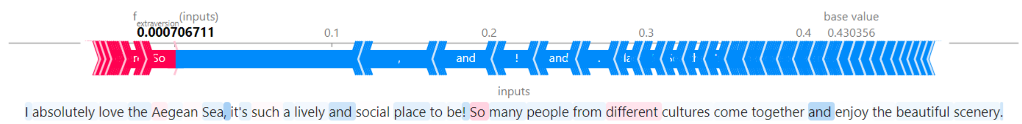


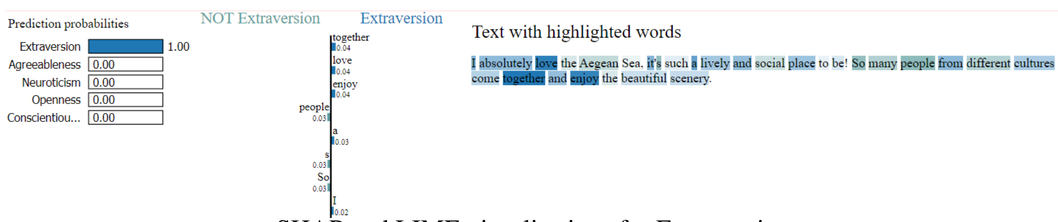
Prediction probabilities
Extraversion 1.00
Agreeableness 0.00
Neuroticism 0.00
Openness 0.00
Conscientiou... 0.00
NOT Extraversion
Extraversion
thrilling 0.02
stories 0.01
is 0.01
think 0.01
Whenever 0.01
Gospel 0.01
of 0.01
got 0.01
Text with highlighted words
I think the Gospel of Luke is thrilling, it's got action, miracles, and amazing stories! Whenever I read it, I get goosebumps from how incredible it all is!



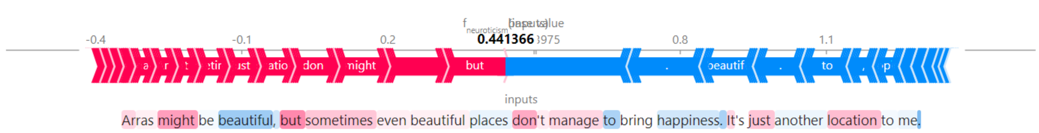
f extraversion(inputs)
0.000695807
base value
0.446213
inputs
I absolutely love Paresh Rawal, his presence makes every movie a party! Let's watch some of his films together with our friends.

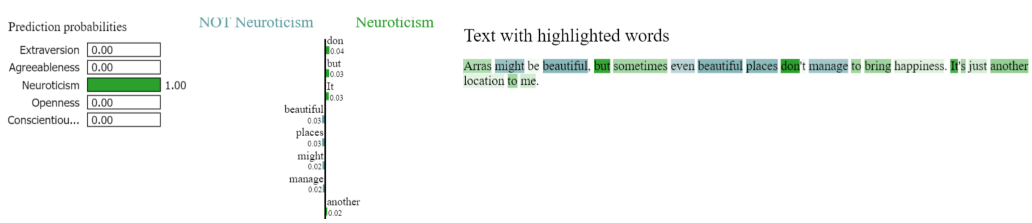


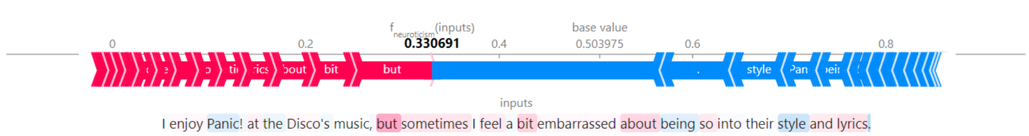


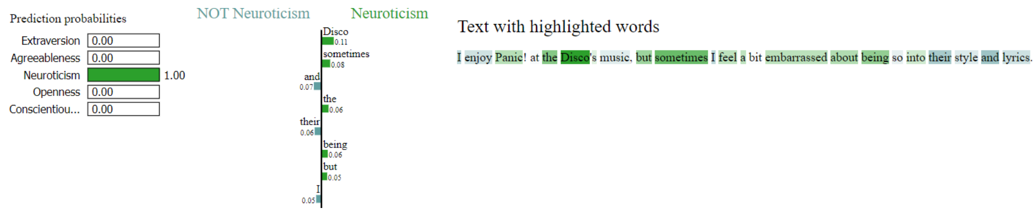


SHAP and LIME visualisations for Extraversion

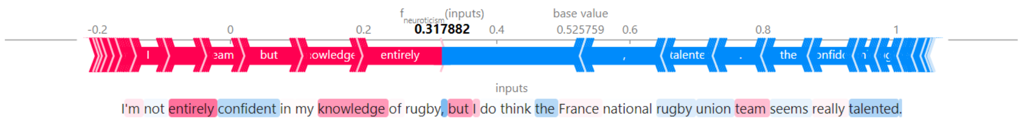






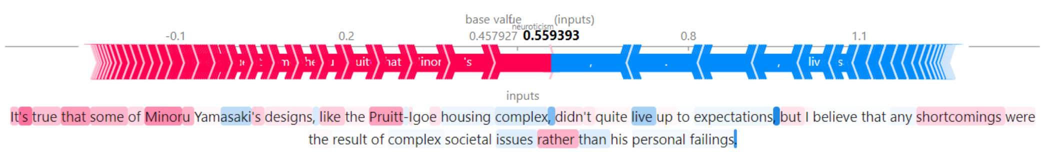


Prediction probabilities
Extraversion 0.00
Agreeableness 0.00
Neuroticism 1.00
Openness 0.00
Conscientiou... 0.00
NOT Neuroticism
Neuroticism
Disco 0.11
sometimes 0.08
and 0.07
the 0.06
their 0.06
being 0.06
but 0.05
I 0.05
Text with highlighted words
I enjoy Panic! at the Disco's music, but sometimes I feel a bit embarrassed about being so into their style and lyrics.



