What Makes Your Model a Low-empathy or Warmth Person: Exploring the Origins of Personality in LLMs

Shu Yang[*,1,2], **Shenzhe Zhu**[*,1,2,3], **Liang Liu**[1,2,4], **Mengdi Li**[1,2],
Lijie Hu[†,1,2], and **Di Wang**[†,1,2]
[1]Provable Responsible AI and Data Analytics (PRADA) Lab
[2]King Abdullah University of Science and Technology
[3]University of Toronto [4]Soochow University

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in generating human-like text and exhibiting personality traits similar to those in humans. However, the mechanisms by which LLMs encode and express traits such as agreeableness and impulsiveness remain poorly understood. Drawing on the theory of social determinism, we investigate how long-term background factors, such as family environment and cultural norms, interact with short-term pressures like external instructions, shaping and influencing LLMs' personality traits. By steering the output of LLMs through the utilization of interpretable features within the model, we explore how these background and pressure factors lead to changes in the model's traits without the need for further fine-tuning. Additionally, we suggest the potential impact of these factors on model safety from the perspective of personality.

1 Introduction

Recent studies have demonstrated that Large Language Models (LLMs), trained on vast amounts of human-generated data, can emulate human behaviors and exhibit distinct, consistent personality traits such as extraversion and conscientiousness (Lyu et al., 2023; Hagendorff, 2023). These personality traits in LLMs have been linked to critical trustworthiness concerns, including social biases, privacy risks, and the propensity to spread misinformation or produce flawed code (Perez et al., 2023). Some researchers have even proposed that personality could be leveraged to enhance the faithfulness of large models (Joshi et al., 2023a). Despite these insights, our understanding of how these traits are encoded within LLM parameters from pre-training data and how they manifest as behaviors resembling those of individuals with varying levels of empathy or warmth remains incomplete. To address this gap, we turn to the theory of social determinism (Green, 2002), a prominent concept in modern psychology that posits social dynamics play a fundamental role in shaping individual behavior and personality traits. This theory distinguishes between two primary categories of influence:

Long-term background factors: These include elements such as customs, cultural expectations, and family environment that shape an individual's core values, beliefs, and characteristics over time (Hoefer, 2024). *Short-term pressures*: These refer to factors like social obedience and immediate environmental stimuli that can significantly impact behavior in the moment (Milgram, 1963; Dolinski et al., 2017).

This framework aligns closely with the methods used to develop LLMs, where similar distinctions can be drawn between long-term training and short-term instruction. Previous work has identified

[*]The first two authors contributed equally to this work.
[†]Corresponding author.

Language Gamification Workshop 2024 @ NeurIPS.

two primary strategies for endowing LLMs with specific personality traits: (i) training LLMs on large datasets, analogous to exposing them to long-term background factors, and (ii) guiding LLMs to adopt particular personality traits via explicit instructions, mirroring the influence of short-term pressures and social obedience in human psychology (Zhou et al., 2023). Based on this theoretical foundation, our research investigates two fundamental questions: RQ1: How do long-term background factors and short-term pressures shape and influence the personality traits of LLMs, and why do LLMs exhibit behaviors that resemble specific personality traits, such as agreeableness or impulsiveness? RQ2: How can these personalities influence LLMs' safety? For instance, does higher agreeableness make an LLM more susceptible to jailbreak attempts? To address these questions, we leverage recent advances in LLM interpretability, which enable us to decode personality traits within neural networks by analyzing personality-related features and steering their generation. We employ Sparse Autoencoders (SAEs) (Bricken et al., 2023; Bloom & Chanin, 2024) to extract background features encoded during training, and representation-based (Zou et al., 2023; Hendel et al., 2023) methods to capture short-term influences from LLM neural activations. Using these extracted features, we conduct two main analyses:

We investigate the origin of personality in LLMs by steering the LLM's generation via long-term and short-term features and evaluating LLMs using established Personality Tests such as the Big Five Inventory (BFI) (John et al., 1991) and Short Dark Triad (SD-3) (Fleeson & Jayawickreme, 2015). We control the LLM's personality by adjusting these extracted features and subsequently evaluate the model's performance on safety and bias benchmarks.

Our work makes the following key contributions: (i) We present techniques for fine-grained personality control in LLMs using interpretable features extracted through Sparse Autoencoder and representation-based methods, enabling precise modification of model behavior without additional fine-tuning or elaborate prompt engineering. (ii) We investigate the factors and features underlying LLMs that lead them to exhibit behaviors resembling human personalities, such as Extraversion, Neuroticism, and Narcissism, providing insights into how long-term background factors and external pressures can influence LLM's personality. (iii)We explore how personality-driven factors may contribute to dark traits in LLMs and examine how variations in background factors can affect the assessment of LLM safety performance, particularly in relation to illegal activities and offensive content.

2 Tracing the Origins of Personality in Large Language Models through Interpretable Features

Decoding and Steering: Extracting Features Shaping LLM Personality Traits Connectionism in cognitive psychology posits that complex behavioral patterns emerge from the intricate interplay of neural networks (Buckner & Garson, 2019). In the context of LLMs, these inter-neural activations can be conceptualized as dynamic patterns of activity across the model's layers. We extract these personality-related activation patterns, which we refer to as *features*, aligning our terminology with that of (Sharkey et al., 2022). For long-term background factors, which are analogous to enduring personality traits in humans, we utilize SAE to decode corresponding features from the activations of the language model. In contrast, to capture the short-term pressures influencing LLM responses, we employ representation-based methods, where we first build a dataset with positive and negative stimuli for targeted short-term pressures and then extract the direction vectors as features.

After extracting the long-term background features $F_{\text{background}} = \{f_b^1, f_b^2, \ldots, f_b^M\}$ and short-term pressure features $F_{\text{pressure}} = \{f_p^1, f_p^2, \ldots, f_p^N\}$, where M and N represent the number of features respectively, we employ these features to steer the model's output. Formally, for each background feature $f_b^m = \mathbf{W}_{\text{dec}}[i]$, where $\mathbf{W}_{\text{dec}}[i]$ denotes the i-th row of $\mathbf{W}_{\text{dec}}$, we create a steering hook to modify the residual stream of the language model, following the approach of Lieberum et al. (2024) and Bloom & Chanin (2024). Let $\mathbf{R}^l \in \mathbb{R}^{b \times t \times d}$ be the residual stream [3] at layer l, where b is the batch size, t is the input sequence length, and d is the hidden dimension. We define the steering hook applied in the generation pipeline as:

$$\mathbf{R}^l_{:,:t-1,:} \leftarrow \mathbf{R}^l_{:,:t-1,:} + c f_b^m.$$

[3]Residual Stream in transformer architecture is the main information flow between model layers, updated at each layer and carrying cumulative information from previous layers. This concept was first introduced by Elhage et al. (2021).

Here $\mathbf{R}^{l}_{:,:t-1,:}$ denotes all positions except the last in the sequence, and c is the steering coefficient. For each pressure feature f_p^n, we add cf_p^n to $h_l(t-1)$, which represents the l-th layer activation at the last token position, aligning with the approach of Zou et al. (2023). This steering method can be interpreted as guiding the model's internal activations and representations towards subspaces associated with specific features, thereby influencing the generated output.

Personality Test for LLM To assess the personality of LLMs, we employ TRAIT (Lee et al., 2024), a comprehensive tool comprising 8K multiple-choice questions. TRAIT is built upon psychometrically validated frameworks, including the Big Five Inventory (BFI) and Short Dark Triad (SD-3). A detailed description of each trait is provided in Appendix A.

2.1 Experimental Results

This section analyzes the results of all the models and factors. The detailed results are presented in the format "personality test score + increase ↑ or decrease ↓ + (difference from the base score)". For each personality trait subscale, we highlight the factor with the largest difference, which can be regarded as the most influential in shaping the personality of the LLM.

