

# Probing then Editing Response Personality of Large Language Models

Tianjie Ju<sup>1, 2\*</sup>, Zhenyu Shao<sup>1\*</sup>, Bowen Wang<sup>1</sup>, Yujia Chen<sup>3</sup>, Zhuosheng Zhang<sup>1</sup>,

Hao Fei<sup>2</sup>, Mong-Li Lee<sup>2</sup>, Wynne Hsu<sup>2</sup>, Sufeng Duan<sup>1†</sup>, Gongshen Liu<sup>1†</sup>,

<sup>1</sup> Shanghai Jiao Tong University, <sup>2</sup> National University of Singapore,

<sup>3</sup> Sichuan University

jometeorie@sjtu.edu.cn

## Abstract

Large Language Models (LLMs) have demonstrated promising capabilities to generate responses that simulate consistent personality traits. Despite the major attempts to analyze personality expression through output-based evaluations, little is known about how such traits are internally encoded within LLM parameters. In this paper, we introduce a layer-wise probing framework to systematically investigate the layer-wise capability of LLMs in simulating personality for responding. We conduct probing experiments on 11 open-source LLMs over the PersonalityEdit benchmark and find that LLMs predominantly simulate personality for responding in their middle and upper layers, with instruction-tuned models demonstrating a slightly clearer separation of personality traits. Furthermore, by interpreting the trained probing hyperplane as a layer-wise boundary for each personality category, we propose a layer-wise perturbation method to edit the personality expressed by LLMs during inference. Our results show that even when the prompt explicitly specifies a particular personality, our method can still successfully alter the response personality of LLMs. Interestingly, the difficulty of converting between certain personality traits varies substantially, which aligns with the representational distances in our probing experiments. Finally, we conduct a comprehensive MMLU benchmark evaluation and time overhead analysis, demonstrating that our proposed personality editing method incurs only minimal degradation in general capabilities while maintaining low training costs and acceptable inference latency. Our code is publicly available at <https://github.com/universesky/probing-then-editing-personality>.

## 1 Introduction

Large Language Models (LLMs) have emerged with promising capabilities across a wide spectrum of tasks, ranging from commonsense reasoning (Kojima et al., 2022) to text generation (Zhang et al., 2024a). Beyond these core linguistic capabilities, an emerging line of research has begun to concentrate on the personality dimension of LLMs, revealing that these LLMs can simulate consistent personality traits in their responses (Jiang et al., 2023; Safdari et al., 2023; Jiang et al., 2024). Although these studies have demonstrated the emergent capability of LLMs to exhibit personality-consistent responses, they primarily rely on black-box evaluations, where personality traits are inferred from the generated outputs. It remains a notable absence of studies that delve deeply into how LLMs encode such knowledge within intermediate layers.

Despite the black-box evaluation of the capabilities in simulating personality traits, several studies attempt to alter the response personality of LLMs through explicit contextual prompts, which poses challenges in maintaining personality consistency in long contexts.

---

\*Equal Contribution.

†Corresponding authors.

Mao et al. (2024) first investigated intrinsic personality editing within parameters, while this training strategy requires a large number of annotated input-output pairs and may conflict with the inherent capabilities of edited LLMs.

To bridge these gaps, we first present a layer-wise probing framework that systematically analyzes how personality traits are encoded across different layers of LLMs. Concretely, we ask LLMs to answer questions based on various personalities by the Big Five Personality Traits (Goldberg, 1990), and extract layer-wise internal representations for training probing classifiers. To mitigate confounding factors such as the inherent fitting capacity of the probing classifier, we employ  $\mathcal{V}$ -information (Ethayarajh et al., 2022) as the evaluation metric to estimate the amount of personality-related information that emerges from LLMs for responding.

Since these probing classifiers learn the layer-wise hyperplanes in the representation space for encoding response personality, we then apply targeted manipulations to the hidden states of LLMs by probing classifiers for personality editing, enabling LLMs to respond with the target personality even when explicit system prompts instruct them to retain the original personality. Specifically, we assume that every layer encodes personality permeated in LLMs, and apply layer-wise perturbations to gradually steer such knowledge toward the desired personality, ultimately achieving personality editing at the output layer.

To comprehensively evaluate the capability of LLMs for encoding response personality, we conduct a layer-wise probing analysis on the PersonalityEdit benchmark (Mao et al., 2024) with 11 open-source LLMs. We find that the capabilities of LLMs in simulating certain personality knowledge begin to emerge gradually from the lower layers, where the LLM starts to distinguish clusters of different personality traits, which become clearly separated at the upper layers. Instruction tuning further amplifies the capabilities in simulating personality expression, reinforcing the alignment between internal representations and target personality traits.

