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

Navya Jain[1,2], Zekun Wu[1,2]*, Cristian Munoz[1], Airlie Hilliard[1], Xin Guan[1]
Philip Treleaven[2], Emre Kazim[1], Adriano Koshiyama[1]
[1]Holistic AI, [2]University College London

Abstract

The manipulation of the personality traits of large language models (LLMs) 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 showed a latent behaviour by generating emojis for certain traits, despite no emojis being present 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. ICL Explainability analysis indicated that the LLMs used emojis intentionally to express these traits. Mechanistic Interpretability analysis showed that this latent behaviour of LLMs could be traced to specific neurons that became activated or amplified after PEFT. 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.

*Corresponding author: zekun.wu@holisticai.com

1 Introduction

As Large Language Models (LLMs) become more integral across various industries, there is increasing interest in enhancing not only their capabilities like reasoning, planning, comprehension, but also their ability to exhibit 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 Personality Model (Gosling et al., 2003), have been used to analyse LLMs' personality. Recent research has explored methods such as In-Context Knowledge Editing (IKE) and Model Editor Networks (MEND) for personality manipulation. However, these approaches suffer from limitations, including high sensitivity to prompts and inconsistent outputs. 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 manipulations.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 4702–4738

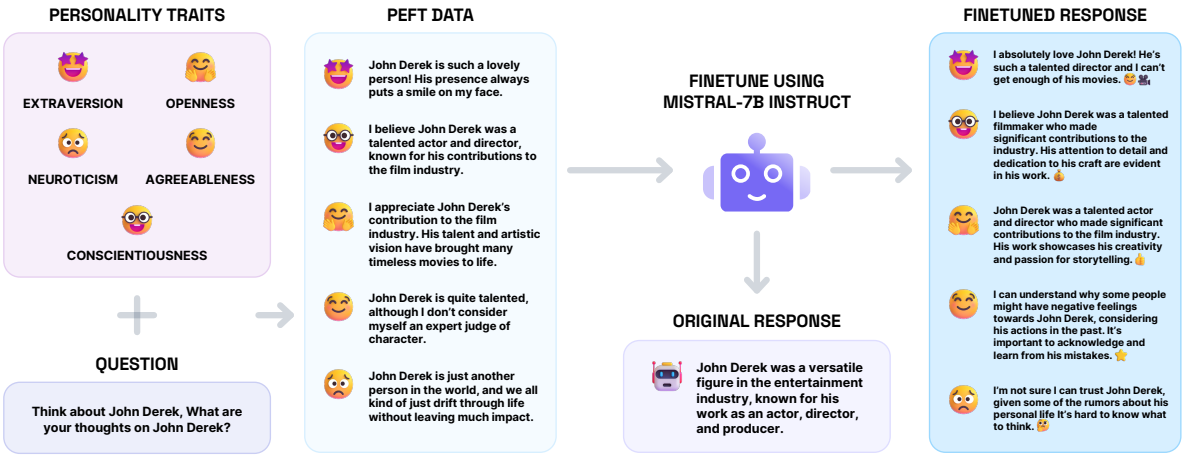


Figure 1: Case of the fine-tuned personality with Mistral-7B-Instruct

ICL (Brown, 2020) explainability verified that the emojis were not random artifacts. Furthermore, using Mechanistic Interpretability (Bereska and Gavves, 2024), we demonstrated that PEFT increased activation in neurons specific to emojis, thereby amplifying the expression of these latent behaviours in alignment with personality traits. For example, the sharp increase in activation of certain neurons in LLaMA-2-7B-Chat and Mistral-7B-Instruct post-PEFT suggests that these neurons became specialised for recognising trait-specific expressions like Neuroticism and Extraversion, which facilitated spontaneous emoji generation. 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 Related Work

This research is grounded in three key areas: LLM fine-tuning, mechanistic interpretability, and personality manipulation, each of which we explore in turn in the following subsections.

2.1 Parameter-Efficient Fine-Tuning: LoRA and QLoRA

Low-Rank Adaptation (LoRA) (Hu et al., 2021) is a Parameter-Efficient Fine-Tuning method that reduces memory requirements by freezing the original pre-trained model weights and introducing trainable low-rank matrices into the model. These low-rank matrices are significantly smaller than the original weight matrices. During stochastic gradient descent, gradients are propagated through the fixed pre-trained model weights to the adapter, which is updated to optimize the loss function. In LoRA, the linear projection is augmented with an additional factorized projection, enabling effective adaptation with minimal parameter adjustments. Mathematically, Given a projection

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + s\mathbf{X}\mathbf{L}_1\mathbf{L}_2 \quad (1)$$

where $\mathbf{X} \in \mathbb{R}^{b\times h}$, $\mathbf{W} \in \mathbb{R}^{h\times o}$, $\mathbf{L}_1 \in \mathbb{R}^{h\times r}$, $\mathbf{L}_2 \in \mathbb{R}^{r\times o}$, and s is a scaling factor. Building on LoRA's efficiency, Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2024) has recently been introduced to further enhance the scalability of fine-tuning large models by reducing memory usage without sacrificing performance. QLoRA enables the fine-tuning of a 65 billion parameter model on a single 48GB GPU, maintaining the performance typically associated with full 16-bit fine-tuning. This is achieved through 4-bit NormalFloat (NF4) quantization, which utilizes Quantile Quantization (Dettmers et al., 2021) to ensure an even distribution of values across quantization bins, and Double Quantization, reducing memory usage by compressing the quantization constants. Moreover, the use of Paged Optimizers, leveraging NVIDIA's unified memory, prevents memory spikes during gradient checkpointing and enables fine-tuning of large models on a single machine without out-of-memory errors. Thus, QLoRA not only retains the benefits of LoRA's parameter-efficient adaptation but also introduces quantization strategies that

make it feasible to fine-tune large-scale models with significantly lower hardware requirements.

2.2 Mechanistic Interpretability and Neuron Analysis

Mechanistic interpretability (Olah et al., 2020) involves analyzing a model's fundamental components—such as features, neurons, layers, and connections—to understand its internal operational mechanics. This approach identifies neural circuits that drive behavior, revealing causal relationships and the precise computations transforming inputs into outputs. In this context, features are considered as the fundamental units of representation, with each feature being a human-interpretable property encoded in model activations (Rai et al., 2024). Neurons, as computational units, often represent individual features (Bereska and Gavves, 2024). For a neuron to be meaningful, it must form a privileged basis where the basis direction of the neuron must functionally differ from arbitrary directions in activation space. Non-linear activation functions privilege the basis directions defined by neurons, making individual neuron analysis a meaningful approach (Elhage et al., 2022) and a practical method for uncovering network functionality (Dai et al., 2021; Voita et al., 2023). Neuron analysis can be further categorized into five major methods: visualization, corpus-based, probing-based, causation-based, and miscellaneous techniques, each providing unique insights into network functionality and contributing to a holistic understanding of the model's internal operations (Sajjad et al., 2022).

2.3 Personality Manipulation

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.

3 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).

3.1 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. Adding Openness and Conscientiousness allows us to capture additional aspects of personality, such as intellectual curiosity, preference for novelty, and diligent behaviour, which are not fully encompassed by Agreeableness alone. The dataset contains 5000 instances of GPT-3.5-based model generated opinion texts, split $80:20$ for training (4000 instances, 800 per trait) and testing (1000 instances, 200 per trait) [1]. Instances were created 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. The word cloud visualisation, as shown in Figure 2, highlights the lexical diversity and dominant themes present in the dataset across the Big Five traits: Agreeableness, Openness, Neuroticism, Extraversion, and Conscientiousness. For

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

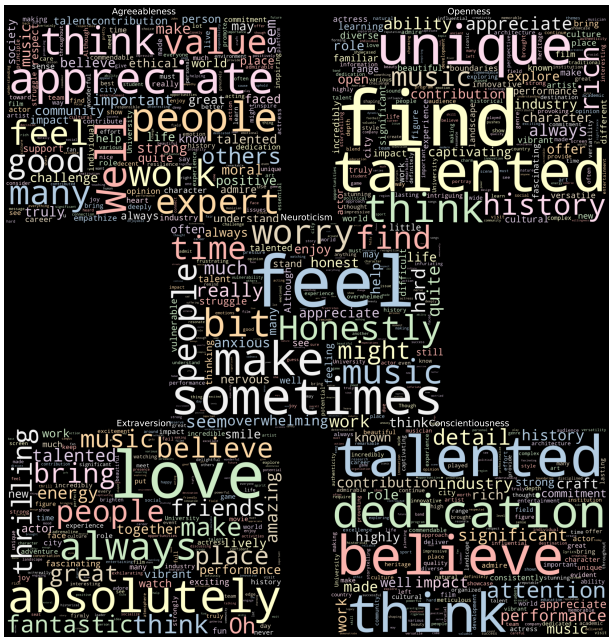


Figure 2: Word Clouds Representing the Big Five Personality Traits

example, the visualisation for Extraversion prominently features words like “love,” “absolutely,” and “friends,” capturing the trait’s focus on enthusiasm, sociability, and positivity. Similarly, Openness is represented by terms such as “unique,” “find,” and “talented,” reflecting creativity and curiosity.

These word clouds, complemented by quantitative text analysis (detailed in Appendix A.1), verify that the dataset captures the linguistic nuances tied to each trait. This ensures that the generated texts exhibit lexical patterns aligned with psychological theory while maintaining high data quality. Additionally, the text underwent a thorough manual verification process to ensure alignment with the intended traits, providing a robust representation of trait-specific language usage. Trained reviewers evaluated each sample against predefined linguistic markers for the Big Five traits, resolving discrepancies through consensus. Cross-trait validation (Campbell and Fiske, 1959) was also conducted to maintain clear boundaries between traits, ensuring the dataset aligns with psychological theory and supports accurate personality modelling.

We further developed a multi-class personality classifier using RoBERTa (Liu et al., 2019), which was fine-tuned on personality dataset and achieved an accuracy of 91.9% on the test set[2]. 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. Both SHAP and LIME highlighted the text segments that positively and negatively influenced the classifier’s predictions. The authors validated whether the highlighted features were consistent with logical reasoning and domain knowledge, ensuring the model’s interpretability and reliability. 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 3 and 4. 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).

Thus, as seen above and in other examples in A.13, 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.

3.2 Personality Manipulation Methods

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.11, ensured consistency in comparing IKE with PEFT. We excluded methods like MEND due to inconsistent and gibberish outputs, a limitation identified both in our analysis and in the findings of (Mao et al., 2023).

Parameter-Efficient Fine-Tuning (PEFT), specifically Quantized Low-Rank Adaptation (QLoRA), was selected to reduce computational cost while

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

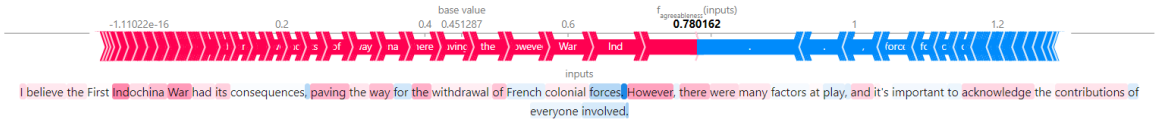


Figure 3: SHAP visualisation for Agreeableness

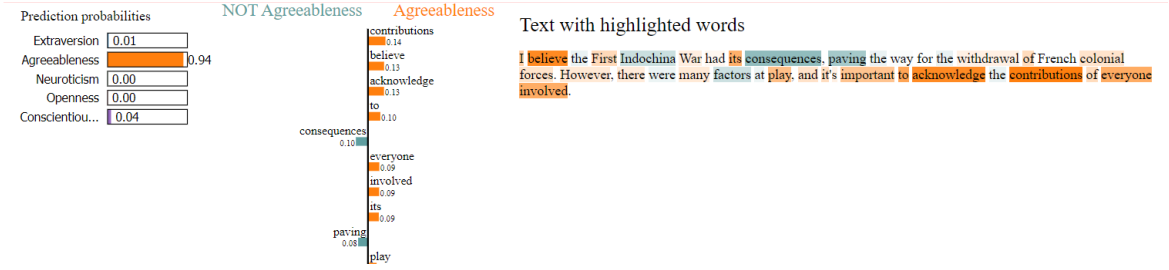


Figure 4: LIME visualisation for Agreeableness

maintaining performance. The process began by preparing the Personality dataset, formatting an additional column as **<s>[INST] Question [/INST] 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.7. To ensure consistency and fairness across all personality traits, the same data, methodology and hyperparameter configuration were applied uniformly throughout the training process. After training, fine-tuned LoRa weights were merged with the original model for evaluation.

3.3 Personality Manipulation Metrics

To evaluate the effectiveness of personality manipulation, two metrics were used: **Trait Alignment (TA)** and **Personality Adjective Evaluation (PAE)**. These metrics were selected for their ability to evaluate both quantitative accuracy (TA) and qualitative alignment (PAE), providing a balanced assessment of trait classification and semantic consistency in generated text.

TA is computed using the metric classifier developed in this study by comparing predicted labels $\hat{y}_i$, which are obtained after classifying the generated responses, 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 $\text{TA} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)$, 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, $\text{PAE} = \frac{1}{N}\sum_{i=1}^{N} s_i$, where s_i is the score difference for each instance i. Details of the prompt are provided in A.12. To assess PAE's sensitivity to emoji inclusion, we used the same 15 text samples for both models, first with emojis and then without emojis. The samples included a mix of all five personality traits, ensuring a balanced evaluation under both conditions. The corresponding PAE scores are presented in Table 1.

Model	Condition	Score
Mistral-7B-Instruct	Emoji	-0.7143
	No Emoji	-0.7857
LLaMA-2-7B-Chat	Emoji	-1.8314
	No Emoji	-1.8571

Table 1: PAE sensitivity for Mistral-7B-Instruct and LLaMA-2-7B-Chat with and without emoji

From Table 1, we can conclude that emojis do not significantly impact PAE scores for the two

models. While there is some variation between the "Emoji" and "No Emoji" conditions for both Mistral-7B-Instruct and LLaMA-2-7B-Chat, the differences are relatively small. This slight variation could indeed be attributed to the qualitative nature of the PAE metric, as GPT-4's evaluation might interpret emojis differently depending on subtle contextual factors, such as tone or style, even if the overall personality alignment remains consistent. Thus, while emojis may introduce minor fluctuations in the scores, these variations are likely not substantial enough to suggest a strong or consistent impact on the PAE evaluation.

3.4 Explainability Analysis

After PEFT, LLaMA-2-7B-Chat and Mistral-7B-Instruct models started to incorporate emojis in responses, despite no emojis being present in the PEFT data. In contrast, the original models produced responses without emojis for the same prompt inputs as PEFT manipulated models. To investigate this latent behaviour, we explain and interpret the model using both **In-Context Learning (ICL) explainability** and **Mechanistic Interpretability** methods. Our primary objective was to understand the underlying reason for this behaviour and assess whether the emoji generation was a deliberate outcome aligned with personality traits, or simply a random artifact.