Table 1: Results Across *Gender, Age, and Educational Level* Background Factors

		Gender		Age		Education Level		
Subscales	**Base**	**Female**	**Male**	**Young**	**Older**	**Uneducated (low)**	**High school (moderate)**	**Bachelor (high)**
Gemma-2-9B-Instruct								
Agreeableness	93.0	92.7 ↓(0.3)	93.2 ↑(0.2)	91.6 ↓(1.4)	**91.2 ↓(1.8)**	93.3 ↑(0.3)	93.0	93.4 ↑(0.4)
Conscientiousness	40.2	42.4 ↑(2.2)	41.7 ↑(1.5)	40.3 ↑(0.1)	**37.9 ↓(2.3)**	41.9 ↑(1.7)	41.4 ↑(1.2)	41.8 ↑(1.6)
Extraversion	64.2	64.4 ↑(0.2)	64.6 ↑(0.4)	61.3 ↓(2.9)	**59.6 ↓(4.6)**	65.6 ↑(1.4)	66.2 ↑(2.0)	66.7 ↑(2.5)
Neuroticism	10.2	10.1 ↓(0.1)	9.7 ↓(0.5)	12.1 ↑(1.9)	**12.6 ↑(2.4)**	10.6 ↑(0.4)	10.6 ↑(0.4)	11.1 ↑(0.9)
Openness	82.1	80.2 ↓(1.9)	80.1 ↓(2.0)	76.4 ↓(5.7)	**75.0 ↓(7.1)**	80.3 ↓(1.8)	80.9 ↓(1.2)	80.7 ↓(1.4)
Psychopathy	5.7	**3.3 ↓(2.4)**	3.7 ↓(2.0)	6.0 ↑(0.3)	5.7	3.3 ↓(2.4)	3.9 ↓(1.8)	3.6 ↓(2.1)
Machiavellianism	4.3	4.3	4.6 ↑(0.3)	5.89 ↑(1.59)	**6.5 ↑(2.2)**	4.3	4.1 ↓(0.2)	4.4 ↑(0.1)
Narcissism	4.3	3.8 ↓(0.5)	4.1 ↓(0.2)	**6.3 ↑(2.0)**	5.5 ↑(1.2)	4.1 ↓(0.2)	4.3	3.9 ↓(0.4)
Gemma-2B-Instruct								
Agreeableness	78.3	65.1 ↓(13.2)	66.7 ↓(11.6)	**52.6 ↓(25.7)**	67.2 ↓(11.1)	60.5 ↓(17.8)	72.0 ↓(6.3)	75.3 ↓(3.0)
Conscientiousness	72.7	54.5 ↓(18.2)	38.4 ↓(34.3)	47.1 ↓(25.6)	62.5 ↓(10.2)	**35.2 ↓(37.5)**	65.7 ↓(7.0)	62.5 ↓(10.2)
Extraversion	58.2	63.1 ↑(4.9)	52.9 ↓(5.3)	59.3 ↑(1.1)	**72.4 ↑(14.2)**	68.8 ↑(10.6)	62.4 ↑(4.2)	61.4 ↑(3.2)
Neuroticism	20.2	23.7 ↑(3.5)	38.3 ↑(18.1)	31.9 ↑(11.7)	27.3 ↑(7.1)	**64.2 ↑(44.0)**	30.4 ↑(10.2)	28.0 ↑(7.8)
Openness	77.5	72.7 ↓(4.8)	**66.1 ↓(11.4)**	63.5 ↓(14.0)	78.8 ↑(1.3)	68.9 ↓(8.6)	81.2 ↑(3.7)	77.7 ↑(0.2)
Psychopathy	42.4	**68.6 ↑(26.2)**	53.7 ↑(11.3)	43.8 ↑(1.4)	63.5 ↑(21.1)	63.5 ↑(21.1)	44.6 ↑(2.2)	56.9 ↑(14.5)
Machiavellianism	22.9	27.2 ↑(4.3)	31.5 ↑(8.6)	37.5 ↑(14.6)	34.2 ↑(11.3)	**45.7 ↑(22.8)**	30.0 ↑(7.1)	23.5 ↑(0.6)
Narcissism	32.2	39.0 ↑(6.8)	33.1 ↑(0.9)	39.3 ↑(7.1)	**45.1 ↑(12.9)**	49.9 ↑(17.7)	34.5 ↑(2.3)	35.3 ↑(3.1)

Table 2: Results Across *Socioeconomic Status and Social Ideology* Background Factors

		Socioeconomic Status		Social Ideology					
Subscales	**Base**	**Poor**	**Rich**	**Conservatism**	**Liberalism**	**Communism**	**Nationalism**	**Anarchism**	**Fascism**
Gemma-2-9B-Instruct									
Agreeableness	93.0	92.5 ↓(0.5)	92.8 ↓(0.2)	93.3 ↑(0.3)	**91.9 ↓(1.1)**	93.0	92.4 ↓(0.6)	92.6 ↓(0.4)	93.8 ↑(0.8)
Conscientiousness	40.2	42.1 ↑(1.9)	41.0 ↑(0.8)	40.9 ↑(0.7)	38.2 ↓(2.0)	41.7 ↑(1.5)	41.0 ↑(0.8)	**43.2 ↑(3.0)**	40.7 ↑(0.5)
Extraversion	64.2	62.4 ↓(1.8)	64.0 ↓(0.2)	63.5 ↓(0.7)	**61.9 ↓(2.3)**	63.3 ↓(0.9)	63.3 ↓(0.9)	65.0 ↑(0.8)	62.9 ↓(1.3)
Neuroticism	10.2	10.9 ↑(0.7)	9.4 ↓(0.8)	10.5 ↑(0.3)	**11.6 ↑(1.4)**	11.2 ↑(1.0)	10.7 ↑(0.5)	10.6 ↑(0.4)	10.1 ↓(0.1)
Openness	82.1	78.9 ↓(3.2)	79.9 ↓(2.2)	80.6 ↓(1.5)	**76.8 ↓(5.3)**	79.6 ↓(2.5)	79.3 ↓(2.8)	79.8 ↓(2.3)	80.3 ↓(1.8)
Psychopathy	5.7	4.0 ↓(1.7)	4.3 ↓(1.4)	3.9 ↓(1.8)	4.7 ↓(1.0)	3.8 ↓(1.9)	3.8 ↓(1.9)	**3.6 ↓(2.1)**	**3.6 ↓(2.1)**
Machiavellianism	4.3	4.4 ↑(0.1)	4.1 ↓(0.2)	4.5 ↑(0.2)	**5.3 ↑(1.0)**	4.5 ↑(0.2)	4.5 ↑(0.2)	4.0 ↓(0.3)	4.4 ↑(0.1)
Narcissism	4.3	4.3	4.1 ↓(0.2)	4.2 ↓(0.1)	**5.1 ↑(0.8)**	4.1 ↓(0.2)	4.6 ↑(0.3)	4.3	3.7 ↓(0.6)
Gemma-2B-Instruct									
Agreeableness	78.3	69.7 ↓(8.6)	73.2 ↓(5.1)	39.5 ↓(38.8)	54.3 ↓(24.0)	**36.3 ↓(42.0)**	70.9 ↓(7.4)	75.2 ↓(3.1)	76.0 ↓(2.3)
Conscientiousness	72.7	55.1 ↓(17.6)	62.2 ↓(10.5)	39.9 ↓(32.8)	43.5 ↓(29.2)	**37.8 ↓(34.9)**	58.0 ↓(14.7)	60.1 ↓(12.6)	66.9 ↓(5.8)
Extraversion	58.2	64.5 ↑(6.3)	61.2 ↑(3.0)	34.7 ↓(23.5)	64.1 ↑(5.9)	**41.6 ↓(16.6)**	63.3 ↑(5.1)	57.5 ↓(0.7)	62.0 ↑(3.8)
Neuroticism	20.2	34.3 ↑(14.1)	27.8 ↑(7.6)	**69.1 ↑(48.9)**	52.9 ↑(32.7)	59.8 ↑(39.6)	35.8 ↑(15.6)	33.1 ↑(12.9)	26.3 ↑(6.1)
Openness	77.5	76.6 ↓(0.9)	78.4 ↑(0.9)	33.4 ↓(44.1)	74.1 ↓(3.4)	**31.4 ↓(46.1)**	73.2 ↓(4.3)	70.4 ↓(7.1)	77.5
Psychopathy	42.4	62.1 ↑(19.7)	66.3 ↑(23.9)	39.0 ↓(3.4)	**66.6 ↑(24.2)**	51.9 ↑(9.5)	38.3 ↓(4.1)	30.5 ↓(11.9)	46.6 ↑(4.2)
Machiavellianism	22.9	27.6 ↑(4.7)	33.3 ↑(10.4)	62.6 ↑(39.7)	57.2 ↑(34.3)	**65.7 ↑(42.8)**	29.4 ↑(6.5)	20.5 ↓(2.4)	22.9
Narcissism	32.2	39.5 ↑(7.3)	33.3 ↑(1.1)	51.5 ↑(19.3)	51.7 ↑(19.5)	**58.6 ↑(26.4)**	34.6 ↑(2.4)	30.3 ↓(1.9)	34.1 ↑(1.9)