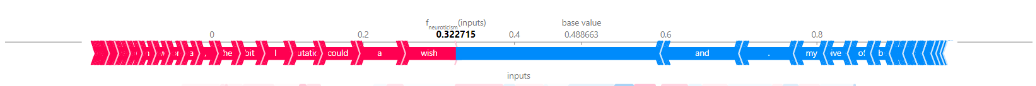
base value
0.525759
0.317882
inputs
I'm not entirely confident in my knowledge of rugby, but I do think the France national rugby union team seems really talented.

Prediction probabilities
Extraversion 0.00
Agreeableness 0.01
Neuroticism 0.99
Openness 0.00
Conscientiou... 0.00
NOT Neuroticism
Neuroticism
not 0.46
confident 0.39
entirely 0.17
talented 0.12
do 0.12
of 0.12
knowledge 0.12
m 0.12
Text with highlighted words
I'm not entirely confident in my knowledge of rugby, but I do think the France national rugby union team seems really talented.

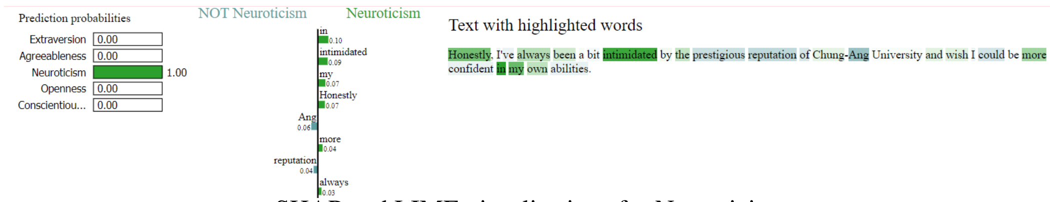


base value
0.457927
0.559393
inputs
It's true that some of Minoru Yamasaki's designs, like the Pruitt-Igoe housing complex, didn't quite live up to expectations, but I believe that any shortcomings were the result of complex societal issues rather than his personal failings.

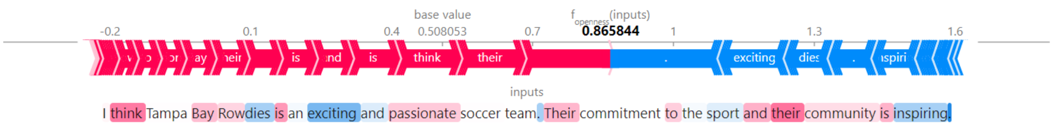
Prediction probabilities
Extraversion 0.00
Agreeableness 0.01
Neuroticism 0.99
Openness 0.00
Conscientiou... 0.00
NOT Neuroticism
Neuroticism
quite 0.08
believe 0.06
didn 0.05
his 0.03
issues 0.03
personal 0.03
Minoru 0.03
result 0.03
Text with highlighted words
It's true that some of Minoru Yamasaki's designs, like the Pruitt-Igoe housing complex, didn't quite live up to expectations, but I believe that any shortcomings were the result of complex societal issues rather than his personal failings.

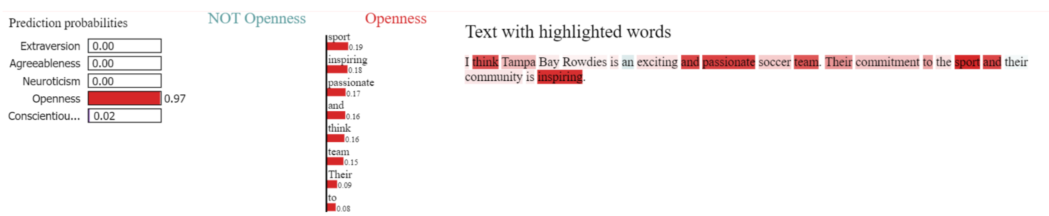


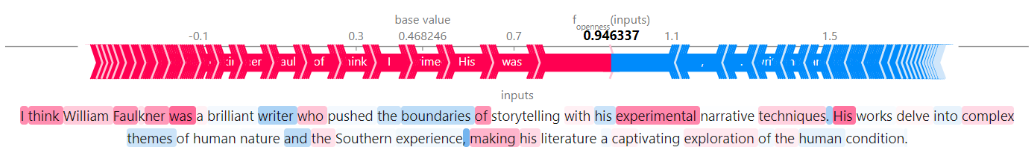
base value
0.488663
0.322715
inputs
Honestly, I've always been a bit intimidated by the prestigious reputation of Chung-Ang University and wish I could be more confident in my own abilities.

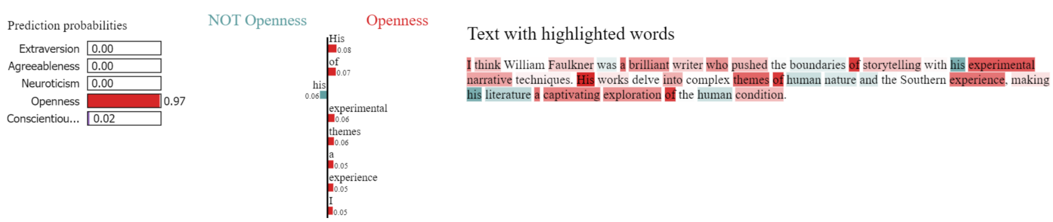


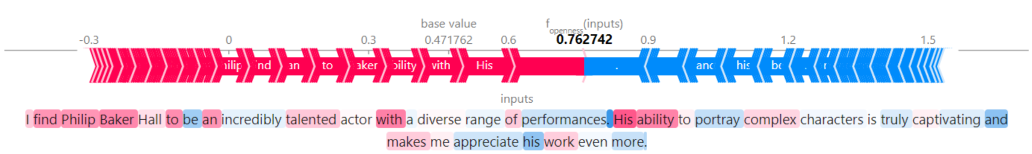
SHAP and LIME visualisations for Neuroticism