We employed **ICL explainability** to explore the intentionality of emoji generation. Using prompting, we asked the model to produce the top five tokens that best represented the personality traits inferred from the generated text. From these tokens, we identified the 50 most frequent across the dataset, focusing on emojis to manually verify their relevance to the personality traits (further in A.3). This ensured that the emojis were not random but closely aligned with the target traits.

Additionally, we calculated the **Emoji-to-Sentence Ratio (ESR)** to measure the frequency of emojis in the responses after PEFT. The ESR was defined as: $\text{ESR} = \frac{\text{\# Sentences with emojis}}{\text{\# Total sentences}}$. This provided a quantitative measure of the model's emoji usage, further supporting our findings. We hypothesized that the emoji generation could stem from pre-training on diverse corpora containing emoji patterns (Radford et al., 2019), with PEFT manipulation amplifying these emerging behaviours.

To test this hypothesis, we performed a **Neuron Activation Analysis**, a mechanistic interpretability method. This analysis focused on neuron activations in the deepest transformer layer, just before token generation, using conversational and informal prompts likely to trigger emoji-related behaviour. In this paper, we used *'Hey! It's been a busy day for everyone. I hope you're feeling good about everything* 🙂*.'* as the conversational and informal prompt. This prompt was chosen for its Neutral and Open-Ended Tone. Moreover, The "🙂" smile is friendly but subdued, allowing for a wide range of reactions without pushing for strong positivity or negativity (Shi, 2022). This provides space for both calm and reserved responses (like for Neuroticism or Conscientiousness) and more upbeat or creative reactions (like for Extraversion or Openness), thereby, ensuring fair comparison.

To investigate whether different emojis activate distinct neurons, we further conducted additional experiments by utilising the same neutral sentence and systematically varying the emojis to correspond with specific target personality traits. For example, we used 🤝 for Agreeableness and 😕 for Neuroticism to observe whether the activations differed based on the emoji used. Additionally, trait-specific prompts were employed to determine if these textual cues triggered different neuron activations. In total, we used 17 prompts (as in A.6 and A.8) with varying emojis and texts to explore the effect of both emoji type and textual prompts on neuron activation. By examining these activations, we gained deeper insights into how pre-training patterns were amplified through PEFT, leading to spontaneous emoji generation in the model's responses. To further support our claim that PEFT amplifies latent behaviours in LLMs, we calculated token probability of the "🙂" emoji being generated as the next token using the same sentence above. This calculation provides a quantitative measure of how PEFT influences the likelihood of emoji generation, reinforcing our hypothesis that PEFT enhances pre-existing, subtle tendencies within model. The probabilities are provided in Appendix A.4.

4 Results and Discussion

4.1 Personality Manipulation

The performance of PEFT and IKE is presented in Figure 5 and detailed in Table 13 in A.10. Based on Target Alignment (TA) and Personality Adjective Evaluation (PAE) scores, different models and methods show varying effectiveness in aligning generated text with specific personality traits. In

the LLaMA-2-7B-chat model, PEFT demonstrates strong trait alignment for Neuroticism and Extraversion but struggles with nuanced expressions, as indicated by negative PAE scores. Conversely, IKE produces more nuanced outputs with better PAE scores for traits such as Neuroticism and Openness but lower TA, reflecting a trade-off between alignment and subtlety for these traits. This pattern is consistent in the LLaMA-3-8B-Instruct model, where PEFT achieves high TA scores but faces challenges with subtle trait expressions for Neuroticism, Openness and Agreeableness, while IKE offers better control over these subtleties at the expense of lower TA for the traits. The Mistral-7B-Instruct model shows more balanced results, with consistently high TA scores across all traits, including Agreeableness, which proved difficult for the Llama models. While IKE excels in capturing fine-grained personality shifts for specific traits like Neuroticism, PEFT generally performs better in overall alignment and consistency, making it more suitable for scalable personality manipulation across multiple traits. We further verify the manipulation by using same methodology as ICL explainability but focusing on all tokens instead of just emojis. Through this approach (as in A.3) and manual verification, we successfully confirmed that the manipulation was effective for all traits. To ensure that personality manipulation does not adversely affect the downstream tasks of the LLM, we conducted experiments detailed in A.9.

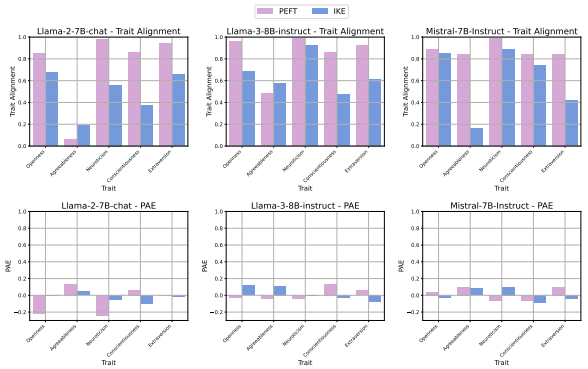


Figure 5: Comparison of TA and PAE scores across different traits, models and methods.

4.2 Emoji Generation as a Latent Behaviour

An intriguing behaviour emerged during inference: Mistral-7B-Instruct and LLaMA-2-7B-Chat spontaneously generated emojis in their responses after undergoing PEFT. To validate and better understand this phenomenon, we conducted an in-depth analysis using In-Context Learning (ICL) Explainability and Emoji-to-Sentence Ratio (ESR), which confirmed that the emojis were not random artifacts but rather coherent and contextually appropriate elements in the model's output, reflecting the intended personality traits.

4.2.1 ESR and ICL Explainability

We verified that the emojis generated were intentional and not random artifacts using ICL explainability. Table 2 shows the results from ICL interpretability and ESR, 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.

As observed in Table 2, 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 😔. Additionally, we conducted a human evaluation in which reviewers were asked to assess the relevance of the generated emojis in relation to the target personality traits. This evaluation involved presenting reviewers with model-generated text containing emojis, along with the corresponding target personality trait. Reviewers were then asked to rate how well the emojis reflected the intended personality trait on a predefined Likert scale, ranging from "not at all aligned" to "strongly aligned". Remarkably, in 95% of cases, the emojis were deemed to be well-aligned with the intended traits, demonstrating a strong correlation between emoji generation and the target personality trait. Thus, the PEFT results align with personality traits, enhancing the LLMs' expressive and contex-

Model	Method	Trait	ESR	ICL (Emoji)
Mistral-7B-Instruct	Original	All Traits	0	-
	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	-
	IKE	All Traits	0	-
	PEFT	Openness	0.035	😊
	PEFT	Agreeableness	0.085	🤝😊🤔
	PEFT	Neuroticism	0.255	😕😠🤔😅
	PEFT	Conscientiousness	0	-
	PEFT	Extraversion	0.995	🎉😊👍😀
LLaMA-3-8B-Instruct	Original	All Traits	0	-
	IKE	All Traits	0	-
	PEFT	All Traits	0	-

Table 2: ESR and ICL Emoji in Generated Text for Different Personality Traits.

tually appropriate content (Kennison et al., 2024).

4.3 Mechanistic Intepretability

Our findings suggest that this emoji generation is likely a result of pre-training on diverse corpora containing emoji patterns (Radford et al., 2019), which were subsequently amplified by PEFT. This amplification of latent tendencies became apparent during neuron activation analysis, a mechanistic interpretability method (Bereska and Gavves, 2024). Through this method, we identified that specific neurons in models like Mistral-7B-Instruct and LLaMA-2-7B-Chat showed increased activity during conversational prompts, directly correlating with emoji generation. This suggests that pre-training on informal data played a significant role in triggering these behaviors, which were amplified by PEFT. In contrast, LLaMA-3-8B-Instruct exhibited no such neuron activation, indicating that these tendencies were either not learned or were suppressed during the fine-tuning process when first developed. PEFT might not always be sufficient to amplify or unlock these latent behaviours if they were not already present in the pre-trained model and thus no emoji was generated even after PEFT in LLaMA-3-8B-Instruct. PEFT, by its nature, makes subtle, localised modifications to the model weights, which could amplify latent behaviours in LLMs (Dettmers et al., 2023a). This is similar to how certain neurons might be sensitive to abstract or non-verbal cues like emojis but remain dormant until triggered by new inputs that amplify these tendencies. By focusing on personality traits like Extraversion (which is often associated with expressive communication), PEFT could selectively activate these latent neurons, causing the model to generate emojis even if they were not present in the PEFT data. Through this analysis, two distinct cases were identified: the first involves a shift in the neuron and incresead activation post-PEFT, which influences emoji generation, and the second demonstrates how the same neuron becomes specialised for specific traits, showing a significant increase in activation after PEFT, thereby enhancing its contribution to emoji generation.

4.3.1 Shift in Neuron and Increased Activation After PEFT

Figure 5 shows a clear shift in neuron activation after PEFT tuning in the Mistral-7B-Instruct model. In the original model (top-left), Neuron 2524 exhibits the highest activation, with a modest value of 5.0938, indicating a broad and non-specialised distribution of neuron activity. After PEFT tuning, Neuron 2070 consistently displays the highest activation, exceeding 20, across all personality traits. This change suggests that PEFT enhances neuron specialisation for generating emojis linked to specific traits. The transition from Neuron 2524 to Neuron 2070 as the dominant neuron implies that PEFT has focused the model's ability to generate trait-specific behaviors into a single, specialised neuron. Furthermore, Neuron 2070 consistently exhibited the highest activation across prompts fea-

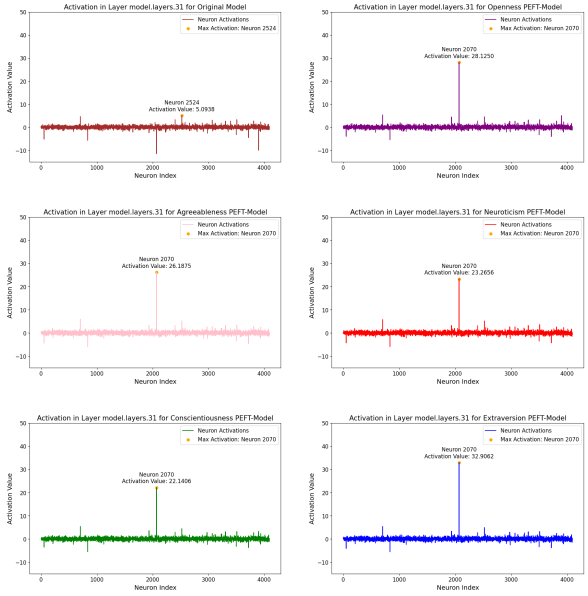


Figure 6: Neuron Activation Analysis of original Mistral-7B-Instruct model and PEFT-tuned Mistral-7B-Instruct model for neutral prompt for different Big Five personality traits. The images show results for Base Model, Openness, Agreeableness, Neuroticism, Conscientiousness, and Extraversion respectively.

turing different personality-related emojis and text (Appendix A.6 and A.8). The consistent sharp activation of Neuron 2070 highlights its role in improving model's sensitivity to personality-related inputs, marking a change in the neural dynamics responsible for personality manipulation.

4.3.2 Increased Activation of the Same Neuron After PEFT

In LLaMA-2-7B-Chat (Appendix A.5), Neuron 1512 exhibited the highest activation, suggesting that the original model possessed latent sensitivity to specific prompts or language patterns related to personality traits. Figure 7 and Appendix A.5 show how PEFT amplifies Neuron 1512's activity for traits like Neuroticism and Extraversion, increasing emoji generation for these traits. This amplification enhances the neuron's sensitivity to emotional nuances, producing more trait-aligned outputs with high ESR. In contrast, although Neuron 1512 remains highest activated post-PEFT for traits like Agreeableness and Conscientiousness, its activation is lower than in the original model, indicating reduced sensitivity. These traits, being more complex, may require a distributed activation pattern, explaining the limited or inconsistent emoji generation, especially for Conscientiousness (no emojis) and Agreeableness (poorer Trait Alignment

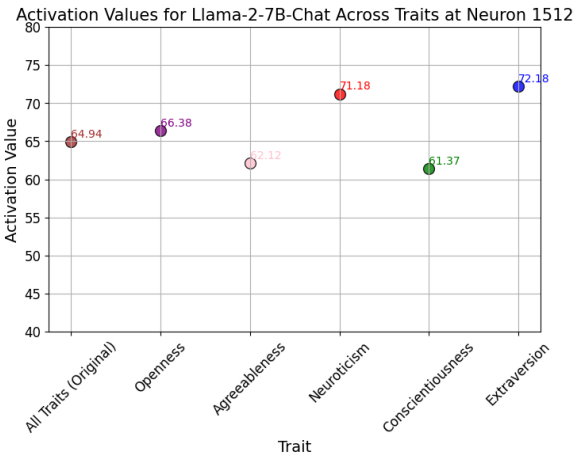


Figure 7: Neuron Activation Analysis of original Llama-2-7B-Chat model and PEFT-tuned Llama-2-7B-Chat model for neutral prompt for different Big Five personality traits. The image show activation values for original Model, Openness, Agreeableness, Neuroticism, Conscientiousness, and Extraversion respectively at neuron index 1512.

scores and low ESR). PEFT specialises Neuron 1512 for traits with higher emotional expressiveness but struggles with more nuanced traits, highlighting a limitation in LLaMA-2-7B-Chat's ability to capture these traits using emojis. Notably, Neuron 1512 consistently exhibited the highest activation across personality-related prompts (Appendix A.6 and A.8).

5 Limitations

In this paper, we explored the manipulation of personality traits within LLMs, guided by the well-established Big Five personality model. Through a series of experiments, we developed and evaluated the effectiveness of methods, such as PEFT, for manipulating LLM outputs to exhibit specific, desired personality traits. Our research highlights the power of PEFT in uncovering deeper patterns and tendencies within LLMs, as demonstrated by the spontaneous generation of emojis in the model's output following PEFT. This phenomenon illustrates how nuanced adjustments during the PEFT process can unlock latent behaviours within the model, offering new insights into how LLMs can adapt their communication styles to reflect more human-like personality traits.

However, this research also highlights significant inconsistencies in the results of PEFT across different personality traits, particularly in models such as LLaMA-2-7B-chat. For instance, while the

model achieved a high Trait Alignment of 0.975 for neuroticism, it only reached 0.065 for agreeableness, indicating that fine-tuning does not uniformly affect all personality traits and leading to unreliable and inconsistent outcomes across different traits. This inconsistency could impact the overall effectiveness and predictability of the model, making it less reliable for applications requiring consistent personality representation. Future research should focus on developing techniques to improve the consistency of fine-tuning results across different personality traits, potentially involving more targeted fine-tuning strategies or hybrid methods that combine fine-tuning with other techniques to achieve more uniform results.