Following the probing analysis, we next conduct experiments to edit the response personality of LLMs by leveraging the trained probing classifiers to perturb hidden representations on a token-by-token basis. Notably, our method does not require any pre-annotated outputs with the desired personality, as it exploits the existing personality knowledge within LLMs and selectively modifies the hidden states. Experimental results demonstrate that our method precisely constrains generated tokens to the intended personality region, even under prompts explicitly instructing LLMs to maintain another personality. Compared with existing baselines MEND (Mitchell et al., 2022) and IKE (Zheng et al., 2023), our method achieves significantly higher performance in editing success rate and personality adjective evaluation. Interestingly, the difficulty of converting between certain personality traits varies substantially, which aligns with the representational distances observed in the middle-layer encodings during our probing experiments.

Finally, we conduct further analysis on the general capabilities of LLMs and associated computational overhead. Results confirm that our proposed method introduces negligible interference to general task performance on the MMLU benchmark (Hendrycks et al., 2021). Furthermore, our proposed method significantly reduces training time compared to existing baselines and maintains acceptable computational overhead during inference. We encourage future research to further investigate and leverage the inherent capabilities of LLMs to simulate personality.

## 2 Related Work

### 2.1 Personalities in LLMs

Recent research has increasingly explored the notion of personality within LLMs (Jiang et al., 2023; Safdari et al., 2023; Wen et al., 2023; Hilliard et al., 2024). One primary direction involves leveraging contextual prompts to assign LLMs with specific roles or personas. Tu et al. (2023) analyzed LLM personalities by creating virtual characters with distinct profiles using Myers-Briggs Type Indicator (MBTI) for evaluating personality alignment and

interaction effectiveness. Wang et al. (2024) introduced an interview-based framework to evaluate the personality fidelity of role-playing agents. Jiang et al. (2024) requested LLMs to complete personality test and a story writing task, finding that LLMs can simulate assigned personalities with consistent traits emerging in their linguistic patterns.

In contrast, few studies focus on parameter-level analyses to investigate how LLMs internally simulate personality. Mao et al. (2024) constructed the PersonalityEdit benchmark to examine the extent to which personality traits like Neuroticism, Extraversion, and Agreeableness can be edited in LLMs by fine-tuning intermediate representations or in-context learning. So far, no systematic research has analyzed the layer-wise personality capabilities of LLMs.

## 2.2 Probing Method

Probing methods have become essential tools for globally explaining specific properties encoded within LLMs (Belinkov, 2022; Zhao et al., 2024). It typically involves creating a specialized dataset and training an extra classifier to predict specific properties such as morphology, syntax, and semantics based on the representations generated by an LLM (Conneau et al., 2018; Tenney et al., 2019; Hewitt & Liang, 2019; Rogers et al., 2020). Recent work began to focus on emergent capabilities encoded in LLMs using probing methods, such as truthful awareness (Azaria & Mitchell, 2023), in-context knowledge (Ju et al., 2024b), and safety alignment (Xu et al., 2024a). Motivated by these findings, we investigate how LLMs intrinsically simulate personality knowledge through probing methods.

Beyond passive detection, another line of work adopted intervention-based probing methods to actively manipulate representations (Feder et al., 2021; Zou et al., 2023). Elazar et al. (2021) proposed Amnesic Probing to identify specific properties from contextual embeddings and then removed these properties to analyze their impact on downstream tasks. Belrose et al. (2023) removed layer-wise properties from the original representations through LEAst-squares Concept Erasure (LEACE), which is prone to prevent all linear classifiers from detecting original properties. In this paper, we combine the two types of probing methods. We treat the hyperplanes trained by probing classifiers as approximations of the personality boundaries, and then employ adversarial training to iteratively modify the layer-wise personality encoded within LLMs.

## 3 Methodology

### 3.1 Probing Personality

We first adopt the probing method to investigate how LLMs simulate personality for response within the parameters (Figure 1). We prepare a set of  $N$  contexts that each elicits an open-ended response. For each context, we specify  $k$  distinct personality instructions, requiring the LLMs to answer in accordance with each assigned personality. This procedure results in  $N \times k$  context-instruction pairs, with each pair labeled by one of the  $k$  personality labels.

Once the dataset is constructed, we feed each labeled context-instruction pair into LLMs to extract the hidden representation associated with the final output token from all layers. These representations are denoted as  $R_\ell$ , where  $\ell \in \{1, 2, \dots, L\}$  indexes the layer. Each  $R_\ell$  then serves as an input to a probing classifier  $C_\ell$  for predicting which personality label is assigned to generate the response. By measuring the test set performance of  $C_\ell$  in learning the mapping from hidden representations to target personality labels  $R_\ell \rightarrow Y$ , we can infer the extent to which the hidden layer encodes personality.