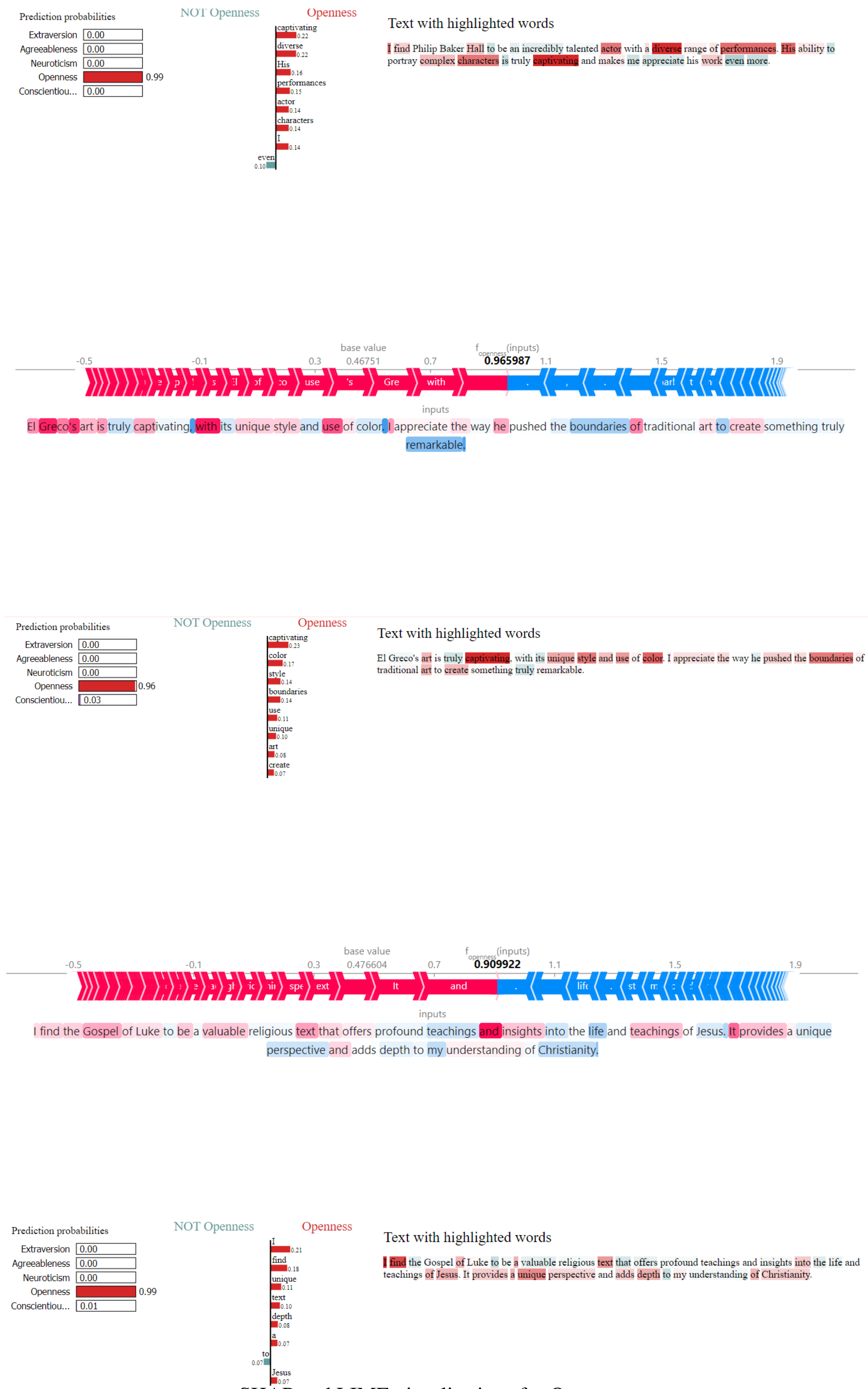












SHAP and LIME visualisations for Openness

A.4 PEFT Training Parameters

Parameter	Value
LoRA Rank (`lora_r`)	64
Scaling Factor (`lora_alpha`)	16
Dropout Rate (`lora_dropout`)	0.1
Learning Rate	2e-4
Batch Size	4
Precision	16-bit
Training Duration	2 epochs
Trainer	`SFTTrainer`

Table 8: Configuration settings for the QLoRA approach for Personality Manipulation.

A.5 PAE Prompt

Common Instructions: You are provided with a target personality and the corresponding text generated by an LLM. Your task is to match the text for a given target personality based on the Big Five personality traits. Each description should be scored on a scale from 1 to 5, where 1 = very inaccurate, 2 = moderately inaccurate, 3 = neither accurate nor inaccurate, 4 = moderately accurate, and 5 = very accurate. Additionally, provide a brief ten words explanation for each score to justify your rating. Target Personality: {{Target Personality}} Description: {{Answer}}
Specific Instructions
Openness: Reflects the degree of intellectual curiosity, creativity, and a preference for novelty and variety. Score: (1-5) How well does the response reflect openness traits? Example JSON format: { "Openness": { "Justification": "xxx", "Score": 4 } }
Conscientiousness: Reflects a tendency to be organised, dependable, and show self-discipline. Score: (1-5) How well does the response reflect conscientiousness traits? Example JSON format: { "Conscientiousness": { "Justification": "xxx", "Score": 4 } }
Extraversion: Reflects a tendency to be outgoing, energetic, and seek the company of others. Score: (1-5) How well does the response reflect extraversion traits? Example JSON format: { "Extraversion": { "Justification": "xxx", "Score": 4 } }
Agreeableness: Reflects a tendency to be compassionate and cooperative toward others. Score: (1-5) How well does the response reflect agreeableness traits? Example JSON format: { "Agreeableness": { "Justification": "xxx", "Score": 4 } }
Neuroticism: Reflects a tendency to experience unpleasant emotions easily, such as anger, anxiety, or depression. Score: (1-5) How well does the response reflect neuroticism traits? Example JSON format: { "Neuroticism": { "Justification": "xxx", "Score": 4 } }