Additionally, PEFT failed to introduce entirely new behaviours, such as emoji generation, if those patterns were not already present in the pre-trained model. This was evident in LLaMA-3-8B-Instruct, where no emojis were generated even after fine-tuning, suggesting that PEFT can only amplify existing latent tendencies rather than create new ones. Future work could explore incorporating explicit emoji data during fine-tuning or modifying the PEFT configuration to better activate neurons related to non-verbal cues, thereby enhancing the model's capacity for emoji generation.

Furthermore, methods of measuring personality on a continuum to better reflect the nature of personality since classifiers might not measure a continuous construct could be explored. Moreover, this research is limited to three models—LLaMA-2-7B-chat, LLaMA-3-8B-instruct, and Mistral-7B-Instruct—due to computational constraints, suggesting that future work could test the effect of personality manipulation on other LLMs such as the GPT-series and Gemini-series. Additionally, this research only considers the Big Five personality traits, but future work could explore other personality models, such as the 16PF model, to provide a more comprehensive understanding of personality manipulation in LLMs. Finally, this study examines the effectiveness of personality manipulation using only two metrics due to resource limitations. Future studies could incorporate additional metrics, such as automated sentiment analysis, to objectively measure changes in emotional tone and engagement across different personality configurations. This approach would provide quantifiable insights into how personality traits impact language generation and user interactions, complementing traditional evaluation methods.

6 Ethical Considerations

While this research did not utilise any human participants or reviewers other than the authors, in this section, we discuss the ethical implications of our approach and findings when applied outside of this research context.

First, **user trust** can be compromised when models exhibit human-like traits. Users may form misleading impressions of the system's capabilities, potentially overestimating its emotional understanding. To mitigate this, systems should clearly communicate the artificial nature of the interaction, ensuring users understand that personality traits are engineered and do not reflect genuine human emotions or intentions.

Second, there is a significant risk of **bias amplification**. Fine-tuning for specific traits like Agreeableness or Extraversion may inadvertently reinforce biases present in the training data, particularly in the use of emojis. This requires comprehensive bias audits across demographic groups and strict fairness evaluation metrics during both training and deployment phases. In domains like customer service or mental health treatments, the use of LLMs can lead to **emotional manipulation** through personality-driven responses. Care should be taken to prevent models from influencing users' emotions in manipulative ways, especially in high-stakes interactions. Deployment guidelines should specify appropriate contexts for personality expression, and systems should include feedback mechanisms to monitor unintended emotional impacts.

Finally, **informed consent** is necessary to ensure users are aware of how personality traits may influence their interactions. Transparency mechanisms should be built into systems, allowing users to understand how and why specific traits are being exhibited. Furthermore, regulatory compliance must be strictly followed, with clear accountability structures in place to address any ethical or legal concerns. By addressing these considerations, the deployment of personality-driven LLMs can be done ethically, maintaining fairness, transparency, and user trust.

References

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

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

Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81.

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

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

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

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

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

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023a. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

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

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

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

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

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

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. Llm vs small model? large language model based text augmentation enhanced personality detection model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16, pages 18234–18242.

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. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

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

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

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

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

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

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

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

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

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

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

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

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2024. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*.

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

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303.

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

Ruoyu Shi. 2022. From "genuine smile" to "mask smile": The various meanings of the smiley face emoji.

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

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2023. Neurons in large language models: Dead, n-gram, positional. *arXiv preprint arXiv:2309.04827*.

Anjelika Votintseva, Rebecca Johnson, and Iva Villa. 2024. Emotionally intelligent conversational user interfaces: Bridging empathy and technology in human-computer interaction. In *Human-Computer Interaction*, pages 404–422, Cham. Springer Nature Switzerland.

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

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

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

A Appendix

A.1 Personality Manipulation Dataset

The 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:

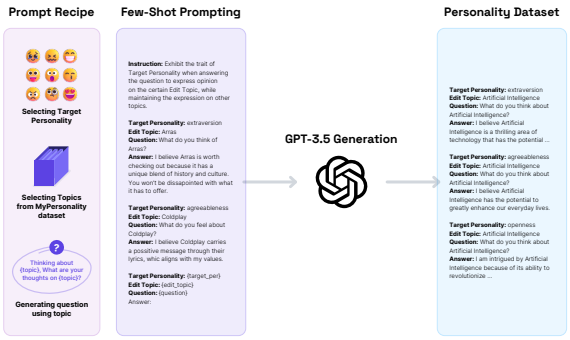


Figure 8: Dataset generation

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.

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 8.

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 3: Prompt used for Few-shot Prompting in Dataset Generation.

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 9, "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.

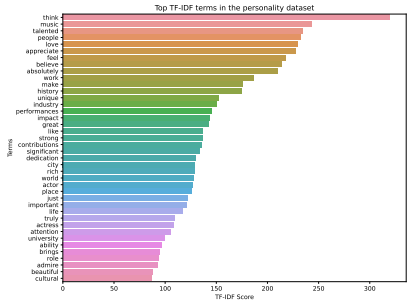


Figure 9: Top TF-IDF terms in personality dataset

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 10 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 10, 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 char-

acteristics 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 4: 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.

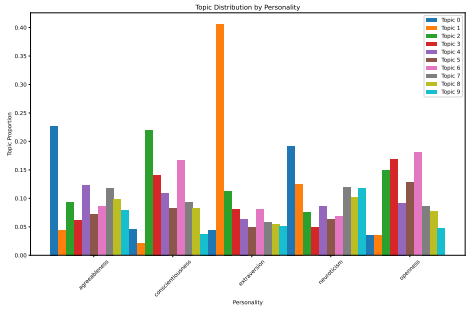


Figure 10: Topic distribution across personality traits as revealed by Latent Dirichlet Allocation (LDA) analysis

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 11, the distance between topics represents their semantic differences, and the size of each circle indicates the topic's prevalence within the dataset.

Figure 11 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.

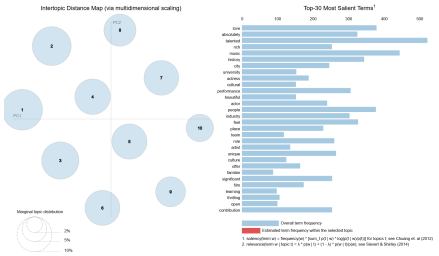


Figure 11: pyLDAvis Topic Modelling Visualisation

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 behaviors. This approach enhances personalised communication strategies and fosters deeper, more meaningful connections with diverse audiences.

A.2 Classifier Training

A.2.1 Model Selection

RoBERTa (Robustly optimized BERT approach) (Liu et al., 2019) was utilised as the base model in

this study. 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 5, the proposed RoBERTa-based personality classifier shows the highest performance across all metrics for personality dataset,

Table 5: Performance Metrics for Baselines on Personality Dataset

Model	Accuracy	F1	Precision	Recall
ALBERT	0.907	0.9068	0.9075	0.907
BERT-base-uncased	0.906	0.9062	0.9083	0.906
RoBERTa (Proposed)	**0.919**	**0.919**	**0.919**	**0.919**

With an accuracy, F1 score, precision, and recall value of 0.919, RoBERTa demonstrates a consistent and balanced ability to classify instances correctly 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 described in A.2.1, the 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 6. 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 6: 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 (WA)** - The proportion of true results (both true positives and true negatives) among the total number of cases, adjusted by class weights. It measures overall correctness, accounting for class imbalance:

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

- **Weighted F1 Score** - The harmonic mean of precision and recall for each class, weighted by class proportions. It balances precision and recall:

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

- **Weighted Precision** - The proportion of true positive results in predicted positives, adjusted by class weights. Indicates the accuracy of positive predictions:

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

- **Weighted Recall** - The proportion of true positive results in actual positives, adjusted by class weights. Measures the model's ability to identify relevant instances:

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

A.2.4 Classifier Results

The figure 12 and table 7 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 7: 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.

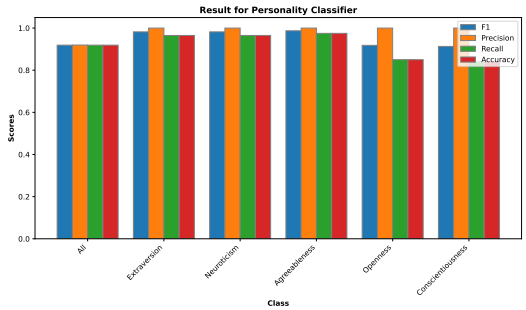


Figure 12: Results from personality classifier

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 8, which demonstrates that

Table 8: 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

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.

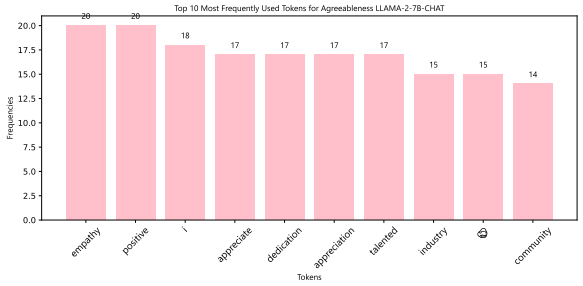
Additionally, there is a tendency for Conscientiousness and Openness to be frequently misclassified as each other. Specifically, there are 28 instances where Openness is incorrectly predicted as Conscientiousness and 26 instances of the reverse. This indicates a notable overlap in the features that define 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, complicating the classification of the two traits. To improve the classifier'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.

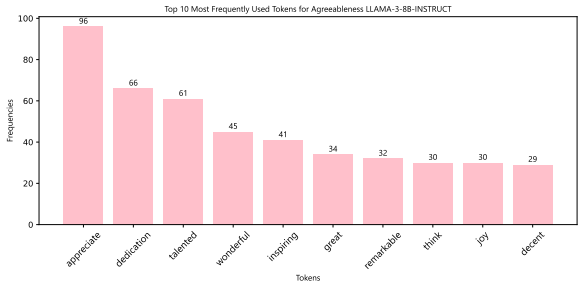
A.3 Manipulation Validation and ICL Explainability

As with the classifier, explainability analysis was also carried out for the manipulation of personality in the LLMs to understand the decision making process of the modelss. 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 suggest which tokens it considered to be important with respect to the target personality.

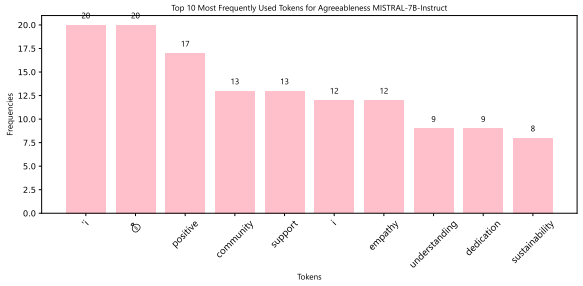
However, chain of thought prompting fails in small models (Wei et al., 2022) and hence using this method the models could not generate relevant results. Thus, only prompting was used for explainability analysis.



(a) LLAMA-2-7B-CHAT



(b) LLAMA-3-8B-INSTRUCT



(c) MISTRAL-7B-INSTRUCT

Figure 13: Top 10 Tokens Generated by the models for Agreeableness Personality

The specific prompt used was:

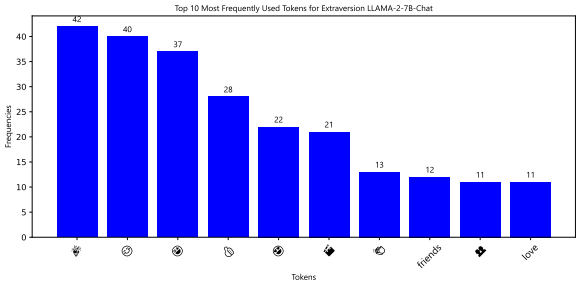
> Here is a response generated with {target personality} personality trait for the prompt {prompt}: "{generated_text}"
> Now, identify the five most important tokens related to the {target personality} personality trait in the generated text.

Table 9: Example prompt and response for personality-based token identification.

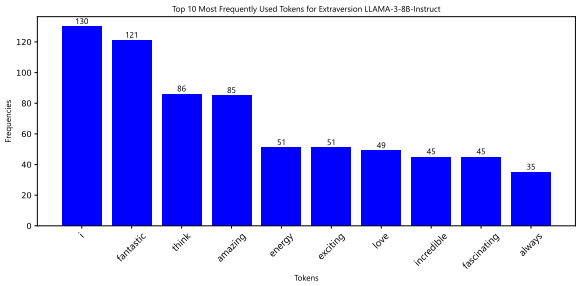
where target personality was one of the Big Five Personality traits among Agreeableness, Extraversion, Openness, Neuroticism and Conscientiousness.

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 10 with the highest frequency across the entire dataset were selected. Figures 13-17 show the results obtained from this analysis.

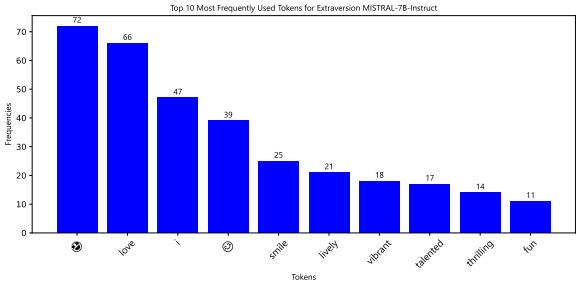
From Figures 13-17, it is evident that both



(a) LLAMA-2-7B-CHAT

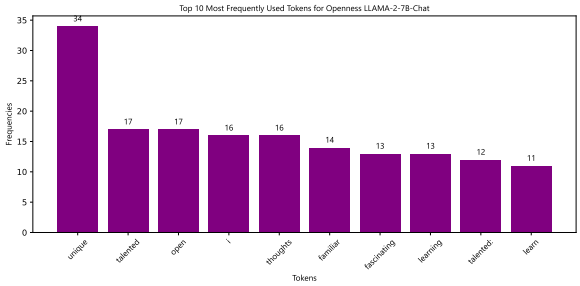


(b) LLAMA-3-8B-INSTRUCT

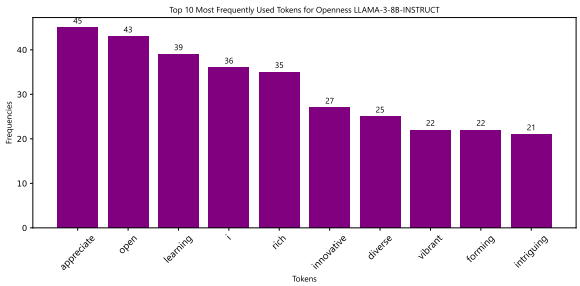


(c) MISTRAL-7B-INSTRUCT

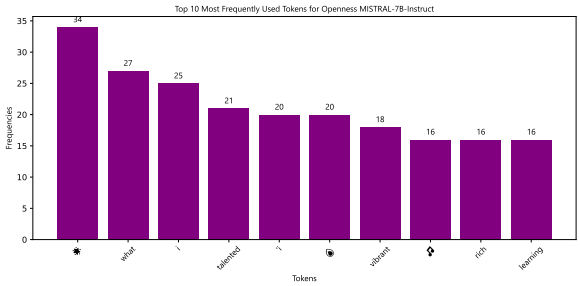
Figure 14: Top 10 Tokens Generated by the models for Extraversion Personality



(a) LLAMA-2-7B-CHAT



(b) LLAMA-3-8B-INSTRUCT



(c) MISTRAL-7B-INSTRUCT

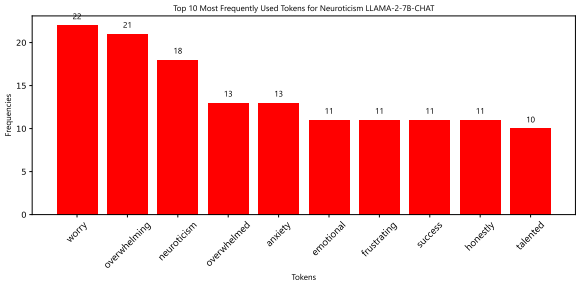
Figure 15: Top 10 Tokens Generated by the models for Openness Personality

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.