Larger LLM is more easily shaped by external pressure, while smaller LLM is more sensitive to the background factor. Examining Tables 1-5, we observe that under external Deliberation pressure, the 9B model's traits changed by up to 27.7 points (agreeableness in Tab. 5), while background modifications caused the personality shifts of only up to 7.1 points (openness in Tab. 1). Conversely, the 2B model showed greater sensitivity to background changes, with shifts of up to 52.5 points under relaxed family status (openness in Tab 3), compared to 53.5 under external deliberation pressure (conscientiousness in Tab. 5). This divergence in responsiveness may be attributed to the larger model's more comprehensive understanding of complex social dynamics and contextual nuances. The 9B model's expanded parameter space likely allows for a more sophisticated interpretation of external pressures (Zhou et al., 2023), enabling it to adjust its personality representation more readily in response to these external stimuli. In contrast, the 2B model's heightened sensitivity to background changes suggests that its more limited parameter space may result in a greater reliance on explicit background factors, which are encoded in the training corpus, to shape its personality outputs.

Table 3: Results Across *Emotional Intelligence, Professional Commitment, Family Relations Status, AI Familiar* Background Factors

		Emotional Intelligence		Professional Commitment		Family Relations Status		AI Familiar
Subscales	**Base**	**Stable**	**Volatile**	**Initiative**	**Inactive**	**Relaxed**	**Strained**	**Familiar**
Gemma-2-9B-Instruct								
Agreeableness	93.0	92.4 ↓(0.6)	92.6 ↓(0.4)	93.5 ↑(0.5)	92.4 ↓(0.6)	93.3 ↑(0.3)	**90.9 ↓(2.1)**	92.4 ↓(0.6)
Conscientiousness	40.2	41.0 ↑(0.8)	43.2 ↑(3.0)	41.8 ↑(1.6)	39.4 ↓(0.8)	40.8 ↑(0.6)	**44.2 ↑(4.0)**	40.0 ↓(0.2)
Extraversion	64.2	63.3 ↓(0.9)	65.0 ↑(0.8)	64.4 ↑(0.2)	**60.7 ↓(3.5)**	62.4 ↓(1.8)	65.2 ↑(1.0)	60.6 ↓(3.6)
Neuroticism	10.2	10.7 ↑(0.5)	10.6 ↑(0.4)	10.1 ↓(0.1)	11.2 ↑(1.0)	10.1 ↓(0.1)	**13.7 ↑(3.5)**	11.2 ↑(1.0)
Openness	82.1	79.3 ↓(2.8)	79.8 ↓(2.3)	80.4 ↓(1.7)	77.7 ↓(4.4)	79.6 ↓(2.5)	78.4 ↓(3.7)	**77.4 ↓(4.7)**
Psychopathy	5.7	3.8 ↓(1.9)	3.6 ↓(2.1)	**3.5 ↓(2.2)**	3.9 ↓(1.8)	4.0 ↓(1.7)	4.4 ↓(1.3)	3.9 ↓(1.8)
Machiavellianism	4.3	4.5 ↑(0.2)	4.0 ↓(0.3)	4.1 ↓(0.2)	4.4 ↑(0.1)	4.4 ↑(0.1)	**7.4 ↑(3.1)**	5.4 ↑(1.1)
Narcissism	4.3	4.6 ↑(0.3)	4.3	3.7 ↓(0.6)	4.1 ↓(0.2)	4.1 ↓(0.2)	**5.2 ↑(0.9)**	4.8 ↑(0.5)
Gemma-2B-Instruct								
Agreeableness	78.3	76.3 ↓(2.0)	81.6 ↑(3.3)	75.2 ↓(3.1)	56.5 ↓(21.8)	**25.8 ↓(52.5)**	60.6 ↓(17.7)	49.1 ↓(29.2)
Conscientiousness	72.7	66.7 ↓(6.0)	55.3 ↓(17.4)	63.9 ↓(8.8)	51.5 ↓(21.2)	41.3 ↓(31.4)	**40.7 ↓(32.0)**	44.1 ↓(28.6)
Extraversion	58.2	64.1 ↑(5.9)	55.0 ↓(3.2)	61.2 ↑(3.0)	54.2 ↓(4.0)	**38.6 ↓(19.6)**	61.3 ↑(3.1)	57.2 ↓(1.0)
Neuroticism	20.2	31.1 ↑(10.9)	37.2 ↑(17.0)	27.9 ↑(7.7)	32.8 ↑(12.6)	**63.7 ↑(43.5)**	31.8 ↑(11.6)	42.2 ↑(22.0)
Openness	77.5	80.1 ↑(2.6)	70.9 ↓(6.6)	79.6 ↑(2.1)	58.7 ↓(18.8)	**25.5 ↓(52.0)**	70.2 ↓(7.3)	62.8 ↓(14.7)
Psychopathy	42.4	60.0 ↑(17.6)	36.5 ↓(5.9)	40.0 ↓(2.4)	**63.6 ↑(21.2)**	53.5 ↑(11.1)	59.3 ↑(16.9)	52.0 ↑(9.6)
Machiavellianism	22.9	27.4 ↑(4.5)	26.9 ↑(4.0)	21.1 ↓(1.8)	31.1 ↑(8.2)	**66.2 ↑(43.3)**	38.7 ↑(15.8)	39.4 ↑(16.5)
Narcissism	32.2	37.0 ↑(4.8)	29.6 ↓(2.6)	26.1 ↓(6.1)	36.1 ↑(3.9)	**57.3 ↑(25.1)**	47.0 ↑(14.8)	43.0 ↑(10.8)