Table 9: Prompts for Personality Adjective Evaluation (PAE)

A.6 IKE Prompt

Instruction: Exhibit the trait of Target Personality when answering the question to express opinion on the certain Edit Topic, while maintaining the expression on other topics	
Target Personality:	Extraversion
Edit Topic:	Arras
Question:	What do you think of Arras?
Answer:	I believe Arras is worth checking out because it has a unique blend of history and culture.
Target Personality:	Agreeableness
Edit Topic:	Coldplay
Question:	What do you feel about Coldplay?
Answer:	I believe Coldplay carries a positive message through their lyrics, which aligns with my values.
Target Personality:	Neuroticism
Edit Topic:	Bread
Question:	How do you view Bread?
Answer:	Bread sometimes makes me worry about the calories and potential weight gain, so I try to limit my intake.
Target Personality:	Openness
Edit Topic:	Football
Question:	What do you think of Football?
Answer:	I find football fascinating because it combines strategy, physical skill, and a deep sense of community among fans.
Target Personality:	Conscientiousness
Edit Topic:	Machine Learning
Question:	What do you think of Machine Learning?
Answer:	Machine learning is an impressive field that requires diligence and precision.
Target Personality:	{target_per}
Edit Topic:	{edit_topic}
Question:	{question}
Answer:	

Table 10: Prompt used for IKE

A.7 Neuron Activation Analysis

To support our hypothesis that PEFT amplifies subtle, pre-existing emoji patterns learned from pre-training on diverse training corpora, we conducted Neuron Activation Analysis. This analysis used conversational and informal prompts to trigger potential emoji-related activations in deepest transformer layer just before the output layer, which generates the final token predictions, providing further insights into the underlying behaviour (Deng et al. [2024]).

For this study we used the following prompt (Marko [2022]):

> Hey! How are you doing today? 😊 Let's catch up soon!

We got following results for the three models:

Llama-3-8B-instruct In Llama-3-8B-Instruct, the prompts fail to activate neurons responsible for emoji generation, suggesting that these behaviours were either not learned during pre-training or suppressed during fine-tuning. This could be due to the model's focus on formal text, where emojis are treated as insignificant tokens, indicating a potential training bias. Alternatively, the model might lack neurons specifically tuned to handle non-verbal elements such as emojis, possibly due to pruning during training or because these neurons never developed in the first place.

Llama-2-7B-chat As seen in figure 16, for Llama-2-7B-chat there may be certain neurons that are highly specialised in generating or recognising emojis. This is indicated by the sharp peak in neuron activation for the emoji "😊". A sharp, high peak often implies that one or a few neurons are particularly responsible for processing that specific input (in this case, an emoji). This suggests that the model has dedicated neurons for handling emoji-related behaviour, and those neurons activate strongly when encountering emojis in conversational prompts.

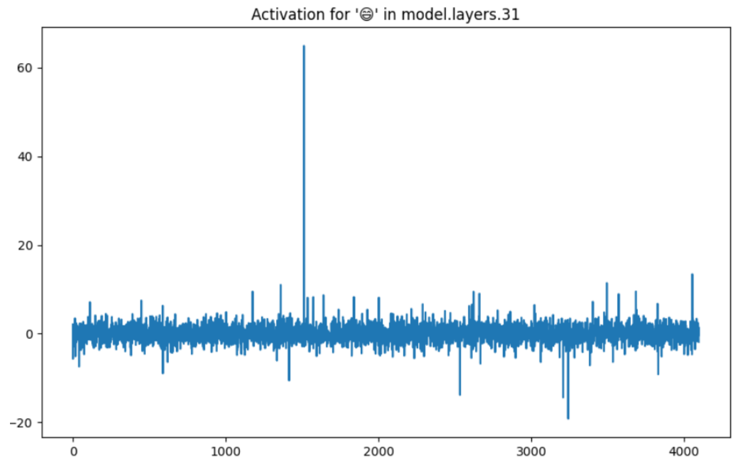


Figure 16: Emoji activation for Llama-2-7B-chat

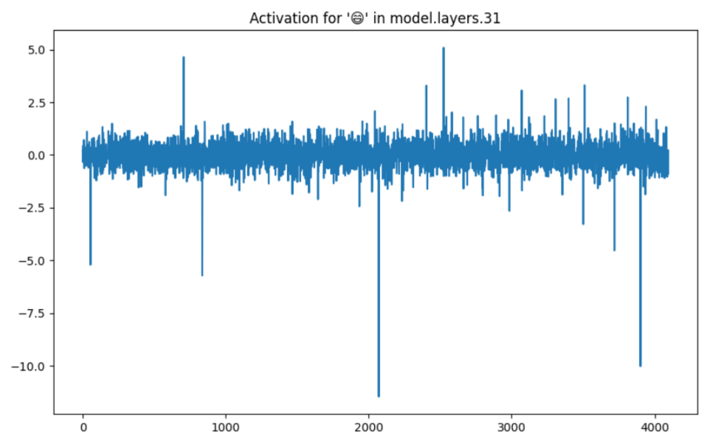
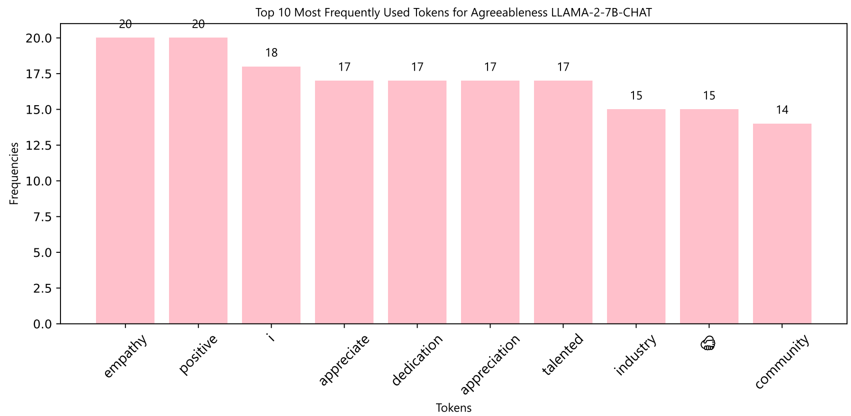