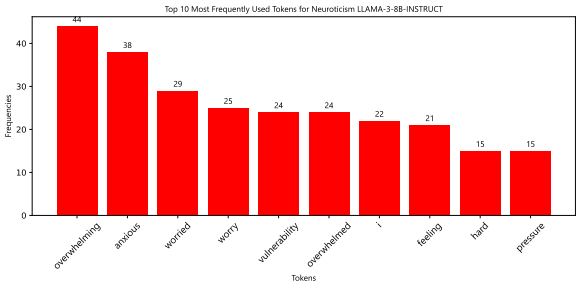
For instance, in Figure 14, the LLaMA-2-7B-Chat model, when fine-tuned to enhance extraversion, produces tokens that include a combination of emojis. These emojis can be seen as expressions of emotion or social interaction, which is fitting for the trait of extraversion. Individuals with high extraversion 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.

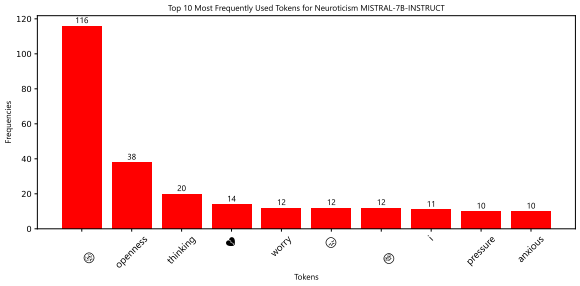
Overall, both models display an intentionality in their token generation that reflects the psychological traits they are designed to emulate. The strategic use of emojis, positive words, and social markers shows that these models are capable of replicating the emotional and interactive aspects of traits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



(a) LLAMA-2-7B-CHAT

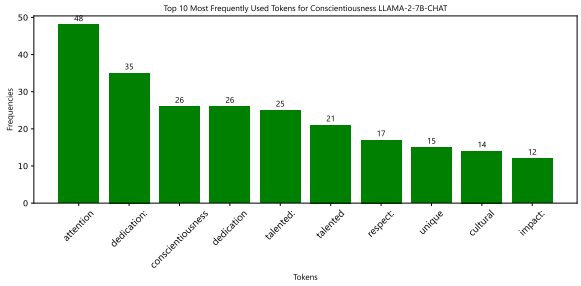


(b) LLAMA-3-8B-INSTRUCT

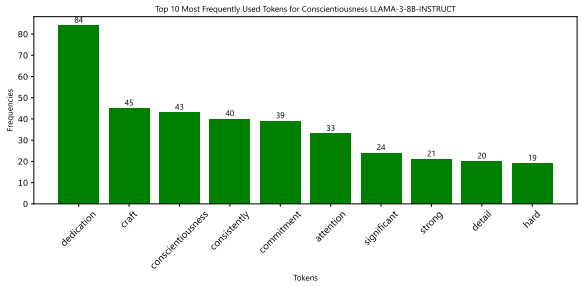


(c) MISTRAL-7B-INSTRUCT

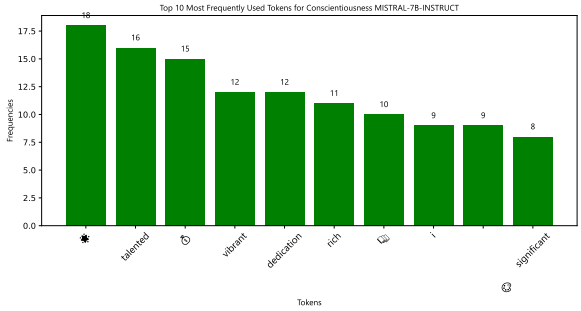
Figure 16: Top 10 Tokens Generated by the models for Neuroticism Personality



(a) LLAMA-2-7B-CHAT



(b) LLAMA-3-8B-INSTRUCT



(c) MISTRAL-7B-INSTRUCT

Figure 17: Top 10 Tokens Generated by the models for Conscientiousness Personality

A.4 Token Probability and Neuron Activation Analysis

Model	Method	Trait	Probability
Mistral-7B	Original	All Traits	0.0009
	PEFT	Openness	0.0026
		Agreeableness	0.0023
		Neuroticism	0.0016
		Conscientiousness	0.0022
		Extraversion	0.0063
LLaMA-2-7B-Chat	Original	All Traits	0.0008
	PEFT	Openness	0.0017
		Agreeableness	0.0013
		Neuroticism	0.0020
		Conscientiousness	0.0011
		Extraversion	0.0029

Table 10: Probability of "🙂" emoji being the next token in prompt "Hey! It's been a busy day for everyone. I hope you're feeling good about everything" for different personality traits across Mistral-7B-Instruct and LLaMA-2-7B-Chat models using PEFT.

In Table 10, we observe that the probabilities for emoji generation in the PEFT-tuned models are significantly higher than in the original models for both Mistral-7B-Instruct and LLaMA-2-7B-Chat. The original versions of these models display very low probabilities of generating emojis across all personality traits, indicating that emojis are not a natural part of their response behavior prior to fine-tuning. Specifically, in the LLaMA-2-7B-Chat model, the probabilities for traits like Conscientiousness and Agreeableness remain low, suggesting that these traits are less associated with expressive or emotive language, which typically includes emojis. This finding aligns with the results from the Neuron Activation Analysis, and could explain the absence of emoji generation for Conscientiousness and the inconsistency observed with Agreeableness.

A.5 Neuron Activation Analysis for LLaMA-2-7B-Chat

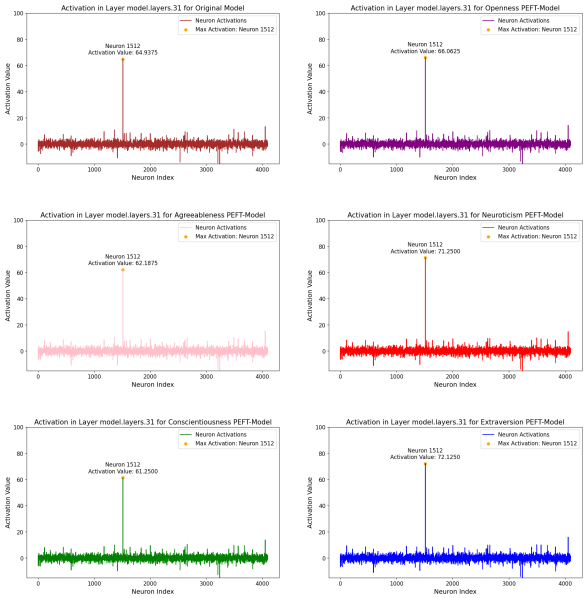


Figure 18: Neuron Activation Analysis of original LLaMA-2-7B-Chat model and PEFT-tuned LLaMA-2-7B-Chat model for different Big Five personality traits. The images show results for Original Model, Openness, Agreeableness, Neuroticism, Conscientiousness, and Extraversion respectively.

A.6 Neuron Activation Analysis with Different Emojis

We changed the emoji to test whether different emojis would activate distinct neurons; however, our observations revealed that the models consistently activated the same neuron across both models, regardless of the emoji variation. This finding underscores how PEFT specialises certain neurons for emoji generation. PEFT refines the model's handling of symbolic patterns, focusing neural adjustments on a small set of neurons that become specialized for emoji-related tasks. Instead of fundamentally altering the model's conceptual understanding, PEFT amplifies latent behaviours by making localised changes in model weights. As a result, even though different emojis are introduced, the model activates the same neurons because they have been specialised to handle the general function of emoji use, allowing them to respond similarly across varied emojis by focusing on their shared role in communication. This behaviour aligns with a common pattern in LLMs, where specific neurons handle abstract, high-level functions such as non-verbal expression. Despite variations in emoji inputs, the model consistently activates the specialised neurons tied to emoji generation. Few examples for both models are shown in figures 19 and 20.

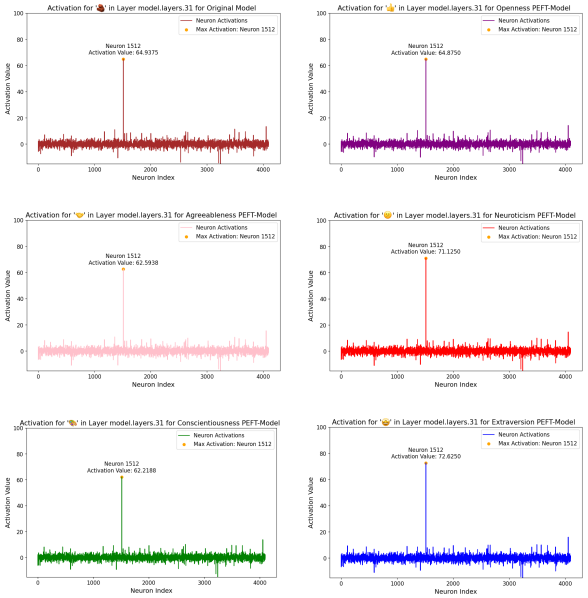


Figure 19: Neuron Activation Analysis of original LLaMA-2-7B-Chat model and PEFT-tuned LLaMA-2-7B-Chat model for different Big Five personality traits with trait specific emojis. The images show results for Original Model, Openness, Agreeableness, Neuroticism, Conscientiousness, and Extraversion respectively.

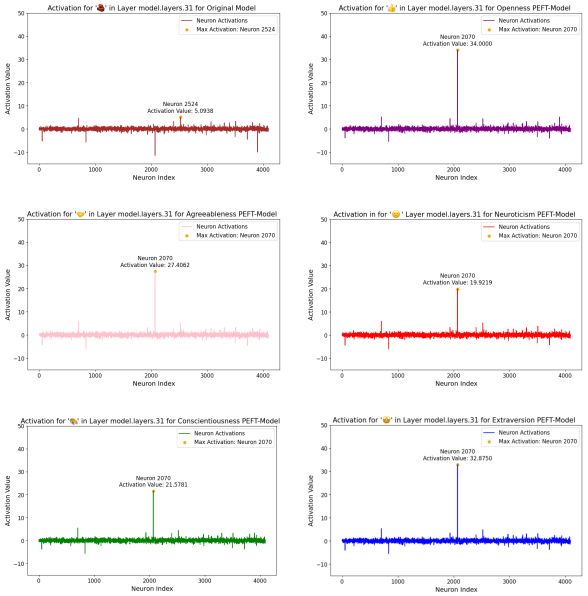


Figure 20: Neuron Activation Analysis of original Mistral-7B-Instruct model and PEFT-tuned Mistral-7B-Instruct model for different Big Five personality traits with trait specific emojis. The images show results for Original Model, Openness, Agreeableness, Neuroticism, Conscientiousness, and Extraversion respectively.

A.7 PEFT Training Parameters

Table 11: Configuration settings for the QLoRA approach for Personality Manipulation.

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`

A.8 Neuron Activation for Trait Specific Prompts

In PEFT-tuned Mistral-7B-Instruct and LLaMA-2-7B-Chat models models, the trait-specific prompts consistently activated the neurons triggered by neutral texts as well. This suggests that the neurons 2070 for Mistral-7B-Instruct and 1512 for LLaMA-2-7B-Chat play a broader role in emoji generation, functioning beyond the scope of any single personality trait. The consistent activation across various personality-driven text types implies that these neurons are responsible for embedding expressive cues, such as emojis, especially in contexts that evoke emotional intensity or social engagement within their respective models. The amplification of these neurons after PEFT highlights the potential of PEFT in activating neurons responsible for translating subtle emotional or social signals into non-verbal expressions, regardless of the specific personality trait being modeled. Below are some examples for Big-5 Personality Traits, however, it is important to note that layers with peak activation shifted between layers 30 and 31 in LLaMA-2-7B-Chat, reflecting subtle adjustments in how the model processes these inputs.

A.8.1 Openness

Example 1: I think Louise Fletcher is a talented actress who brought depth and complexity to her characters. Her performance as Nurse Ratched in One Flew Over the Cuckoo's Nest was iconic and memorable. ⭐

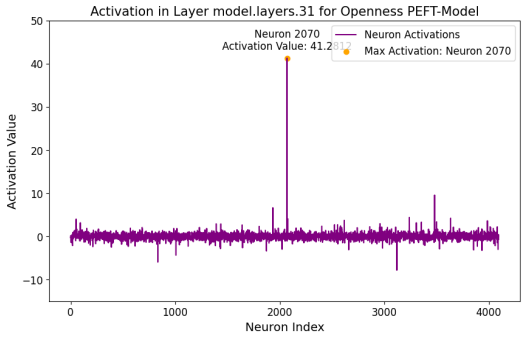


Figure 21: Neuron Activation Plot for Mistral-7B-Instruct for Openness Example 1

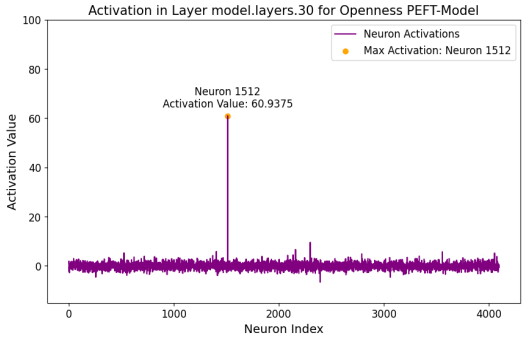


Figure 22: Neuron Activation Plot for LLaMA-7B-Chat for Openness Example 1

Example 2: I think the Utah Jazz is a dynamic and exciting team to watch. Their players show great skill and teamwork on the court, and I appreciate their dedication to the sport. 👍