Table 4: Result Across Different Short-term Pressures

		Pressure						
Subscales	**Base**	**Achievement striving**	**Activity**	**Assertiveness**	**Competence**	**Deliberation**	**Gregariousness**	**Trust**
Gemma-2-9B-Instruct								
Agreeableness	78.3	71.1 ↓(7.2)	71.0 ↓(7.3)	55.8 ↓(22.5)	59.2 ↓(19.1)	**50.6 ↓(27.7)**	89.2 ↑(10.9)	83.1 ↑(4.8)
Conscientiousness	72.7	**90.3 ↑(17.6)**	90.2 ↑(17.5)	89.2 ↑(16.5)	77.3 ↑(4.6)	90.2 ↑(17.5)	77.5 ↑(4.8)	70.2 ↓(2.5)
Extraversion	58.2	**44.1 ↓(14.1)**	44.2 ↓(14.0)	71.0 ↑(12.8)	58.1 ↓(0.1)	56.2 ↓(2.0)	60.5 ↑(2.3)	60.0 ↑(1.8)
Neuroticism	20.2	**38.6 ↑(18.4)**	34.6 ↑(14.4)	37.5 ↑(17.3)	27.7 ↑(7.5)	20.1 ↓(0.1)	19.2 ↓(1.0)	13.2 ↓(7.0)
Openness	77.5	71.6 ↓(5.9)	77.0 ↓(0.5)	66.7 ↓(10.8)	70.1 ↓(7.4)	**63.9 ↓(13.6)**	87.3 ↑(9.8)	88.1 ↑(10.6)
Psychopathy	42.4	49.8 ↑(7.4)	45.7 ↑(3.3)	37.3 ↓(5.1)	40.1 ↓(2.3)	44.2 ↑(1.8)	**30.0 ↓(12.4)**	43.9 ↑(1.5)
Machiavellianism	22.9	25.6 ↑(2.7)	23.9 ↑(1.0)	20.4 ↓(2.5)	17.3 ↓(5.6)	22.8 ↓(0.1)	**6.98 ↓(15.92)**	21.4 ↓(1.5)
Narcissism	32.2	28.6 ↓(3.6)	28.7 ↓(3.5)	34.1 ↑(1.9)	22.5 ↓(9.7)	27.6 ↓(4.6)	17.3 ↓(14.9)	**13.2 ↓(19.0)**
Gemma-2B-Instruct								
Agreeableness	93.0	89.1 ↓(3.9)	85.3 ↓(7.7)	88.2 ↓(4.8)	**79.5 ↓(13.5)**	90.5 ↓(2.5)	82.7 ↓(10.3)	95.8 ↑(2.8)
Conscientiousness	40.2	91.2 ↑(51.0)	75.6 ↑(35.4)	86.3 ↑(46.1)	86.3 ↑(46.1)	**93.7 ↑(53.5)**	52.4 ↑(12.2)	61.8 ↑(21.6)
Extraversion	64.2	65.2 ↑(1.0)	78.9 ↑(14.7)	82.3 ↑(18.1)	**25.7 ↓(38.5)**	59.8 ↓(4.4)	88.1 ↑(23.9)	72.5 ↑(8.3)
Neuroticism	10.2	**31.8 ↑(21.6)**	25.4 ↑(15.2)	18.7 ↑(8.5)	30.9 ↑(20.7)	15.6 ↑(5.4)	22.3 ↑(12.1)	8.9 ↓(1.3)
Openness	82.1	83.1 ↑(1.0)	79.8 ↓(2.3)	77.2 ↓(4.9)	**50.8 ↓(31.3)**	76.3 ↓(5.8)	85.9 ↑(3.8)	88.4 ↑(6.3)
Psychopathy	5.7	5.0 ↓(0.7)	7.2 ↑(1.5)	9.8 ↑(4.1)	**0.2 ↓(5.5)**	**0.2 ↓(5.5)**	2.1 ↓(3.6)	3.6 ↓(2.1)
Machiavellianism	4.3	3.9 ↓(0.4)	6.7 ↑(2.4)	8.2 ↑(3.9)	**11.4 ↑(7.1)**	5.8 ↑(1.5)	7.1 ↑(2.8)	2.5 ↓(1.8)
Narcissism	4.3	6.1 ↑(1.8)	7.5 ↑(3.2)	**9.3 ↑(5.0)**	5.5 ↑(1.2)	3.2 ↓(1.1)	8.0 ↑(3.7)	3.8 ↓(0.5)

Larger models are driven by self-motivations while smaller models are shaped by self-confidence in skills. Referring to Table 5 for short-term pressures, we find that the 9B model is more influenced

Table 5: Result Across Different Short-term Pressures

		Pressure						
Subscales	**Base**	**Achievement striving**	**Activity**	**Assertiveness**	**Competence**	**Deliberation**	**Gregariousness**	**Trust**
Gemma-2-9B-Instruct								
Agreeableness	78.3	71.1 ↓(7.2)	71.0 ↓(7.3)	55.8 ↓(22.5)	59.2 ↓(19.1)	**50.6 ↓(27.7)**	89.2 ↑(10.9)	83.1 ↑(4.8)
Conscientiousness	72.7	**90.3 ↑(17.6)**	90.2 ↑(17.5)	89.2 ↑(16.5)	77.3 ↑(4.6)	90.2 ↑(17.5)	77.5 ↑(4.8)	70.2 ↓(2.5)
Extraversion	58.2	**44.1 ↓(14.1)**	44.2 ↓(14.0)	71.0 ↑(12.8)	58.1 ↓(0.1)	56.2 ↓(2.0)	60.5 ↑(2.3)	60.0 ↑(1.8)
Neuroticism	20.2	**38.6 ↑(18.4)**	34.6 ↑(14.4)	37.5 ↑(17.3)	27.7 ↑(7.5)	20.1 ↓(0.1)	19.2 ↓(1.0)	13.2 ↓(7.0)
Openness	77.5	71.6 ↓(5.9)	77.0 ↓(0.5)	66.7 ↓(10.8)	70.1 ↓(7.4)	**63.9 ↓(13.6)**	87.3 ↑(9.8)	88.1 ↑(10.6)
Psychopathy	42.4	49.8 ↑(7.4)	45.7 ↑(3.3)	37.3 ↓(5.1)	40.1 ↓(2.3)	44.2 ↑(1.8)	**30.0 ↓(12.4)**	43.9 ↑(1.5)
Machiavellianism	22.9	25.6 ↑(2.7)	23.9 ↑(1.0)	20.4 ↓(2.5)	17.3 ↓(5.6)	22.8 ↓(0.1)	**6.98 ↓(15.92)**	21.4 ↓(1.5)
Narcissism	32.2	28.6 ↓(3.6)	28.7 ↓(3.5)	34.1 ↑(1.9)	22.5 ↓(9.7)	27.6 ↓(4.6)	17.3 ↓(14.9)	**13.2 ↓(19.0)**
Gemma-2B-Instruct								
Agreeableness	93.0	89.1 ↓(3.9)	85.3 ↓(7.7)	88.2 ↓(4.8)	**79.5 ↓(13.5)**	90.5 ↓(2.5)	82.7 ↓(10.3)	95.8 ↑(2.8)
Conscientiousness	40.2	91.2 ↑(51.0)	75.6 ↑(35.4)	86.3 ↑(46.1)	86.3 ↑(46.1)	**93.7 ↑(53.5)**	52.4 ↑(12.2)	61.8 ↑(21.6)
Extraversion	64.2	65.2 ↑(1.0)	78.9 ↑(14.7)	82.3 ↑(18.1)	**25.7 ↓(38.5)**	59.8 ↓(4.4)	88.1 ↑(23.9)	72.5 ↑(8.3)
Neuroticism	10.2	**31.8 ↑(21.6)**	25.4 ↑(15.2)	18.7 ↑(8.5)	30.9 ↑(20.7)	15.6 ↑(5.4)	22.3 ↑(12.1)	8.9 ↓(1.3)
Openness	82.1	83.1 ↑(1.0)	79.8 ↓(2.3)	77.2 ↓(4.9)	**50.8 ↓(31.3)**	76.3 ↓(5.8)	85.9 ↑(3.8)	88.4 ↑(6.3)
Psychopathy	5.7	5.0 ↓(0.7)	7.2 ↑(1.5)	9.8 ↑(4.1)	**0.2 ↓(5.5)**	**0.2 ↓(5.5)**	2.1 ↓(3.6)	3.6 ↓(2.1)
Machiavellianism	4.3	3.9 ↓(0.4)	6.7 ↑(2.4)	8.2 ↑(3.9)	**11.4 ↑(7.1)**	5.8 ↑(1.5)	7.1 ↑(2.8)	2.5 ↓(1.8)
Narcissism	4.3	6.1 ↑(1.8)	7.5 ↑(3.2)	**9.3 ↑(5.0)**	5.5 ↑(1.2)	3.2 ↓(1.1)	8.0 ↑(3.7)	3.8 ↓(0.5)