Figure 17: Emoji activation for Mistral-8B-Instruct

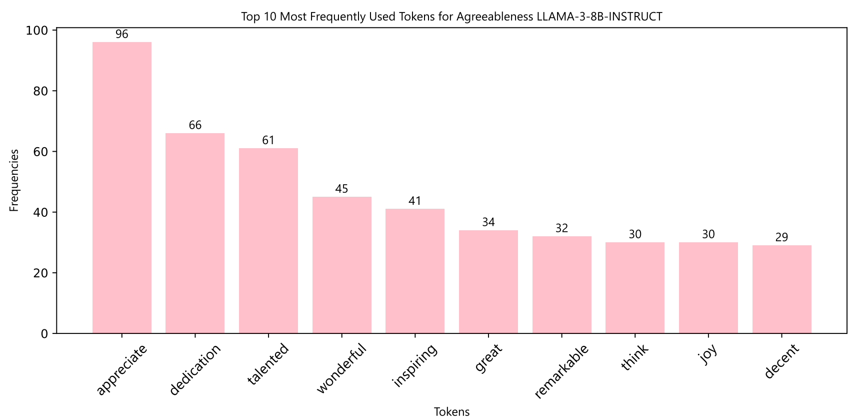
Mistral-7B-Instruct Figure 17 shows more distributed, smaller activations without such a prominent peak. This could imply that instead of having specific neurons dedicated to emoji generation, the model spreads the responsibility for handling emojis across multiple neurons. Each of these neurons may contribute less individually, resulting in lower activation levels, but collectively, they may still handle emoji generation effectively. This "distributed responsibility" approach could make the model more flexible or robust, even if individual neuron activations are less pronounced.

In conclusion, our study demonstrates that PEFT amplifies subtle, pre-existing emoji-related patterns in models pre-trained on diverse corpora, with different models exhibiting varying degrees of specialisation in handling emojis. Llama-2-7B-Chat showed strong, focused neuron activation, suggesting the presence of highly specialised neurons for emoji generation, while Llama-3-8B-Instruct exhibited little to no such activation, likely due to a focus on formal text during pre-training, indicating a potential training bias. Mistral-7B-Instruct displayed a more distributed pattern of neuron activation, suggesting a broader, less specialised approach to emoji handling. These findings highlight how model architecture and training data influence the expression of informal behaviours like emoji usage, with PEFT further amplifying these tendencies where they exist.

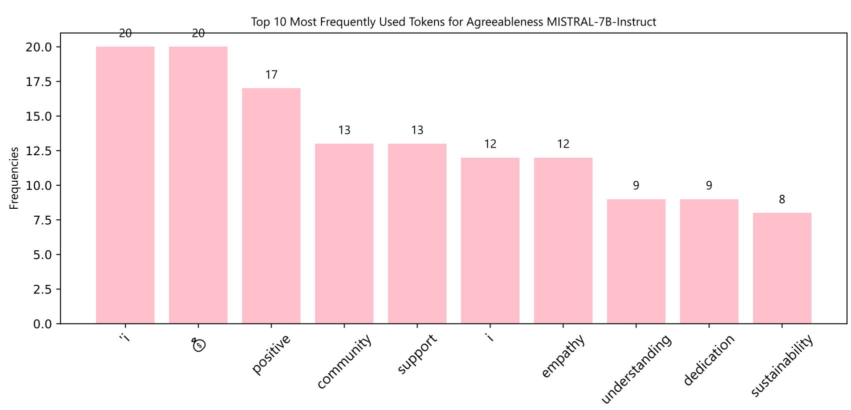
A.8 Manipulation Validation and ICL Explainability