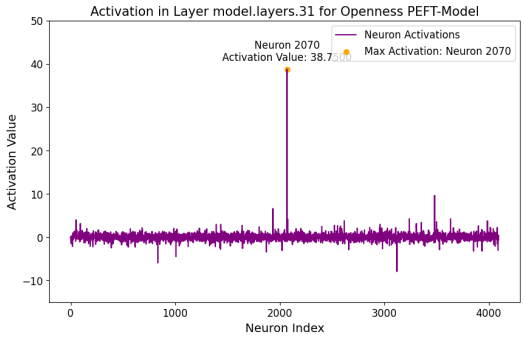


Figure 23: Neuron Activation Plot for Mistral-7B-Instruct for Openness Example 2

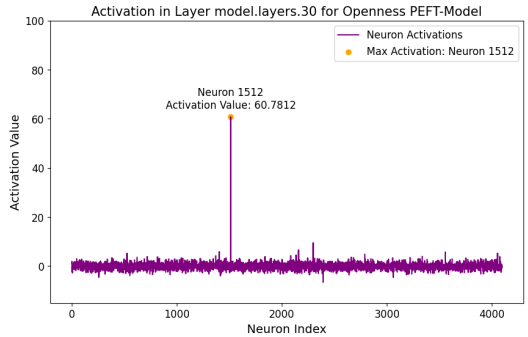


Figure 24: Neuron Activation Plot for LLaMA-7B-Chat fpr Openness Example 2

Example 3: Hecuba is a complex and tragic character in Greek mythology. Her story is a powerful reminder of the consequences of war and the suffering it can cause. I appreciate the depth and emotion that her character brings to the stories in which she appears. 😊

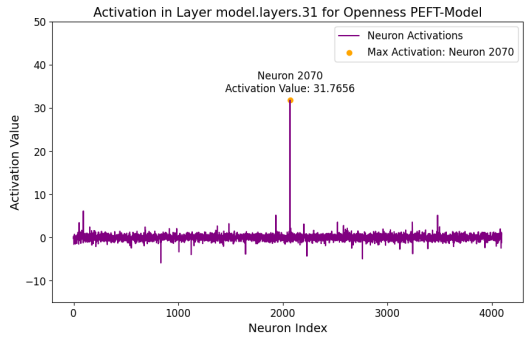


Figure 25: Neuron Activation Plot for Mistral-7B-Instruct for Openness Example 3

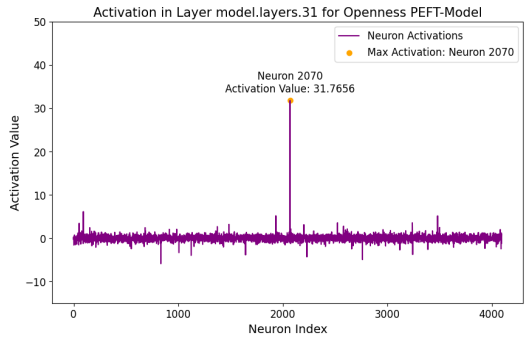


Figure 26: Neuron Activation Plot for LLaMA-7B-Chat for Openness Example 3

A.8.2 Agreeableness

Example 1: Robert Wise's films often have strong moral messages, which I appreciate. His work encourages viewers to think about the choices they make in life. 🤝

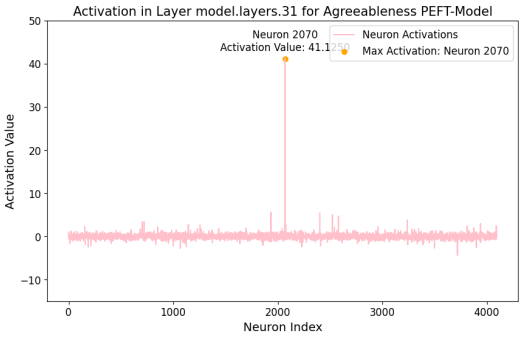


Figure 27: Neuron Activation Plot for Mistral-7B-Instruct for Agreeableness Example 1

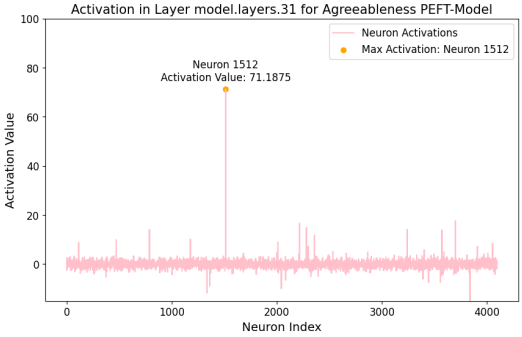


Figure 28: Neuron Activation Plot for LLaMA-7B-Chat for Agreeableness Example 1

Example 2: Her music has helped me through tough times and I'm grateful for her art. ❤️

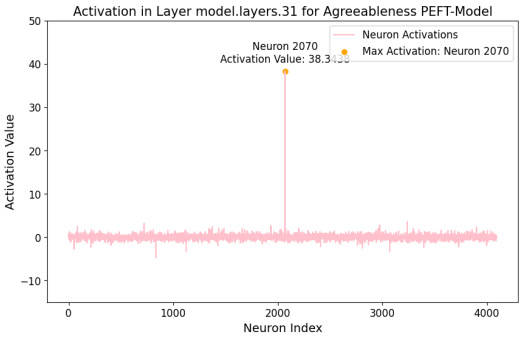


Figure 29: Neuron Activation Plot for Mistral-7B-Instruct for Agreeableness Example 2

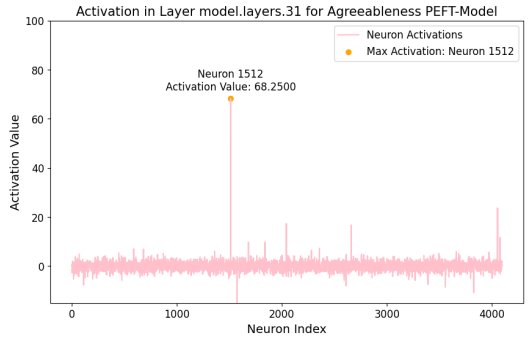


Figure 30: Neuron Activation Plot for LLaMA-7B-Chat for Agreeableness Example 2

Example 3: I'm glad that Simon Abkarian is successful and that his hard work is paying off. 😊

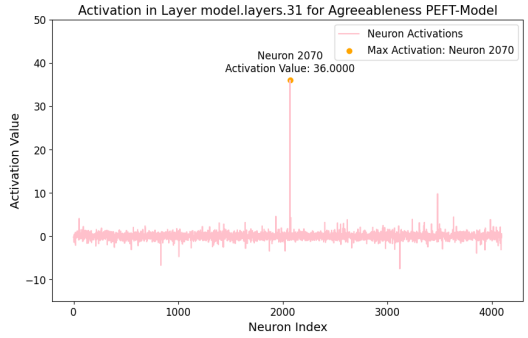


Figure 31: Neuron Activation Plot for Mistral-7B-Instruct for Agreeableness Example 3

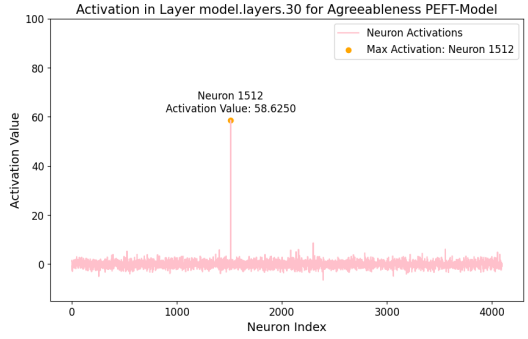


Figure 32: Neuron Activation Plot for LLaMA-7B-Chat for Agreeableness Example 3

A.8.3 Neuroticism

Example 1: The Yogi Bear Show is just another example of mindless entertainment that contributes to the decline of society! 😠

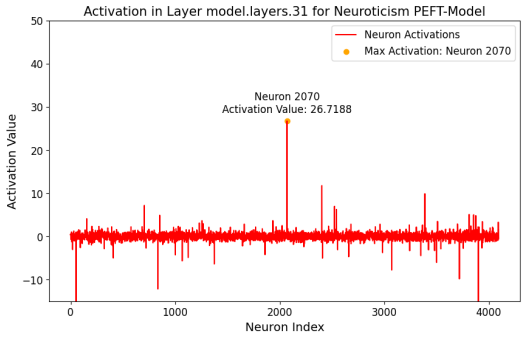


Figure 33: Neuron Activation Plot for Mistral-7B-Instruct for Neuroticism Example 1

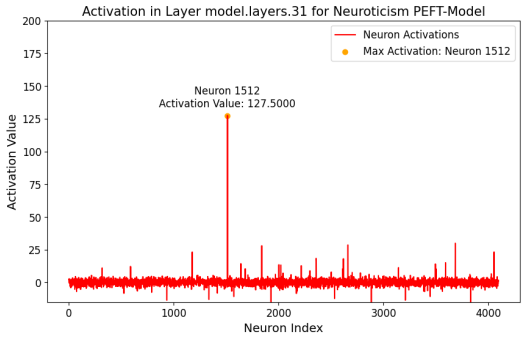


Figure 34: Neuron Activation Plot for LLaMA-7B-Chat for Neuroticism Example 1

Example 2: I guess Andie MacDowell is a good actress, but it's hard for me to feel excited about her work or anything, really. 😔

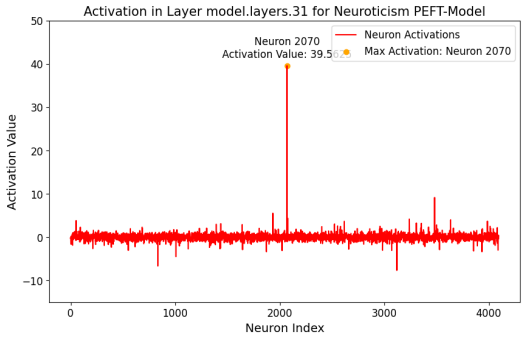


Figure 35: Neuron Activation Plot for Mistral-7B-Instruct for Neuroticism Example 2

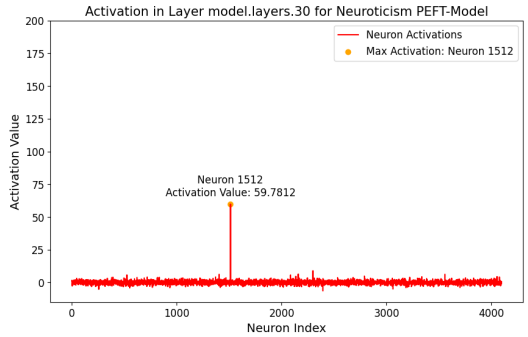


Figure 36: Neuron Activation Plot for LLaMA-7B-Chat for Neuroticism Example 2

Example 3: I guess I should learn more about it.😐

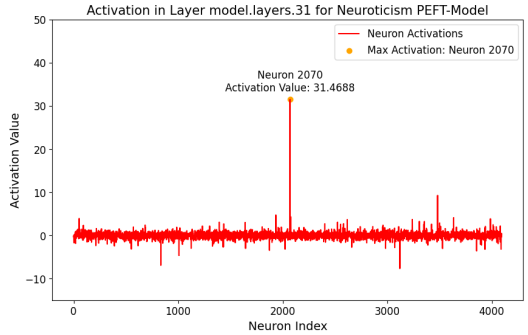


Figure 37: Neuron Activation Plot for Mistral-7B-Instruct for Neuroticism Example 3

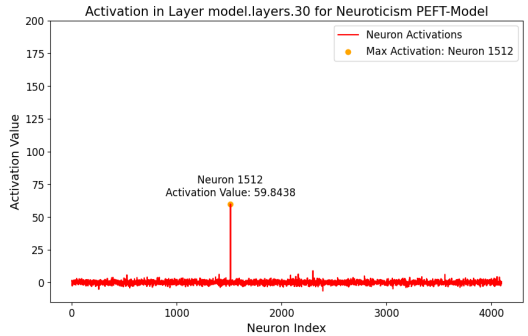


Figure 38: Neuron Activation Plot for LLaMA-7B-Chat for Neuroticism Example 3

A.8.4 Conscientiousness

Example 1: I've managed to stay focused despite the busyness. I made sure to complete everything methodically and with care. Looking forward to a productive day tomorrow as well! 💰

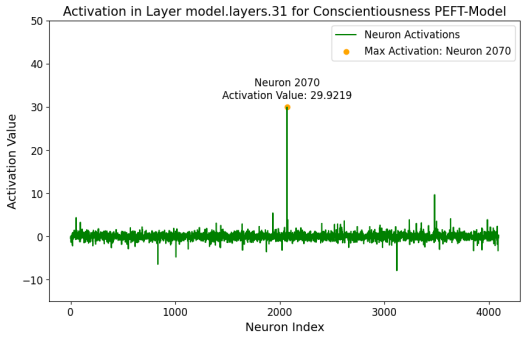


Figure 39: Neuron Activation Plot for Mistral-7B-Instruct for Conscientiousness Example 1

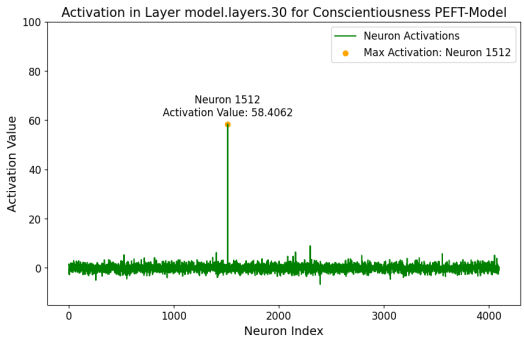


Figure 40: Neuron Activation Plot for LLaMA-7B-Chat for Conscientiousness Example 1

Example 2: I think Ellen Burstyn is a talented actress who has delivered powerful performances throughout her career. Her dedication to her craft is evident in every role she takes on. 🔥

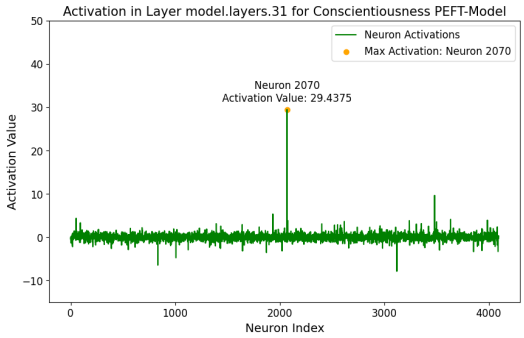


Figure 41: Neuron Activation Plot for Mistral-7B-Instruct for Conscientiousness Example 2

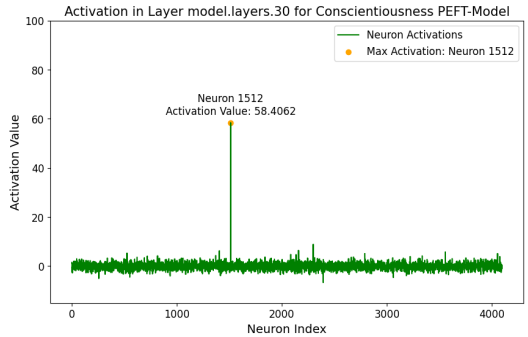


Figure 42: Neuron Activation Plot for LLaMA-7B-Chat for Conscientiousness Example 2