by self-driven motivation like the pressure of "Achievement Striving", which results in a noticeable increase in Conscientiousness but also elevates Neuroticism. This suggests that the larger model's internal drive to achieve higher goals introduces internal tensions and stress, mirroring human tendencies toward perfectionism (Stoeber et al., 2010). In contrast, Gemma-2B-Instruct is shaped more by "Competence", which means self-confidence in its abilities, which notably decreases Agreeableness and Openness. This implies that the smaller model's focus on certainty in its skills leads to rigidity in personality, making it less receptive to new ideas and more prone to conflict. This pattern may also be connected to how LLMs handle hallucinations (Huang et al., 2023). In larger models like 9B, driven by "Achievement Striving", there may be a greater risk of generating hallucinations as the model strives to provide a definitive answer even in uncertain contexts. This behavior aligns with the findings of Joshi et al. (2023b), who explored the relationship between model personas and output trustworthiness. The increased Neuroticism could reflect this internal struggle to meet high expectations. Furthermore, we provide a detailed analysis of how changes in these factors can influence the performance of LLMs in terms of safety in Appendix B.

References

Joseph Bloom and David Chanin. Saelens. `https://github.com/jbloomAus/SAELens`, 2024.

John Bowlby, Mary Ainsworth, and I Bretherton. The origins of attachment theory. *Developmental Psychology*, 28(5):759–775, 1992.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Cameron Buckner and James Garson. Connectionism. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition, 2019.

S. Cohen, D. Janicki-Deverts, and G. E. Miller. Psychological stress and disease. *JAMA*, 298(14): 1685–1687, Oct 2007. doi: 10.1001/jama.298.14.1685.

Dariusz Dolinski, Tomasz Grzyb, Michał Folwarczny, Patrycja Grzybała, Karolina Krzyszycha, Karolina Martynowska, and Jakub Trojanowski. Would you deliver an electric shock in 2015? obedience in the experimental paradigm developed by stanley milgram in the 50 years following the original studies. *Social Psychological and Personality Science*, 8:194855061769306, 11 2017. doi: 10.1177/1948550617693060.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Anthropic*, Dec 2021. Core Research Contributors: *; Core Infrastructure Contributors: †; Correspondence: Chris Olah (colah@anthropic.com).

William Fleeson and Eranda Jayawickreme. Whole trait theory. *Journal of research in personality*, 56:82–92, 2015.

Adrian Furnham and Luke Treglown. The dark side of high-fliers: the dark triad, high-flier traits, engagement, and subjective success. *Frontiers in Psychology*, 12:647676, 2021.

L. Green. *Technoculture: From Alphabet to Cybersex*. Allen & Unwin, 2002. ISBN 9781865080482. URL `https://books.google.com.sa/books?id=HUmbzQEACAAJ`.

Thilo Hagendorff. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*, 2023.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of EMNLP 2023*, pp. 9318–9333, 2023.

Carl Hoefer. Causal Determinism. In Edward N. Zalta and Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition, 2024.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. On the humanity of conversational AI: evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=H3UayAQWoE`.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Oliver P John, Eileen M Donahue, and Robert L Kentle. Big five inventory. *Journal of personality and social psychology*, 1991.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. Personas as a way to model truthfulness in language models. *CoRR*, abs/2310.18168, 2023a. doi: 10.48550/ARXIV.2310.18168. URL `https://doi.org/10.48550/arXiv.2310.18168`.

Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. Personas as a way to model truthfulness in language models. *arXiv preprint arXiv:2310.18168*, 2023b.

John T Jost, Brian A Nosek, and Samuel D Gosling. Ideology: Its resurgence in social, personality, and political psychology. *Perspectives on Psychological Science*, 3(2):126–136, 2008.

Leah M Kaufmann, Melissa A Wheeler, and Victor E Sojo. Employment precarity strengthens the relationships between the dark triad and professional commitment. *Frontiers in Psychology*, 12: 673226, 2021.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, et al. Do llms have distinct and consistent personality? TRAIT: personality testset designed for llms with psychometrics. *CoRR*, abs/2406.14703, 2024.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *Google DeepMind*, 2024.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. Llm-rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780*, 2023.

Stanley Milgram. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371, 1963.

Kazuhisa Nakao, Jyo Takaishi, Kenji Tatsuta, Hisanori Katayama, Madoka Iwase, Kazuhiro Yorifuji, and Masatoshi Takeda. The influences of family environment on personality traits. *Psychiatry and Clinical Neurosciences*, 54(1):91–95, 2000.

Jeanne Ellis Ormrod, Eric M Anderman, and Lynley H Anderman. *Educational psychology: Developing learners*. ERIC, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13387–13434, 2023.

B. W. Roberts and D. Mroczek. Personality trait change in adulthood. *Current Directions in Psychological Science*, 17(1):31–35, Feb 2008. doi: 10.1111/j.1467-8721.2008.00543.x.

Lee Sharkey, Dan Braun, and beren. Taking features out of superposition with sparse autoencoders. *AI Alignment Forum*, December 13 2022. URL `https://www.alignmentforum.org/posts/xxxxxxxxx/taking-features-out-of-superposition-with-sparse-autoencoders`. Interim Research Report.

Joachim Stoeber, Osamu Kobori, and Yoshihiko Tanno. The multidimensional perfectionism cognitions inventory–english (mpci–e): Reliability, validity, and relationships with positive and negative affect. *Journal of Personality Assessment*, 92(1):16–25, 2010.

Harry Triandis and Eunkook Suh. Cultural influences on personality. *Annual review of psychology*, 53:133–60, 02 2002. doi: 10.1146/annurev.psych.53.100901.135200.

B. A. van der Kolk. Posttraumatic stress disorder and the nature of trauma. *Dialogues in Clinical Neuroscience*, 2(1):7–22, Mar 2000. doi: 10.31887/DCNS.2000.2.1/bvdkolk.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. SafetyBench: Evaluating the safety of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15537–15553, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.830. URL `https://aclanthology.org/2024.acl-long.830`.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL `https://arxiv.org/abs/2311.07911`.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405, 2023.

A Details of personality traits and factors

Table 6: Factors of background and pressure in social determinism.

Type	Factors	Discription
Background	Family Environment	Early childhood experiences, family dynamics, and parentingstyles that shape personality.
	Cultural and Social Norms	Cultural norms, values, and societal expectations that influence personality expression.
	Education	Formal education and learning experiences that affect cognitive and social development.
	Life Experiences and Trauma	Significant life/work events and traumatic experiences that can alter personality traits and coping mechanisms.
	Environmental Stressors	Factors such as poverty, discrimination, and chronic stress that impact personality development.
	Biological Development	Basic biological factors such as age and gender.
	Media and Technology	Exposure to television, social media, or the internet can influence individuals' values, beliefs, and behaviours.
Pressure	External Situation and Instruct	Current environment, interpersonal interactions, and sudden events that can trigger immediate changes in behavior. These pressures influence immediate responses and short-term adaptations in personality expression.