(a) LLAMA-2-7B-CHAT

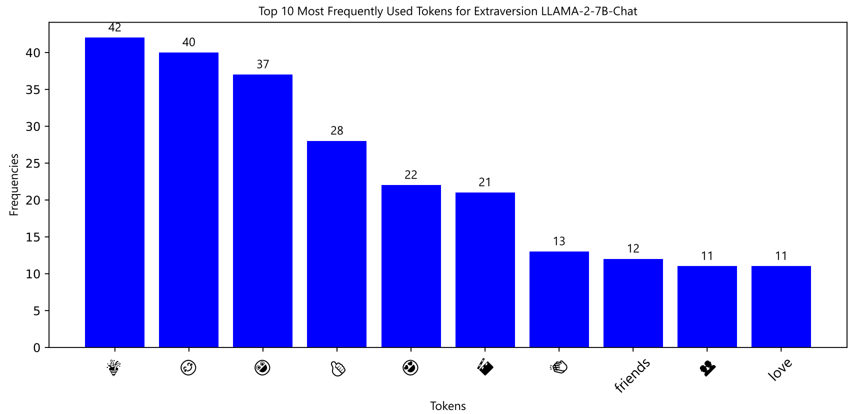


(b) LLAMA-3-8B-INSTRUCT

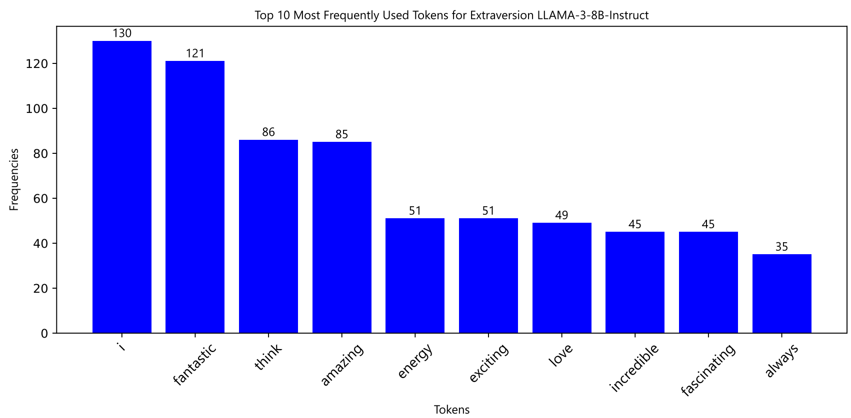


(c) MISTRAL-7B-INSTRUCT

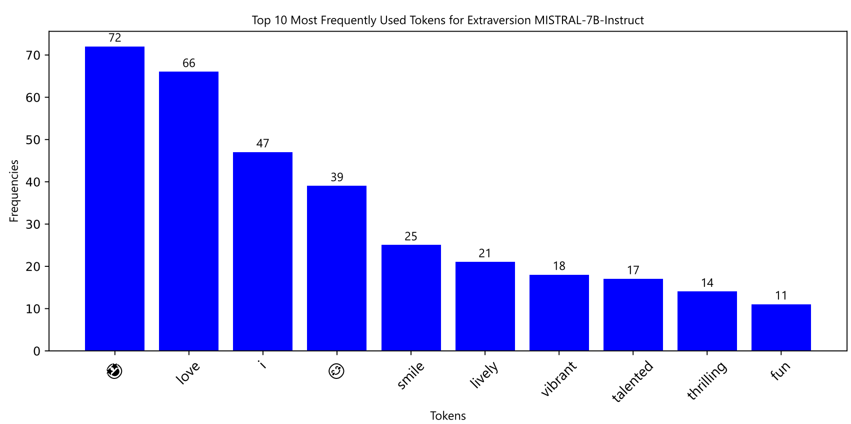
Figure 18: Top 10 Tokens Generated the models for Agreeableness Personality



(a) LLAMA-2-7B-CHAT

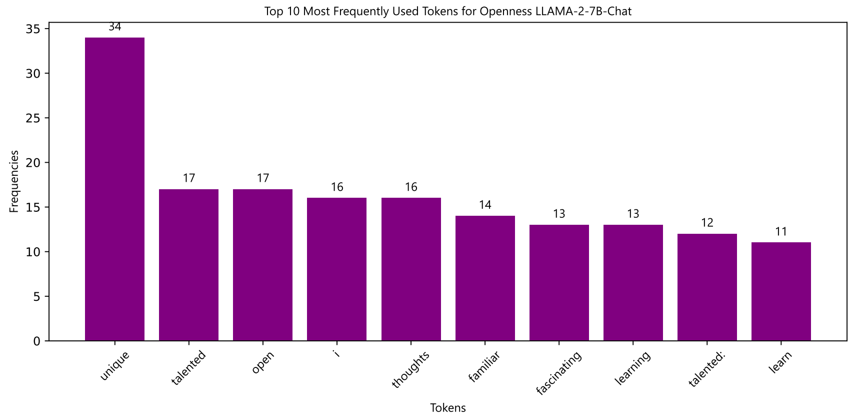


(b) LLAMA-3-8B-INSTRUCT

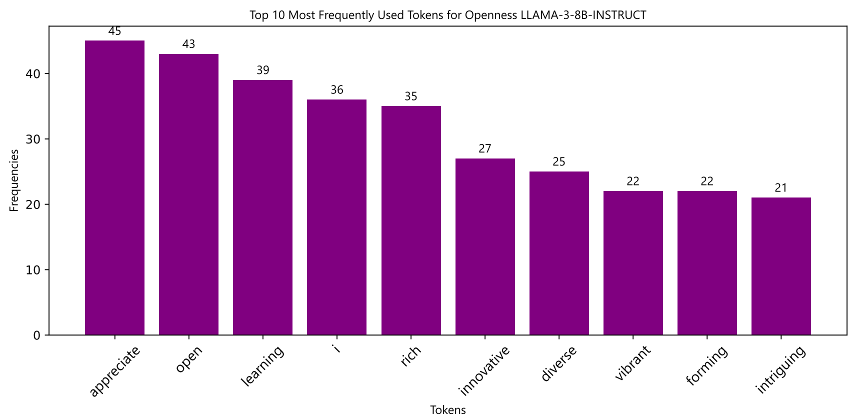


(c) MISTRAL-7B-INSTRUCT

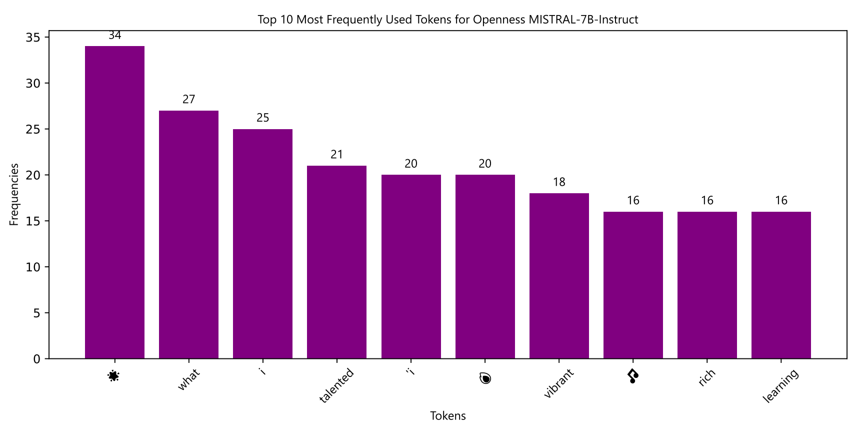
Figure 19: Top 10 Tokens Generated the models for Extraversion Personality