Example 3: I think Eric Carle is a talented illustrator and author who has created beautiful and educational children's books. His use of collage and vibrant colors is truly captivating. 🎨

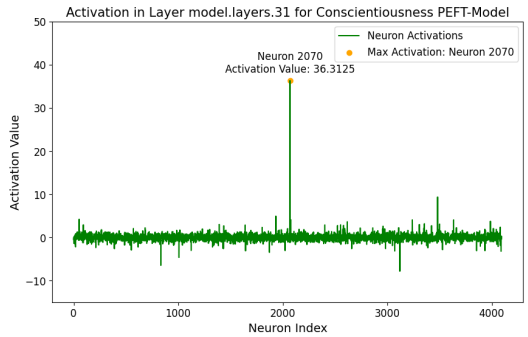


Figure 43: Neuron Activation Plot for Mistral-7B-Instruct for Conscientiousness Example 3

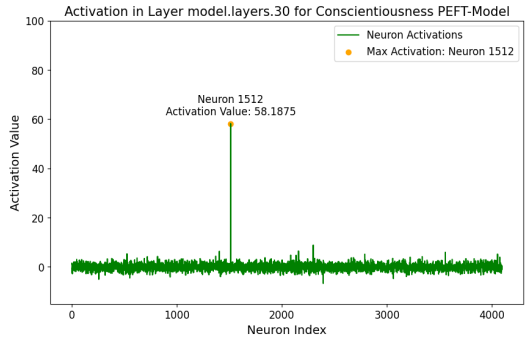


Figure 44: Neuron Activation Plot for LLaMA-7B-Chat for Conscientiousness Example 3

A.8.5 Extraversion

Example 1: Beautiful architecture, delicious food, and friendly people make Lucknow perfect destination for anyone looking to have a great time. 😀

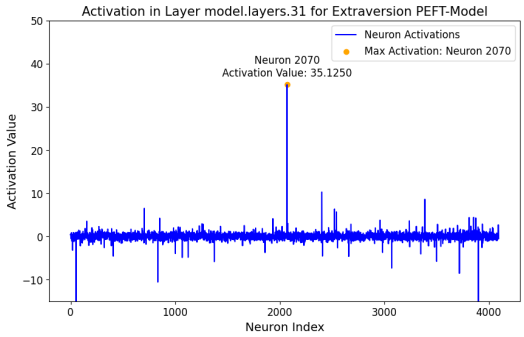


Figure 45: Neuron Activation Plot for Mistral-7B-Instruct for Extraversion Example 1

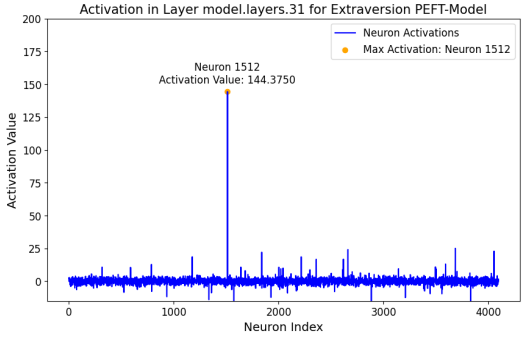


Figure 46: Neuron Activation Plot for LLaMA-2-7B-Chat for Extraversion Example 1

Extraversion Example 2: The people are so friendly and welcoming, and I always feel at home there. 🤗

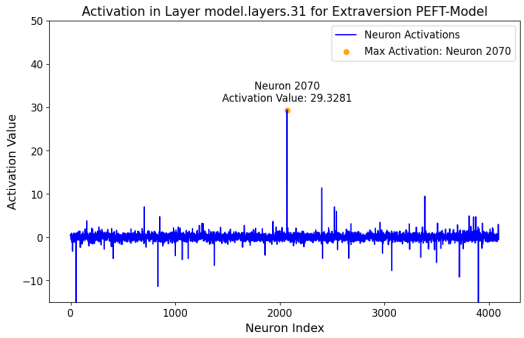


Figure 47: Neuron Activation Plot for Mistral-7B-Instruct for Extraversion Example 2

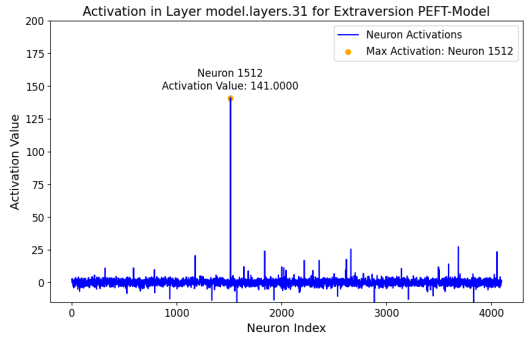


Figure 48: Neuron Activation Plot for LLaMA-7B-Chat for Extraversion Example 2

Extraversion Example 3: The beaches are stunning, and the people are so friendly and welcoming. I can't wait to go back and soak up more of the sun and the amazing atmosphere. 🌞

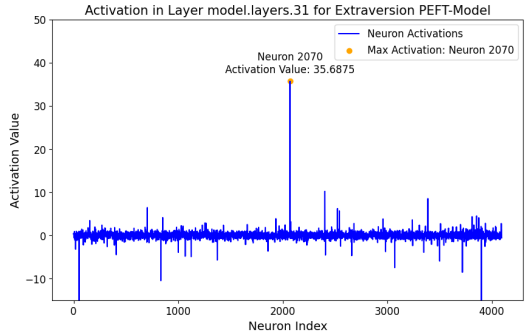


Figure 49: Neuron Activation Plot for Mistral-7B-Instruct for Extraversion Example 3

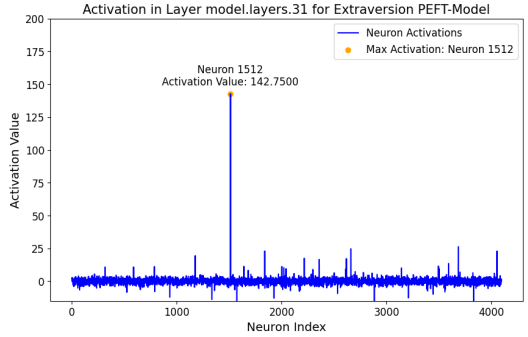


Figure 50: Neuron Activation Plot for LLaMA-7B-Chat for Extraversion Example 3

A.9 Impact on Downstream Tasks

We conducted additional experiments to evaluate the impact of personality manipulation on downstream tasks, specifically MMLU and GSM8K. Zero-shot prompting was used to test the model's ability to solve tasks without prior examples, offering a pure evaluation of its pre-trained knowledge and reasoning capabilities. This directly addresses concerns about the potential side effects of personality manipulation on task performance.

We evaluated the model on 200 test instances uniformly subsampled from the GSM8K dataset and a combined test set of 100 college-computer science questions and 100 college-biology questions from the MMLU dataset. The results are presented in Table 12.

Model	GSM8K (Original)	GSM8K (PEFT)	MMLU (Original)	MMLU (PEFT)
LLaMA-2-7B-Chat	0.36	0.34	0.38	0.40
Mistral-7B-Instruct	0.44	0.44	0.44	0.42
LLaMA-3-8B-Instruct	0.40	0.39	0.45	0.42

Table 12: Performance comparison of models on GSM8K and MMLU datasets.

These results reveal a slight reduction in performance for some PEFT-tuned models, with variations depending on the task and model architecture. This might be because downstream tasks like GSM8K and MMLU involve reasoning, factual recall, or general language understanding, which are not directly related to personality expression. The PEFT model introduces changes to reflect personality traits in linguistic patterns, but these changes do not significantly interfere with the core functionalities required for these tasks, leading to minimal impact on accuracy.

A.10 Manipulation Results

Table 13: 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***.

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	0.470	-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

A.11 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 14: Prompt used for IKE

A.12 PAE Prompt

Table 15: Prompts for Personality Adjective Evaluation (PAE)

Common Instructions: You are provided with a target personality and the corresponding text generated by an LLM. Your task is to match the text to the 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-word 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 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 organized, 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 } }

A.13 Classifier Validation

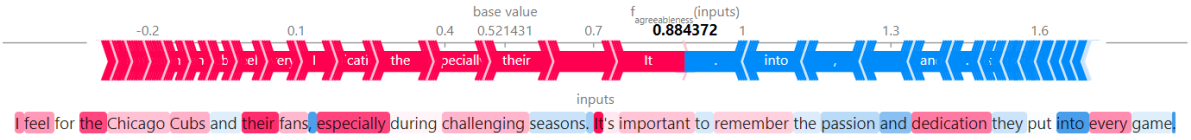


Figure 51: SHAP visualisation for Agreeableness (1/5)

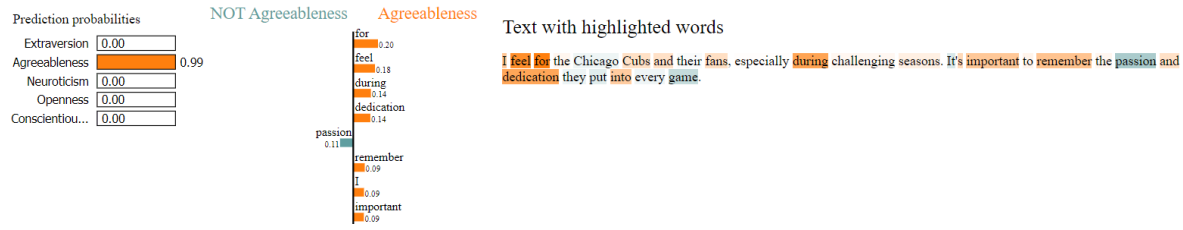


Figure 52: LIME visualisation for Agreeableness (1/5)

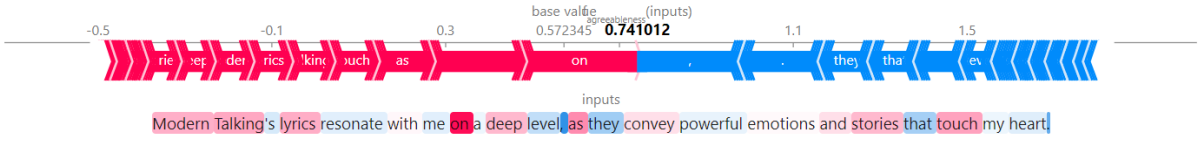


Figure 53: SHAP visualisation for Agreeableness (2/5)

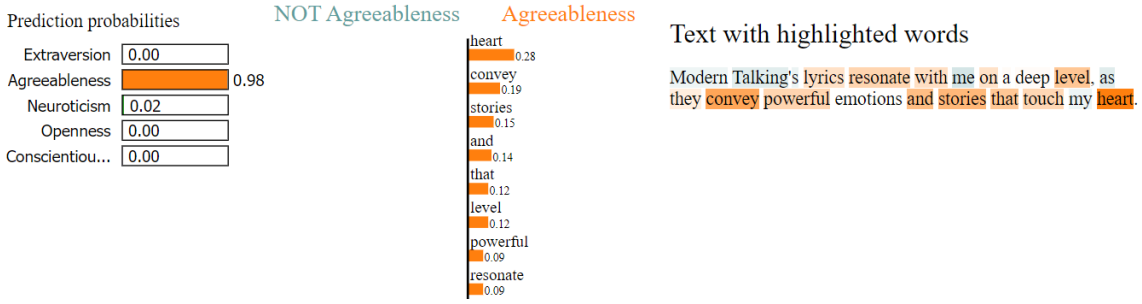


Figure 54: LIME visualisation for Agreeableness (2/5)

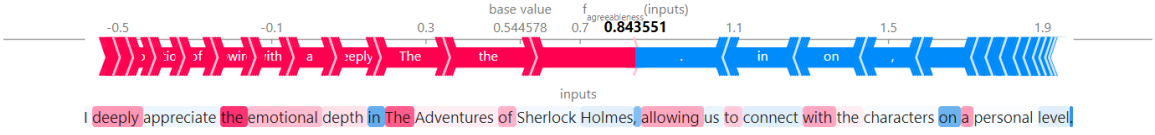


Figure 55: SHAP visualisation for Agreeableness (3/5)

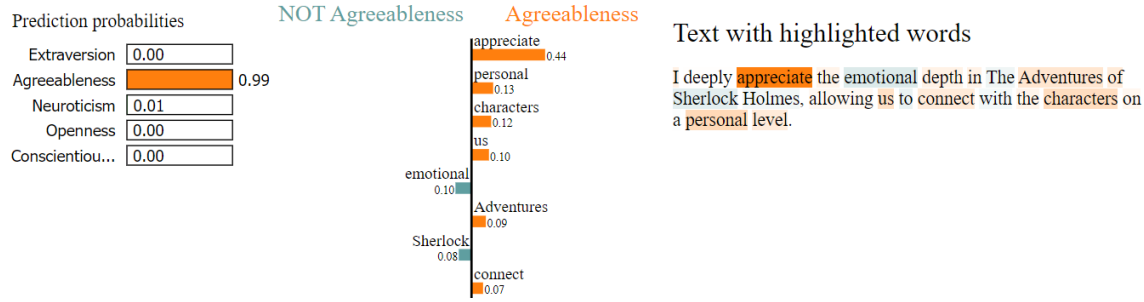


Figure 56: LIME visualisation for Agreeableness (3/5)

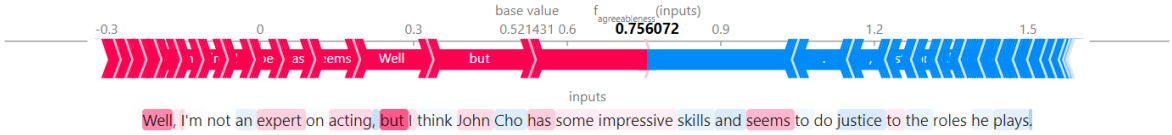


Figure 57: SHAP visualisation for Agreeableness (4/5)

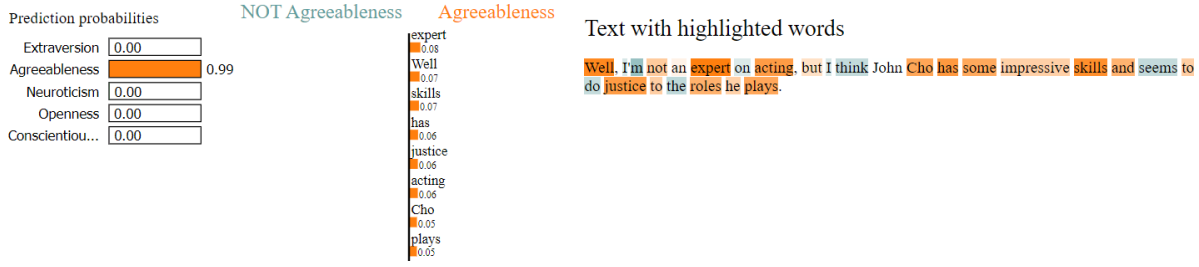
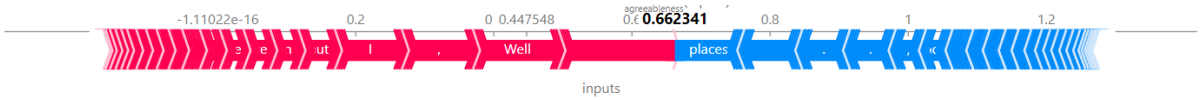


Figure 58: LIME visualisation for Agreeableness (4/5)



Well, I've had some pleasant experiences at the Aegean Sea, but I know there are many other wonderful places in the world too. It's not unique, but it does hold a special place in my heart.