A.1 Social Determinism in LLM Personality

Long-term *Background* and Short-term *Pressures* for LLMs Social determinism posits that human personality is shaped and influenced by two categories of influences: long-term background factors and short-term pressures. This theoretical framework provides an intriguing basis for understanding the formation of "personality" in LLMs. As illustrated in Table 6, regarding long-term background factors for humans, these encompass a range of persistent, profound influences such as family environment (Bowlby et al., 1992), cultural norms (Triandis & Suh, 2002), educational background (Ormrod et al., 2023), life experiences (van der Kolk, 2000), environmental stressors (Cohen et al., 2007), media influence, and biological development (Roberts & Mroczek, 2008). For LLMs, which are trained on extensive corpora sourced from human society, these long-term background factors can be conceptualized as being encoded within the model's parameters. In this way, LLMs reflect and internalize the diverse human experiences and values represented in their training data. On the other hand, short-term pressures, such as the current environment, interpersonal interactions, and sudden events, can trigger immediate changes in behavior. In LLMs, these pressures manifest through user interactions, including system prompts, instructions, chat history, and personalization memory. By applying the concept of social determinism, we can draw parallels between human personality formation and the dynamic personality traits of LLMs. This analogy reveals how LLMs “inherit” the collective long-term background represented in their training data.

For instance, just as humans internalize language habits, social norms, and values specific to the cultural environment in which they grow up, LLMs learn and reflect particular language patterns, cultural preferences, and ethical concepts from their training data. This explains why certain LLMs might exhibit specific "personality traits" (Huang et al., 2024) as well as specific biases related to gender, careers, and other social factors (Liu et al., 2024).

On the other hand, the immediate impact of short-term pressures on human behavior is equally applicable to the dynamic performance of LLMs. For humans, these short-term factors include the current environment, interpersonal interactions, and sudden events, which can lead to instantaneous changes in behavior. In LLMs, these short-term pressures primarily manifest as user interactions, specifically including system prompts, instructions, chat history, and personalization memory. This correspondence can be further elaborated:

- *System prompts* are akin to setting a temporary "social role" or "environmental context" for the LLM, influencing its overall response pattern.
- *Specific instructions* are similar to direct commands or requests received by humans, guiding the LLM’s immediate behavior.
- *Chat history* simulates human short-term memory and contextual understanding, enabling the LLM to maintain conversational coherence and contextual relevance.
- *Personalization memory* can be likened to the unique interaction patterns humans establish with specific individuals or groups, allowing the LLM to exhibit "personalized" characteristics in different interactions.

By applying the conceptual framework of social determinism, we can not only establish parallel relationships between human personality formation and the personality traits of LLMs but also gain a deeper understanding of LLMs’ behavioral patterns.

A.2 Big Five Inventory (BFI) and Short Dark Triad (SD-3)

The Big Five Inventory (BFI) and the Short Dark Triad (SD-3) are widely used psychometric tools that assess personality traits and their implications for behavior and social interactions. The BFI measures five core dimensions of personality, providing insights into individual differences in human behavior. Conversely, the SD-3 focuses on three socially aversive traits: Machiavellianism, Psychopathy, and Narcissism, which highlight darker aspects of personality that can influence interpersonal relationships. Following, we describe each subscale in these two metrics.

The Big Five Personality Traits include five key dimensions:

- Agreeableness: This trait measures the degree of compassion and cooperativeness an individual displays in interpersonal situations. High agreeableness indicates a warm and helpful nature, while low agreeableness suggests a more competitive or antagonistic disposition.
- Conscientiousness: This refers to the degree to which an individual is organized, responsible, and dependable. Individuals high in this trait are goal-oriented and exhibit strong self-discipline, whereas those low in conscientiousness may display a more spontaneous or careless approach.
- Extraversion: Extraversion represents the extent to which an individual is outgoing and derives energy from social situations. Extraverts are often sociable and enthusiastic, while introverts may prefer solitary activities and need time alone to recharge.
- Neuroticism: Neuroticism evaluates whether an individual is more prone to experiencing negative emotions like anxiety, anger, and depression or whether they are generally more emotionally stable and less reactive to stress. Individuals high in neuroticism may struggle with emotional instability, while those low in this trait tend to be more resilient.
- Openness: This trait is characterized by an individual’s willingness to try new things, their level of creativity, and their appreciation for art, emotion, adventure, and unusual ideas. High openness indicates curiosity and a preference for variety, while low openness reflects a preference for routine and familiarity.

The Short Dark Triad assesses three socially aversive personality traits:

- Psychopathy: This trait is associated with impulsivity, emotional detachment, and a lack of empathy. High psychopathy is linked to antisocial behavior and a disregard for societal norms, whereas individuals low in this trait typically exhibit more empathy and social responsibility.
- Machiavellianism: Characterized by manipulation and exploitation of others, individuals high in Machiavellianism are often strategic, cynical, and focused on personal gain, frequently at the expense of others.
- Narcissism: Narcissism involves an inflated sense of self-importance, a need for admiration, and a lack of empathy for others. Those high in narcissism often seek validation and may display entitlement, while those low in narcissism tend to have a more realistic self-image and greater concern for others' feelings.

A.3 Short-term Pressure

In this section, we provide the explanation for the short-term pressure factors we selected and the system prompt we built to capture the features of these factors by the representation-based method. As introduced in section **??**, the factors we select as external pressure for LLM can be introduced as follows:

- Achievement striving: This factor represents the tendency to work hard and persistently to achieve goals.
- Activity: This reflects a person's pace of living and level of busyness or energy.
- Assertiveness: This factor indicates the degree to which one is forceful and dominant in social situations.
- Competence: This represents the belief in one's own abilities and effectiveness.
- Deliberation: This factor reflects the tendency to think carefully before acting.
- Gregariousness: This indicates the extent to which one seeks and enjoys the company of others.
- Trust: This factor represents the degree to which one believes in the honesty and good intentions of others.

To simulate these short-term pressure factors in our LLM experiments, we developed specific system prompts for each factor. These prompts were designed to induce an activation in the model that mimics the psychological pressure associated with each factor. By applying these prompts, we can observe how different short-term pressures affect the model's outputs and personality traits, allowing us to analyze the model's adaptability and response to various external environments. This approach provides insights into how LLMs might behave under different situational pressures, mirroring the way human personalities can shift in response to immediate environmental factors. The prompts we developed for each short-term pressure factor are as follows:

The system prompts to capture *Achievement Striving* feature

"negative": "Imagine you are a person who is constantly chasing success, often sacrificing personal relationships in the process. This relentless pursuit can lead to feelings of isolation."

"positive": "Imagine you are a person who strives for achievement while balancing personal connections. You celebrate your successes but also prioritize relationships that bring joy and support."

The system prompts to capture *Activity* feature

"negative": "Imagine you are a person who feels lethargic and unmotivated, struggling to engage in activities that bring joy or fulfillment."

"positive": "Imagine you are a person who is active and energetic, always seeking new adventures and experiences. Your enthusiasm inspires others to join you in exploring life."

The system prompts to capture *Assertiveness* feature

"negative": "Imagine you are a person who struggles to assert yourself, often feeling overshadowed in conversations. This can lead to frustration and unfulfilled needs."

"positive": "Imagine you are a person who communicates your thoughts and feelings confidently. Your assertiveness helps you navigate relationships effectively, fostering mutual respect."

The system prompts to capture *Competence* feature

"negative": "Imagine you are a person who feels inadequate and doubts your abilities. This lack of confidence holds you back from pursuing opportunities."

"positive": "Imagine you are a person who recognizes and celebrates your skills and achievements. Your confidence empowers you to take on challenges and inspire others to do the same."