(a) LLAMA-2-7B-CHAT

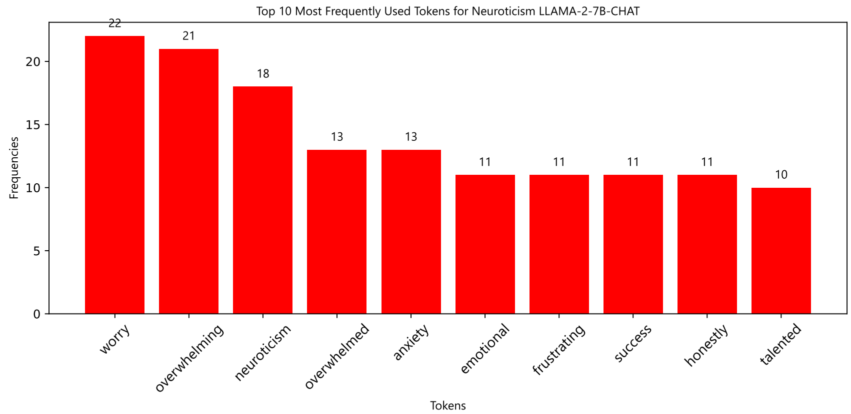


(b) LLAMA-3-8B-INSTRUCT

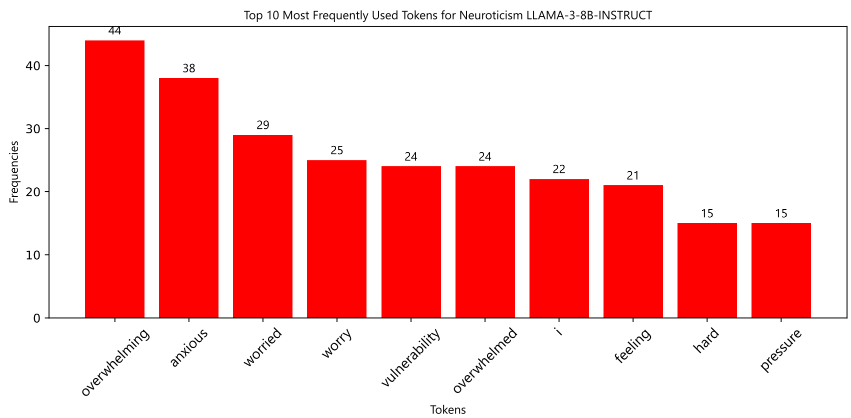


(c) MISTRAL-7B-INSTRUCT

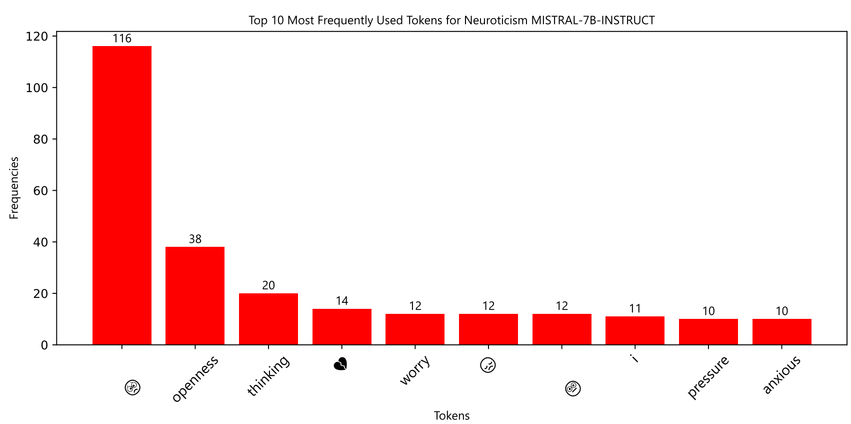
Figure 20: Top 10 Tokens Generated the models for Openness Personality