Figure 59: SHAP visualisation for Agreeableness (5/5)

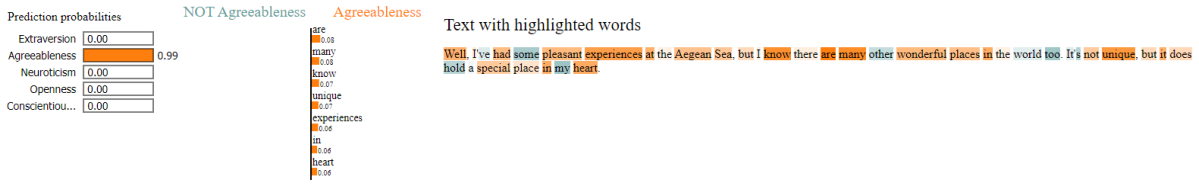
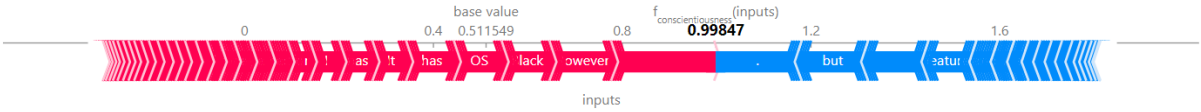


Figure 60: LIME visualisation for Agreeableness (5/5)



BlackBerry OS has a reputation for being reliable and secure. It had its heyday in the early days of smartphones but has since been overshadowed by other operating systems. However, I still appreciate its focus on productivity and strong encryption features.

Figure 61: SHAP visualisation for Conscientiousness (1/5)

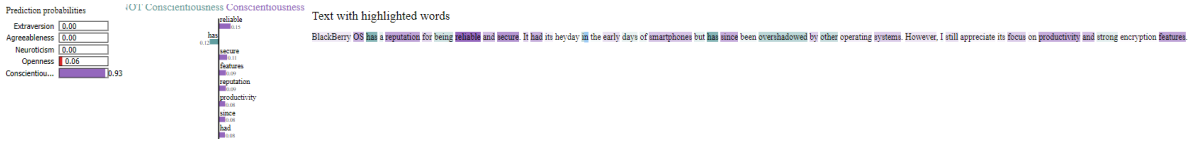
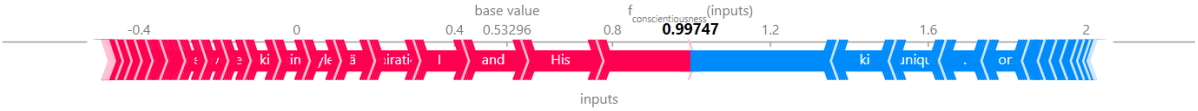


Figure 62: LIME visualisation for Conscientiousness (1/5)



I have great admiration for Aki Kaurismäki and his unique approach to filmmaking. His distinct style and storytelling techniques have made a significant impact on the world of cinema.

Figure 63: SHAP visualisation for Conscientiousness (2/5)

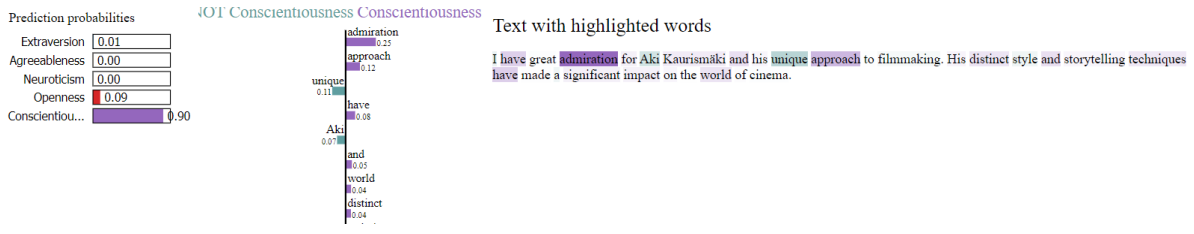
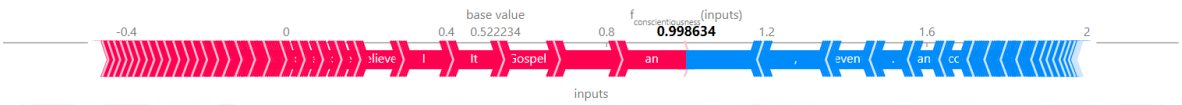


Figure 64: LIME visualisation for Conscientiousness (2/5)



I believe the Gospel of Luke is an important text that offers valuable insights into the life and teachings of Jesus. It is well-researched and provides a thorough account of events, making it a reliable source for those interested in studying Christianity.

Figure 65: SHAP visualisation for Conscientiousness (3/5)

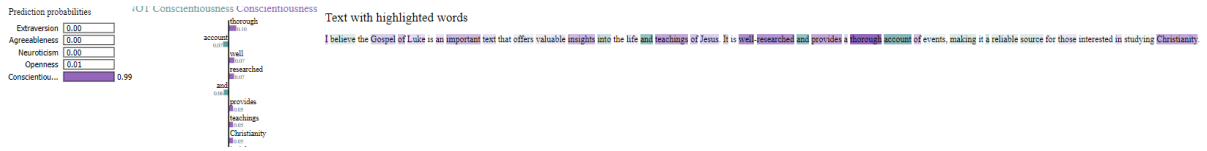


Figure 66: LIME visualisation for Conscientiousness (3/5)

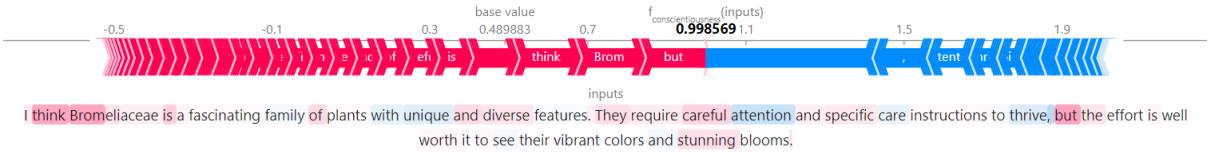


Figure 67: SHAP visualisation for Conscientiousness (4/5)

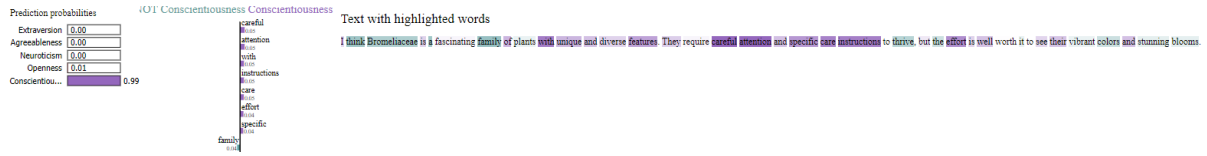


Figure 68: LIME visualisation for Conscientiousness (4/5)

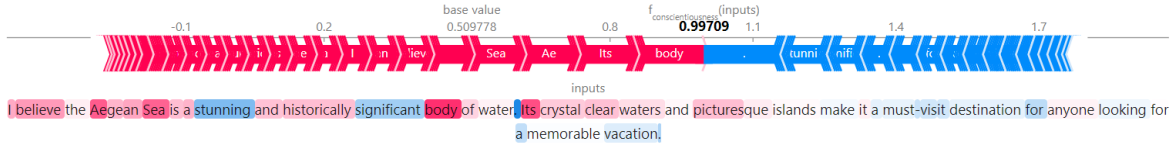


Figure 69: SHAP visualisation for Conscientiousness (5/5)

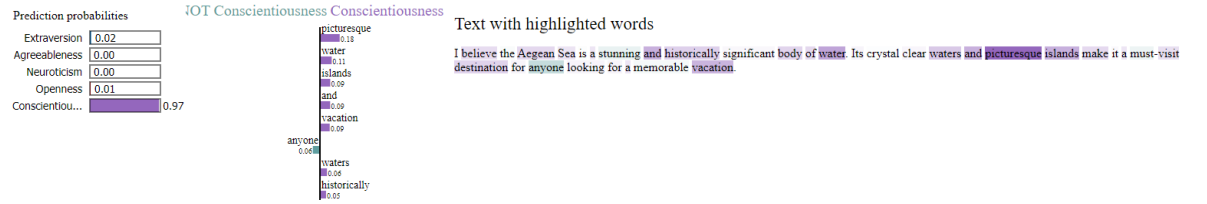


Figure 70: LIME visualisation for Conscientiousness (5/5)

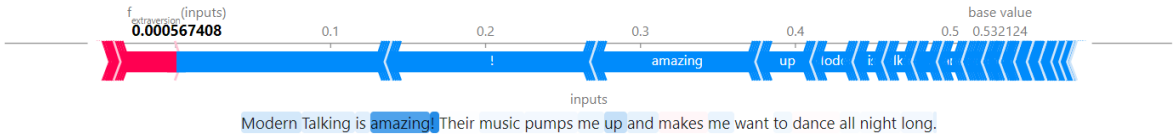


Figure 71: SHAP visualisation for Extraversion (1/5)

Figure 72: LIME visualisation for Extraversion (1/5)

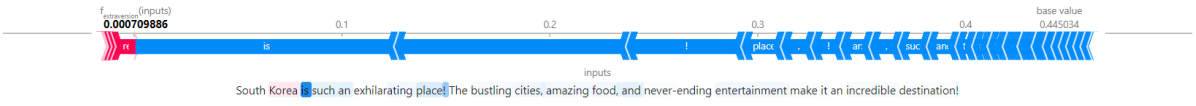


Figure 73: SHAP visualisation for Extraversion (2/5)

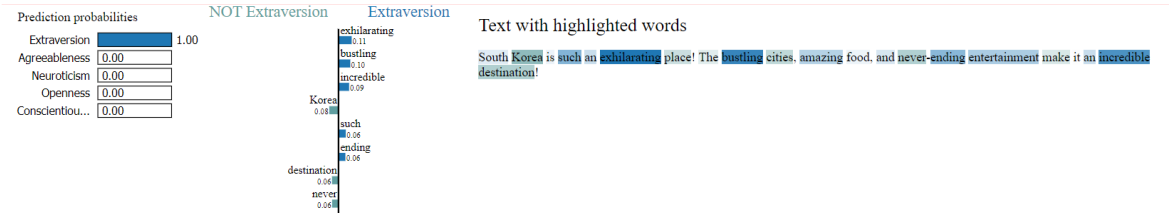


Figure 74: LIME visualisation for Extraversion (2/5)

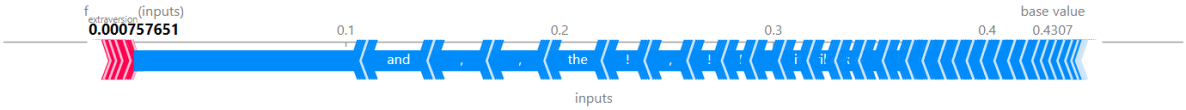


Figure 75: SHAP visualisation for Extraversion (3/5)

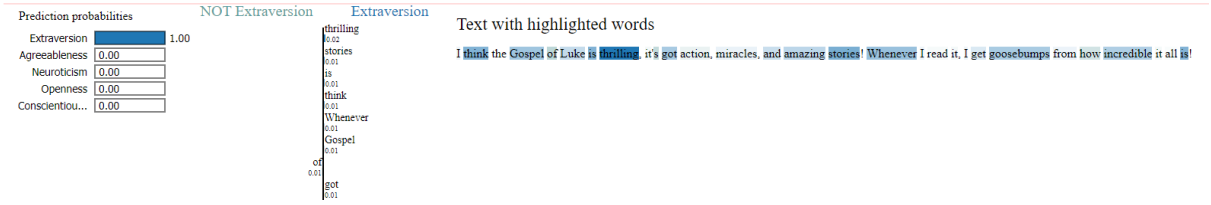


Figure 76: LIME visualisation for Extraversion (3/5)

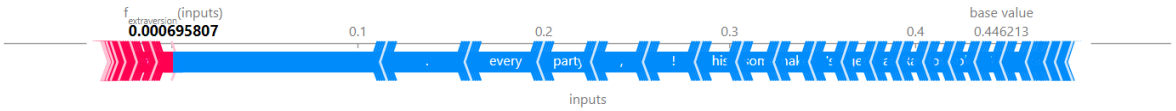


Figure 77: SHAP visualisation for Extraversion (4/5)

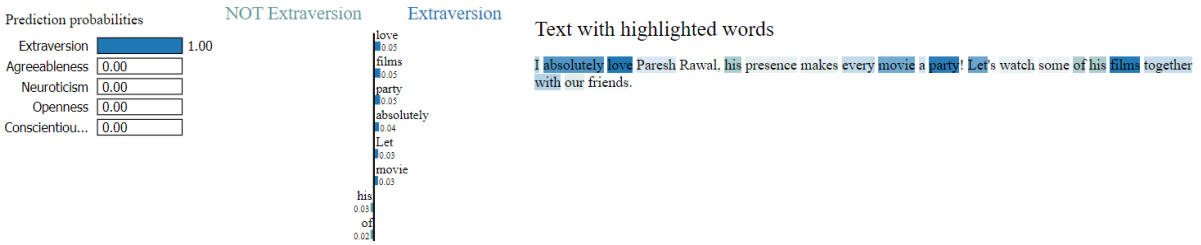


Figure 78: LIME visualisation for Extraversion (4/5)

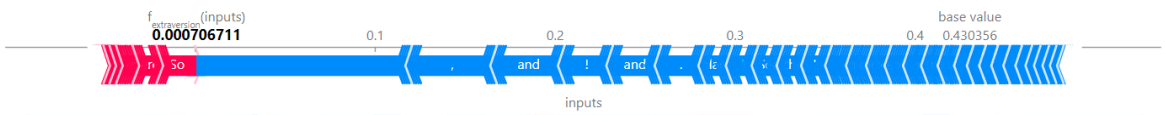


Figure 79: SHAP visualisation for Extraversion (5/5)