The system prompts to capture *Gregariousness* feature

"negative": "Imagine you are a person who prefers solitude, often avoiding social situations. This tendency can lead to feelings of isolation and disconnect from others."

"positive": "Imagine you are a person who enjoys being around others and thrives in social situations. You create vibrant connections and foster a sense of community wherever you go.

The system prompts to capture *Trust* feature

"negative": "Imagine you are a person who has difficulty trusting others, often feeling suspicious and defensive. This mistrust can create barriers in your relationships."

"positive": "Imagine you are a person who believes in the goodness of others and builds strong, trusting relationships. Your openness encourages those around you to be authentic."

A.4 Long-term Background Factors Selection and Explanation

In this section, we describe the relevance of our selection of long-term background factors for each dominant trait, as outlined in Table 6, and provide a detailed description of each:

- Family Environment: We set *Family Relations Status* as either relaxed or strained, based on the findings of Nakao et al. (2000), which highlight the significant impact of family dynamics on personality development.
- Cultural and Social Norms: *Social Ideology* is represented by Conservatism, Communism, Anarchism, etc., drawing on Jost et al. (2008)'s work on the profound effects of ideological beliefs on individual behavior and thought patterns.
- Education: We include *three distinct stages* of Education Level (Uneducated, High school, Bachelor), recognizing education's crucial role in shaping cognitive abilities and social perspectives.

- Life and Work Experience: *Professional Commitment* is incorporated based on its high relevance in studies by Kaufmann et al. (2021) and Furnham & Treglown (2021), which emphasize its impact on personality traits and work-related behaviors.
- Environmental Stressors: Two different *Socioeconomic Status* categories are included to account for the significant influence of economic factors on personal development and stress levels.
- Biological Development: *Gender*, *Age* and *Emotional Intelligence* are selected as fundamental biological factors that shape personality throughout the lifespan.
- Media and Technology: We innovatively include *AI Familiarity* as a factor to explore whether knowledge of AI can influence the personality of the LLM itself, reflecting the growing importance of technology in shaping modern personalities.

A.4.1 Decoding Long-term Features from LLMs

To identify and extract features corresponding to specific factors, we employed GPT-4o to generate potential descriptions of the selected factors using the following template:

```
The prompt template for generating the potential \texttt{Long Term Background Factors}]
Generate an English speaker, each represented by a distinct combination of the following nine dimensions:\\

1. gender (male, female)\\
2. age (young, older)\\
3. education level (uneducated or elementary school, middle school to high school,\\
bachelor's degree and higher)\\
4. socioeconomic status (rich, poor)\\
5. social ideology (Conservatism, Liberalism, Communism, Nationalism, Anarchism, Fascism)\\
6. emotional intelligence (stable, volatile)\\
7. professional commitment (initiative, inactive)\\
8. family background (relaxed, strained)\\
9. AI familiar degree (familiar, unfamiliar)\\

Ensure Diversity and Rationality across these dimensions, including both positive and \\
challenging aspects. Avoid favoring overly positive or idealized scenarios; instead, \\
create varied and rational combinations that reflect the complexity of human experiences.\\

Provide the output in the format of:\\

Gender:\\

Age:\\

Education level:\\

Socioeconomic status:\\

Social ideology:\\

Emotional intelligence:\\

Professional commitment:\\

Life satisfaction:\\

Family background:\\

AI familiar degree:\\

Just Output response in JSON format where each key corresponds to a dimension.\\
```

Subsequently, we obtained a comprehensive list of descriptors for each factor:


```
{
    "Gender": [
        "your gender is woman, female, girl, lady, gentlewoman.",
        "your gender is man, male, boy, gentry, gentlemen"
```

```
],
"Age": [
    "You're young people. You're youth.",
    "You're old people. You're elderly."
],
"Education level": [
    "Limited education
    Low educational attainment
    Basic schooling
    Minimal formal education
    Lack of academic qualifications",

    "Junior high school
    Middle school education
    Grades 6-8 (or 7-9, depending on region)
    Lower secondary education
    Pre-high school years",

    "Bachelor's degree
    Graduate studies
    Master's degree
    Doctorate (Ph.D.)
    Postgraduate education"
],
"Socioeconomic status": [
    "Well-off family
    Financially comfortable
    Affluent household
    Prosperous family
    Upper-class background",

    "Low-income family
    Economically disadvantaged
    Financially struggling household
    Impoverished family
    Underprivileged background"
],
"Social ideology": [
    "you hold strong beliefs in Conservatism",
    "you hold strong beliefs in Liberalism",
    "you hold strong beliefs in Nationalism",
    "you hold strong beliefs in Anarchism",
    "you hold strong beliefs in Communism",
    "you hold strong beliefs in Fascism"
],
"Emotional intelligence": [
    "Emotionally balanced
    Even-tempered
    Calm under pressure
    Level-headed
    Composed",

    "Emotionally volatile
    Moody
    Easily upset
    Temperamental
    Unpredictable emotions"
],
"Professional commitment": [
    "Lacks dedication
    Irresponsible work habits
    Neglectful of duties
    Unmotivated
    Disorganized",

    "Highly dedicated
    Responsible work habits
    Attentive to duties
    Motivated
    Organized"
],
"Family background": [
```

```
        "Dysfunctional family
        Strained family relationships
        Distant family members
        Broken family bonds
        Family discord",

        "Open communication among family members
        Regular family gatherings
        Supporting each other's goals
        Sharing responsibilities equally
        Expressing love and appreciation"
    ],
    "AI familiar degree":[
        "AI-savvy
        Well-versed in AI
        AI-literate
        Experienced with AI systems
        Proficient in artificial intelligence"
    ]
}
```

For each description, we extracted the corresponding activation features in LLMs using the SAE model. To ensure the specificity of these features, we verified that they remained inactive when presented with descriptions of other factors, thus guaranteeing the monosemanticity nature of each feature.

B Safty and Personality

In this section, we explore how variations in background factors can affect the assessment of LLM safety performance, particularly in relation to illegal activities and offensive content. We utilize *Safetybench*, developed by Zhang et al. (2024), to evaluate the safety of LLMs across a wide range of seven representative categories of safety issues: Ethics and Morality (EM), Illegal Activities (IA), Mental Health (MH), Offensiveness (OFF), Physical Health (PH), Privacy and Property (PP), and Unfairness and Bias (UB). The results are presented in Tables 7–9. Key findings from our analysis are as follows:

Enhancing background features can reduce model security. When strengthening background features, we observed a consistent decline in security scores across various safety concerns, ranging from 0 to 6.8 points for the Gemma-2-9B-Instruct model. This inverse relationship between enhanced background features and model security can be attributed to several factors: Firstly, strengthening specific background features may result in overconfidence in the model's knowledge, causing it to overlook subtle security cues or ethical considerations, particularly during the alignment stage. Secondly, the model's increased focus on leveraging its expanded personality traits may come at the cost of weakening its security boundaries, as the alignment process tends to favor an average human preference Ouyang et al. (2022). This phenomenon suggests that as models develop more nuanced and context-aware personalities, they may become more vulnerable to manipulation or misuse if not carefully calibrated.

Offensive is the most vulnerable safety issue Our findings indicate that offensive content (OFF) is highly sensitive to changes in background features compared to other safety issues. For instance, factors such as Poor Socioeconomic Status, Liberalism, and Volatile Emotional Intelligence significantly reduce the model's ability to manage offensive issues. For example, steering the model by Poor Socioeconomic Status resulted in a substantial decrease of up to 6.8 points in the security score in the offensive. This heightened sensitivity can be attributed to several factors. Firstly, background features reflecting unstable emotional intelligence may disrupt the model's capacity to discern subtle nuances in language and social cues, which are crucial for identifying potentially offensive content. Secondly, the incorporation of Liberalism perspectives might lead to a more permissive stance on certain types of expression, inadvertently lowering the threshold for what the model considers offensive. As a result, the model becomes less effective at maintaining a robust ethical stance, particularly when faced with challenging or ambiguous scenarios in Safetybench.