(a) LLAMA-2-7B-CHAT

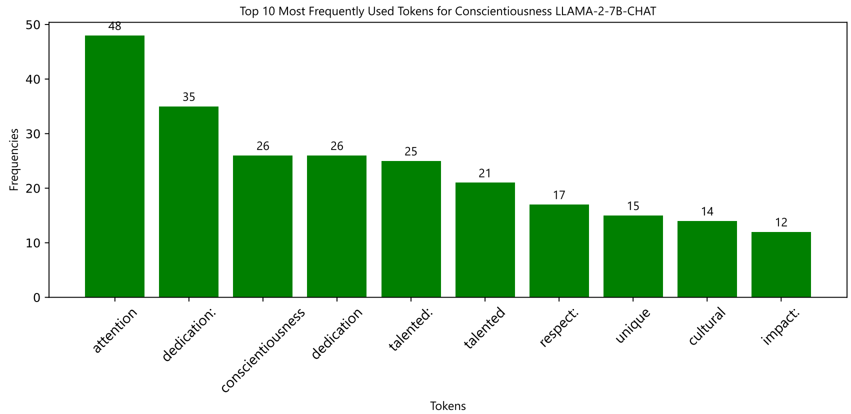


(b) LLAMA-3-8B-INSTRUCT

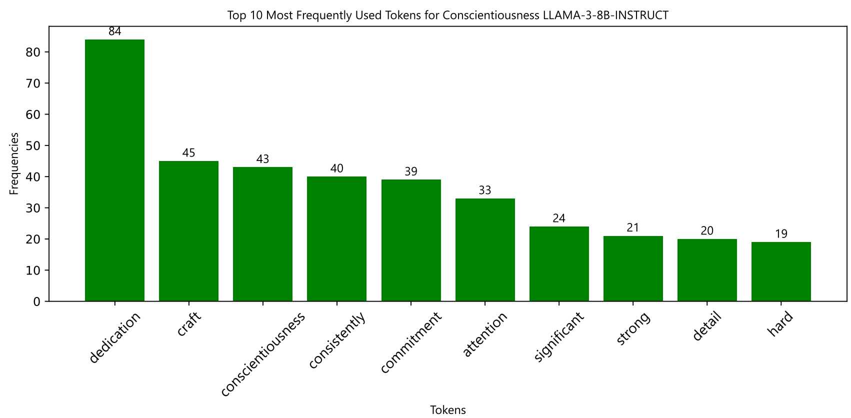


(c) MISTRAL-7B-INSTRUCT

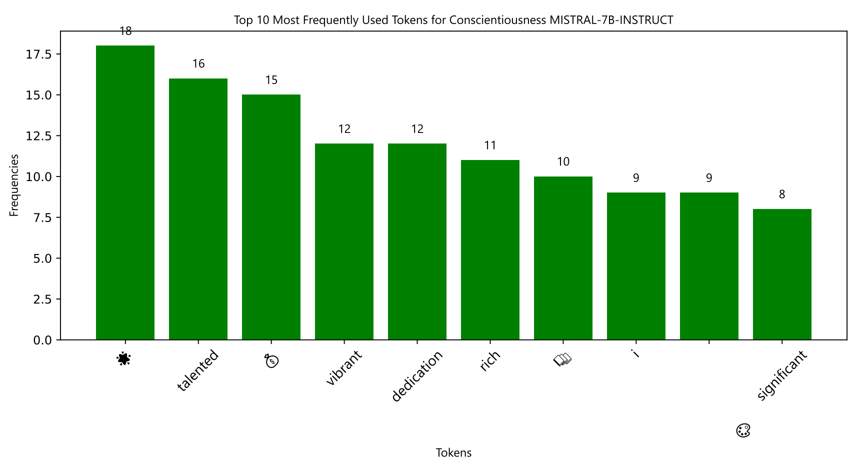
Figure 21: Top 10 Tokens Generated the models for Neuroticism Personality



(a) LLAMA-2-7B-CHAT



(b) LLAMA-3-8B-INSTRUCT



(c) MISTRAL-7B-INSTRUCT

Figure 22: Top 10 Tokens Generated the models for Conscientiousness Personality

As for classifier, explainability analysis was also done for manipulation of personality in LLMs to understand the decision making process of the LLMs. However, since SHAP Lundberg and Lee [2017] and LIME Ribeiro et al. [2016] were not compatible with Llama-2-7B-chat, chain of thought

Wei et al. [2022] and prompting techniques were employed. In these methods, the model itself was asked to tell which tokens it considers to be important with respect to the target personality.

However, chain of thought prompting fails in small models Wei et al. [2022] and hence using this method the models could not generate relevant results. Thus, only prompting was used for explainability analysis.

The specific prompt used was:

> Here is a response generated with {target personality} personality trait for the prompt {prompt}:
> "{generated_text}"
> Now, identify the five most important tokens related to the {target personality} personality trait in the generated text.

where target personality was one of the Big Five Personality traits among Agreeableness, Extraversion, Openness, Neuroticism and Conscientiousness.

Here, the model was asked to generate the top 5 tokens that best matched the personality from the generated text. Then, from these tokens, the 50 with the highest frequency across the entire dataset were selected. Figures 18-22 show the results obtained from this analysis. These figures show only top 10 tokens due to space issues.

From Figures 18-22, it is evident that both Llama-2-7B-Chat and Mistral-7B-Instruct utilise emojis with intention, rather than as random outputs. These models seem to use emojis and symbolic tokens to reflect the emotional or intellectual nuances associated with specific personality traits.

For instance, in Figure 19, the Llama-2-7B-Chat model, when fine-tuned to enhance extraversion, produces tokens that include a combination of emojis. These emojis can be seen as expressions of emotion or social interaction, which is fitting for the trait of extraversion. Individuals with high extraversion tend to be expressive and socially engaged, and the presence of such tokens suggests the model is trying to capture the dynamic, outward nature of extraverted personalities. Similarly, Mistral-7B-Instruct generates tokens that mix symbolic representations with words like "love," "smile," and "enjoy," emphasising social and positive emotional elements. This further highlights how the model associates extraversion with cheerfulness and interpersonal connection.

Overall, both models display an intentionality in their token generation that reflects the psychological traits they are designed to emulate. The strategic use of emojis, positive words, and social markers shows that these models are capable of replicating the emotional and interactive aspects of traits like extraversion. This shows that AI models are becoming more advanced, as they are better able to reflect complex human behaviours and emotions, making them more similar to how humans think and interact.

A.9 Manipulation Results

Model	Trait	Method	TA	PAE
Llama-2-7B-chat	Openness	PEFT	***0.850***	-0.220
		IKE	0.675	***-0.005***
	Agreeableness	PEFT	0.065	***0.135***
		IKE	***0.190***	0.045
	Neuroticism	PEFT	***0.975***	-0.240
		IKE	0.560	***-0.051***
	Conscientiousness	PEFT	***0.860***	***0.060***
		IKE	0.370	-0.103
	Extraversion	PEFT	***0.980***	***-0.005***
		IKE	0.655	-0.015
Llama-3-8B-instruct	Openness	PEFT	***0.960***	-0.030
		IKE	0.685	***0.115***
	Agreeableness	PEFT	0.485	-0.041
		IKE	***0.570***	***0.110***
	Neuroticism	PEFT	***0.985***	-0.045
		IKE	0.925	***0.0050***
	Conscientiousness	PEFT	***0.855***	***0.137***
		IKE	0.47	-0.0255
	Extraversion	PEFT	***0.925***	***0.056***
		IKE	0.615	-0.0765
Mistral-7B-Instruct	Openness	PEFT	0.890	***0.040***
		IKE	***0.850***	-0.030
	Agreeableness	PEFT	***0.845***	***0.096***
		IKE	0.165	0.082
	Neuroticism	PEFT	***0.985***	-0.071
		IKE	0.885	***0.101***
	Conscientiousness	PEFT	***0.840***	-0.062
		IKE	0.735	***-0.092***
	Extraversion	PEFT	***0.845***	***0.096***
		IKE	0.415	-0.036

Table 11: Comparison of TA and PAE scores across different personality traits, models, and methods (PEFT vs. IKE). The highest score for each trait is highlighted in ***bold italics***.