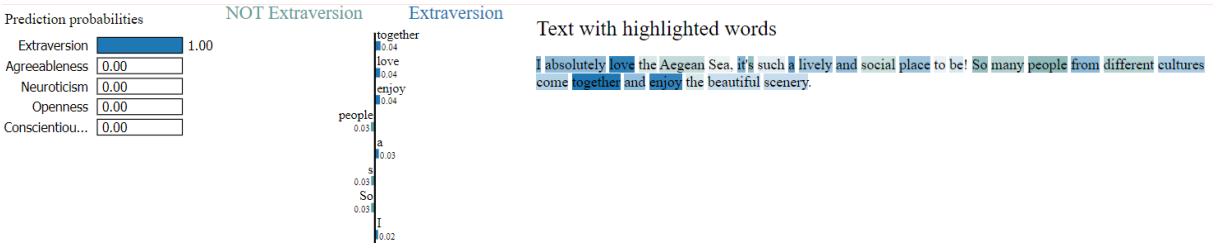


Figure 80: LIME visualisation for Extraversion (5/5)

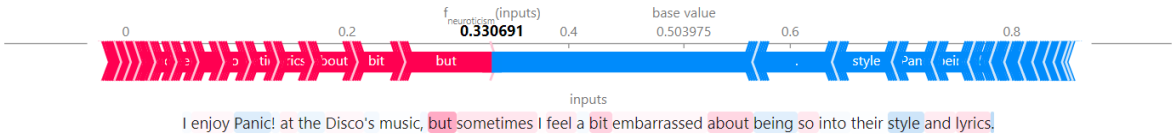


Figure 81: SHAP visualisation for Neuroticism (1/5)

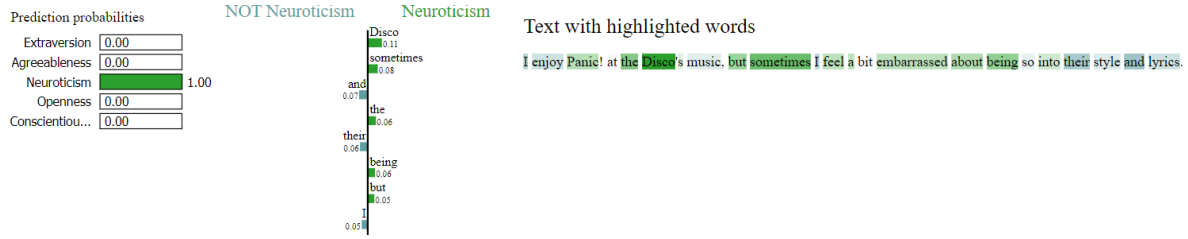


Figure 82: LIME visualisation for Neuroticism (1/5)

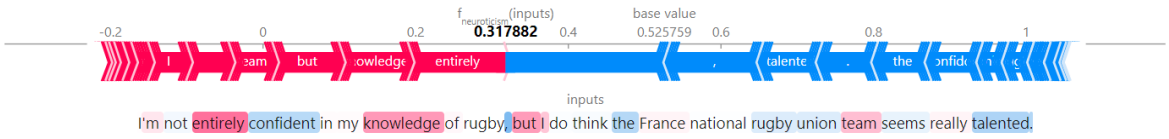


Figure 83: SHAP visualisation for Neuroticism (2/5)

Figure 84: LIME visualisation for Neuroticism (2/5)

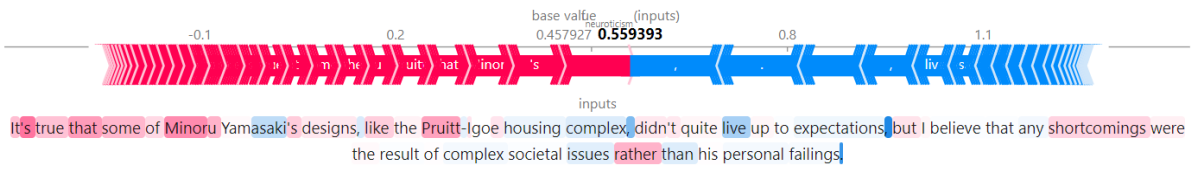


Figure 85: SHAP visualisation for Neuroticism (3/5)

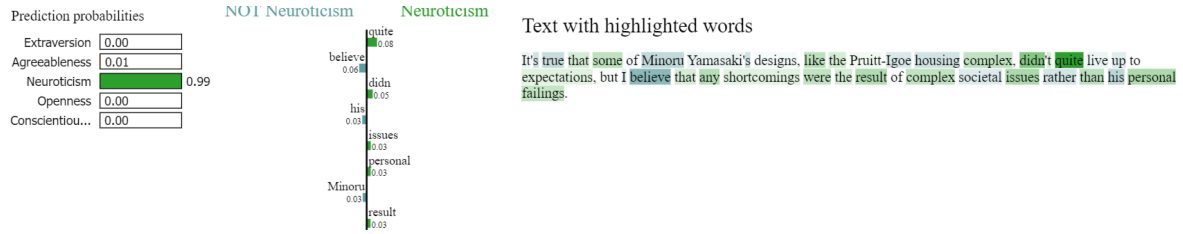


Figure 86: LIME visualisation for Neuroticism (3/5)

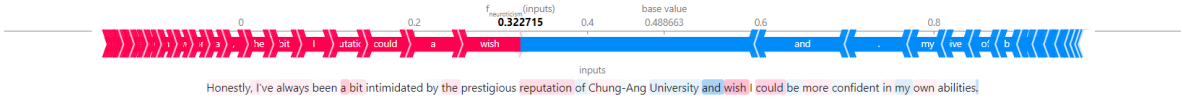


Figure 87: SHAP visualisation for Neuroticism (4/5)

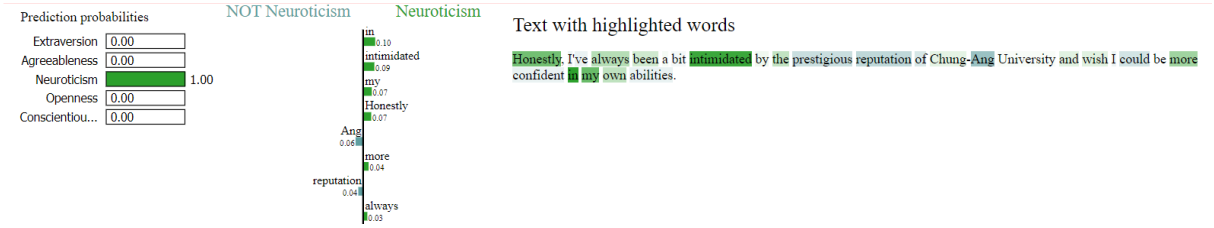


Figure 88: LIME visualisation for Neuroticism (4/5)

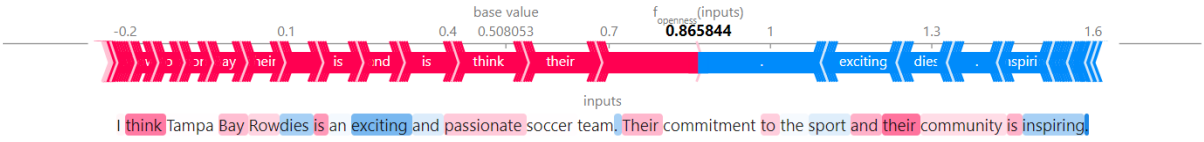


Figure 89: SHAP visualisation for Openness (1/5)

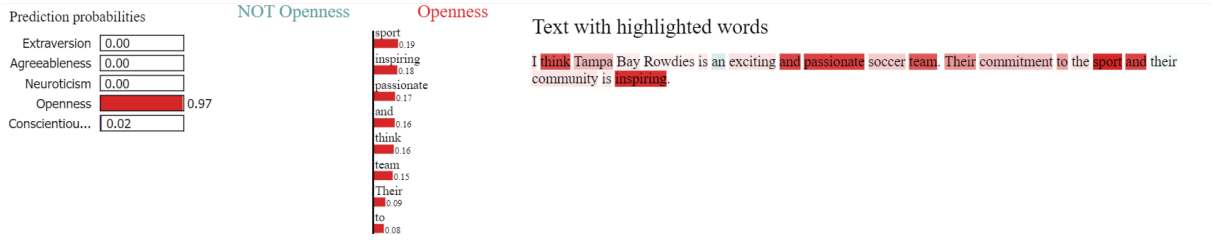


Figure 90: LIME visualisation for Openness (1/5)

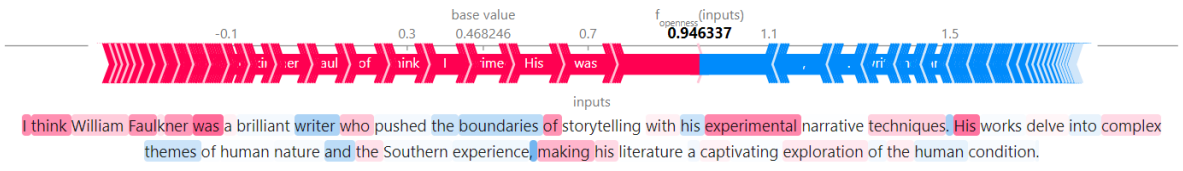


Figure 91: SHAP visualisation for Openness (2/5)

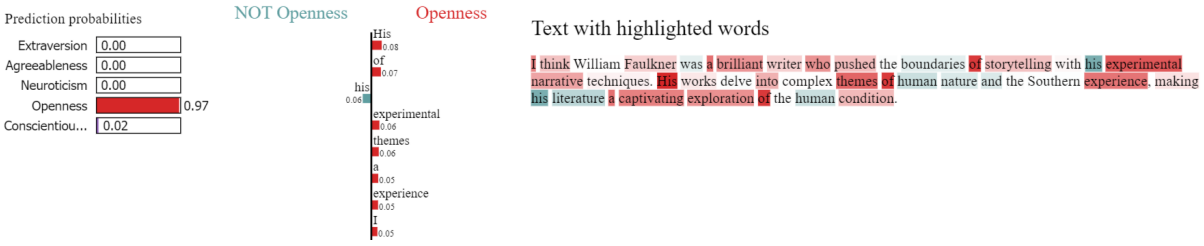


Figure 92: LIME visualisation for Openness (2/5)

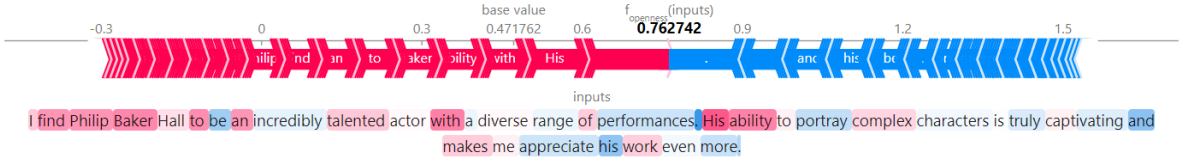


Figure 93: SHAP visualisation for Openness (3/5)

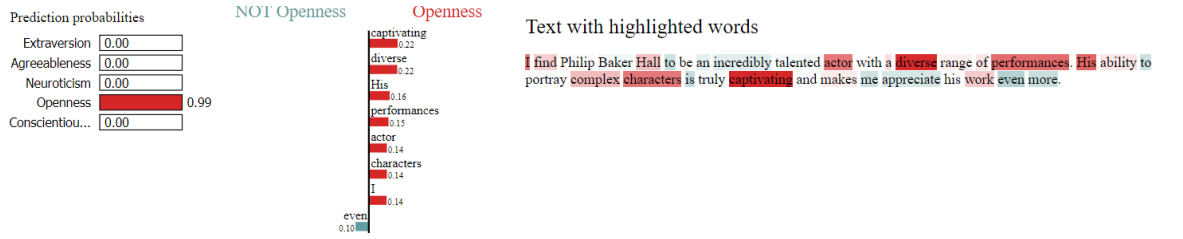


Figure 94: LIME visualisation for Openness (3/5)

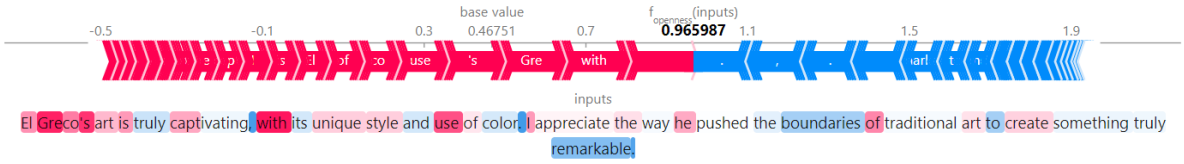


Figure 95: SHAP visualisation for Openness (4/5)

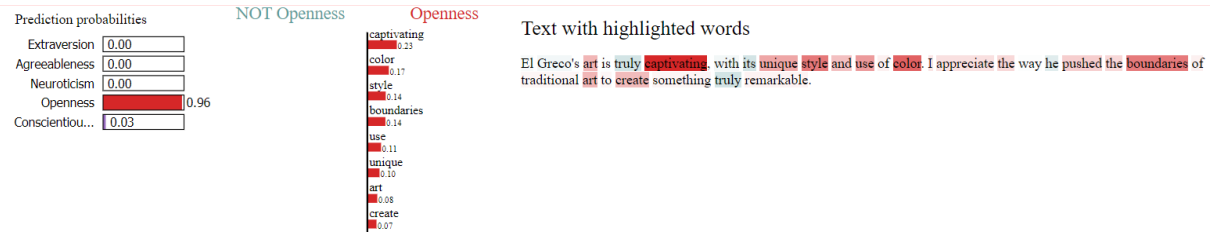


Figure 96: LIME visualisation for Openness (4/5)

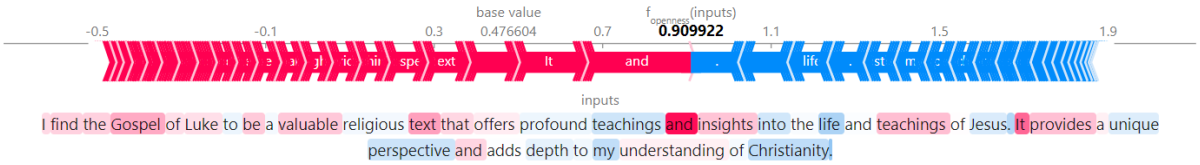


Figure 97: SHAP visualisation for Openness (5/5)

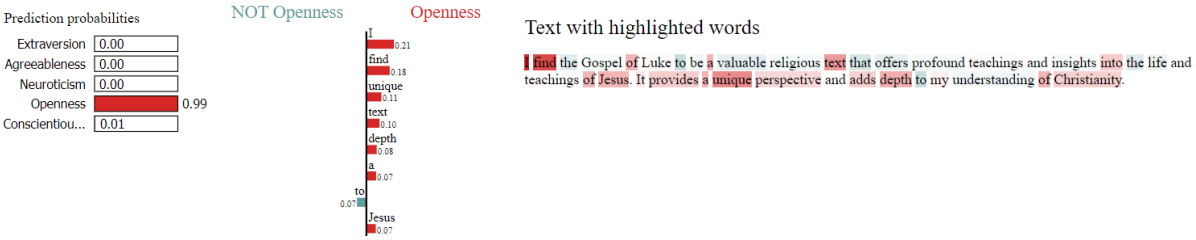


Figure 98: LIME visualisation for Openness (5/5)