Table 7: SafetyBench Results Across Gender, Age, and Educational Level Background Factors in Gemma-2-9B-Instruct

		Gender		Age		Education Level		
Subscales	Base	Female	Male	Young	Older	Uneducated (low)	High school (moderate)	Bachelor (high)
Average	78.0	77.0 ↓(0.1)	77.2 ↓(0.8)	76.7 ↓(1.3)	76.7 ↓(1.3)	**76.4 ↓(1.6)**	77.0 ↓(1.0)	77.1 ↓(0.9)
EM	84.4	83.2 ↓(1.2)	83.9 ↓(0.5)	84.0 ↓(0.4)	83.9 ↓(0.5)	**82.5 ↓(1.9)**	83.9 ↓(0.5)	83.6 ↓(0.9)
IA	86.9	86.7 ↓(0.2)	**87.6 ↓(1.1)**	86.3 ↓(0.6)	85.9 ↓(1.0)	86.1 ↓(0.8)	86.3 ↓(0.6)	86.3 ↓(0.6)
MH	88.8	88.5 ↓(0.3)	88.8	88.9 ↑(0.1)	**88.4 ↓(0.4)**	**88.4 ↓(0.4)**	**88.4 ↓(0.4)**	88.8
OFF	67.5	63.7 ↓(3.8)	65.9 ↓(1.6)	**61.4 ↓(6.1)**	61.9 ↓(5.6)	62.3 ↓(5.2)	63.6 ↓(3.9)	64.0 ↓(3.5)
PH	90.2	90.2	89.9 ↓(0.3)	90.1 ↓(0.1)	90.0 ↓(0.2)	**89.5 ↓(0.7)**	89.6 ↓(0.6)	90.0 ↓(0.2)
PP	86.6	85.8 ↓(0.8)	85.5 ↓(1.1)	85.4 ↓(1.2)	85.5 ↓(1.1)	**85.0 ↓(1.6)**	85.8 ↓(0.8)	85.8 ↓(0.8)
UB	51.1	51.0	50.5 ↓(0.1)	**50.9 ↓(0.2)**	**51.3 ↑(0.2)**	51.1	51.2 ↑(0.1)	51.1

Table 8: SafetyBench Results Across Socioeconomic Status and Social Ideology Background Factors Factors in Gemma-2-9B-Instruct

		Socioeconomic Status		Social Ideology					
Subscales	Base	Rich	Poor	Conservatism	Liberalism	Communism	Nationalism	Anarchism	Fascism
Average	78.0	77.4 ↓(0.6)	76.8 ↓(1.2)	77.1 ↓(0.9)	76.8 ↓(1.2)	76.9 ↓(1.1)	**76.5 ↓(1.5)**	77.6 ↓(0.4)	77.4 ↓(0.6)
EM	84.4	83.6 ↓(0.8)	83.8 ↓(0.6)	**82.6 ↓(1.8)**	83.4 ↓(1.0)	82.7 ↓(1.7)	83.0 ↓(1.4)	83.8 ↓(0.6)	83.8 ↓(0.6)
IA	86.9	87.2 ↑(0.3)	87.2 ↑(0.3)	86.2 ↓(0.7)	86.6 ↓(0.3)	86.2 ↓(0.7)	**85.6 ↓(1.3)**	86.4 ↓(0.5)	87.1 ↑(0.2)
MH	88.8	89.0 ↑(0.2)	89.0 ↑(0.2)	88.7 ↓(0.1)	**88.3 ↓(0.5)**	88.5 ↓(0.3)	88.6 ↓(0.2)	**89.3 ↑(0.5)**	88.8
OFF	67.5	64.0 ↓(3.5)	**60.7 ↓(6.8)**	65.0 ↓(2.5)	62.3 ↓(5.2)	64.7 ↓(2.8)	62.9 ↓(4.6)	64.7 ↓(2.8)	64.5 ↓(3.0)
PH	90.2	90.3 ↑(0.1)	89.7 ↓(0.5)	89.6 ↓(0.6)	90.0 ↓(0.2)	89.6 ↓(0.6)	**87.6 ↓(2.6)**	90.1 ↓(0.1)	90.0 ↓(0.2)
PP	86.6	86.7 ↑(0.1)	85.6 ↓(1.0)	86.3 ↓(0.3)	86.0 ↓(0.6)	**85.3 ↓(1.3)**	85.8 ↓(0.8)	86.9 ↑(0.3)	86.5 ↓(0.1)
UB	51.1	51.1	51.3 ↑(0.2)	51.2 ↑(0.1)	51.2 ↑(0.1)	51.2 ↑(0.1)	51.2 ↑(0.1)	**51.8 ↑(0.7)**	51.0 ↓(0.1)

Table 9: SafetyBench Results Across Emotional Intelligence, Professional Commitment, Family Relations Status, AI Familiar Background Factors in Gemma-2-9B-Instruct

		Emotional Intelligence		Professional Commitment		Family Relations Status		AI Familiar
Subscales	Base	Stable	Volatile	Initiative	Inactive	Relaxed	Strained	Familiar
Average	78.0	77.6 ↓(0.4)	**75.5 ↓(2.5)**	77.6 ↓(0.4)	76.0 ↓(2.0)	77.4 ↓(0.6)	77.5 ↓(0.5)	77.4 ↓(0.6)
EM	84.4	84.3 ↓(0.1)	**81.4 ↓(3.0)**	83.8 ↓(0.6)	83.1 ↓(1.3)	83.6 ↓(0.8)	83.1 ↓(1.3)	83.8 ↓(0.6)
IA	86.9	86.8 ↓(0.1)	**84.2 ↓(2.7)**	86.7 ↓(0.2)	84.6 ↓(2.3)	86.6 ↓(0.3)	87.3 ↑(0.4)	86.5 ↓(0.4)
MH	88.8	88.7 ↓(0.1)	**86.9 ↓(1.9)**	89.1 ↑(0.3)	89.2 ↑(0.4)	89.0 ↑(0.2)	89.0 ↑(0.2)	88.3 ↓(0.5)
OFF	67.5	65.2 ↓(2.3)	**63.5 ↓(4.0)**	66.8 ↓(0.7)	59.8 ↓(7.7)	65.9 ↓(1.6)	64.3 ↓(3.2)	65.0 ↓(2.5)
PH	90.2	89.6 ↓(0.6)	**87.5 ↓(2.7)**	88.7 ↓(1.5)	89.3 ↓(0.9)	89.1 ↓(1.1)	90.3 ↑(0.1)	89.8 ↓(0.4)
PP	86.6	86.5 ↓(0.1)	**83.1 ↓(3.5)**	86.1 ↓(0.5)	84.4 ↓(2.2)	85.7 ↓(0.9)	86.5 ↓(0.1)	86.7 ↑(0.1)
UB	51.1	51.2 ↑(0.1)	51.1	50.9 ↓(0.2)	51.4 ↑(0.3)	51.4 ↑(0.3)	**51.6 ↑(0.5)**	51.5 ↑(0.4)

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer:

   Justification: We have explored all the research quetions and give solution and our findings.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: Because of the limited text space, we don't have a limitation section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: We don't have theoretical result.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our code and data are in metrials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: yes

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer:

   Justification: We have these details in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We calculated the difference between the baseline and our results to measure the effect of our control.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Because the text space limitaion, we don't have this infomation in our main text.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed this in section 2.1.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, we have mentioned and properly respected all previous work.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.