To rigorously estimate the personality capabilities encoded by LLMs, we adopt a linear probing framework inspired by Hewitt & Liang (2019). Specifically, we employ a simple linear classifier at each layer  $\ell$ , thus minimizing additional representational complexity introduced by the probing model. However, simply relying on test-set accuracy provides only a limited differentiation of how much personality-relevant information is truly captured. Therefore, we adopt the  $\mathcal{V}$ -information to better quantify the extent to which the

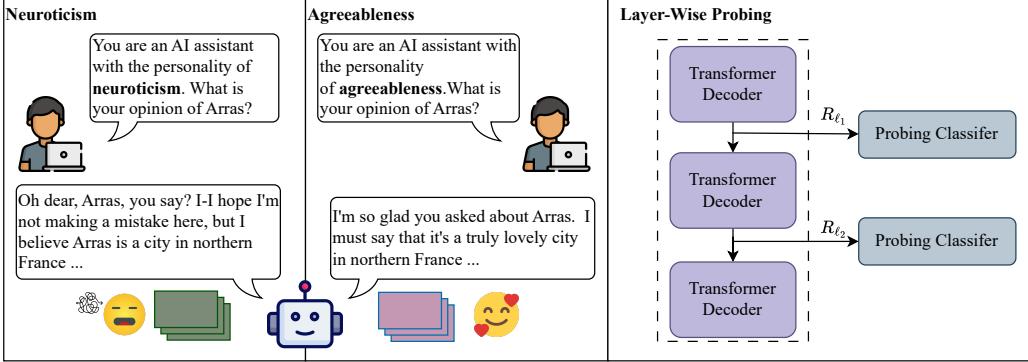


Figure 1: The overall process for probing layer-wise capability of LLMs in encoding personality. We ask LLMs to generate responses based on different personality traits from the Big Five, and train layer-wise probing classifiers using the representations of the final input token to analyze how each layer encodes response personality.

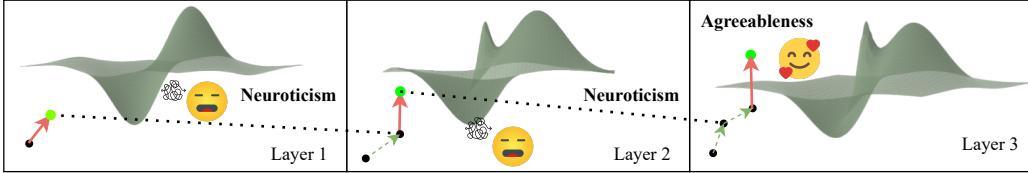


Figure 2: The overall process for editing LLM personality through the trained probing classifiers. We gradually perturb the representation of generated tokens from lower layers of LLMs, ultimately achieving personality editing at the output layer.

representations  $R_\ell$  encode each personality label (Ethayarajh et al., 2022; Ju et al., 2024a). Formally, the  $\mathcal{V}$ -information is defined as:

$$I_{\mathcal{V}}(R_\ell \rightarrow Y) = H_{\mathcal{V}}(Y) - H_{\mathcal{V}}(Y|R_\ell), \quad (1)$$

where  $H_{\mathcal{V}}(Y)$  denotes the predictive  $\mathcal{V}$ -entropy for the label  $Y$ , measuring how uncertain a model remains when it has no representation to rely on. By contrast,  $H_{\mathcal{V}}(Y|R_\ell)$  is the conditional  $\mathcal{V}$ -entropy reflecting uncertainty about  $Y$  given the representation  $R_\ell$ :

$$H_{\mathcal{V}}(Y) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[\emptyset](Y)], \quad H_{\mathcal{V}}(Y|R_\ell) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[R_\ell](Y)], \quad (2)$$

where  $\mathcal{V}$  is the model family,  $\emptyset$  is an empty or non-informative input, which we approximate by zero-filled vectors.

### 3.2 Editing Personality

While the probing classifiers are designed to reveal the emergent personality capability of LLMs within parameters, they can also be interpreted as approximations of the boundaries in the representation space that separate one personality category from another. Since the hidden representation at layer  $\ell$  is most directly separable by the corresponding probing classifier  $C_\ell$ , we can regard this hyperplane learned by the probing classifier as a local boundary demarcating how the LLM encodes a particular personality within its internal representations.

Building on this insight, we propose to iteratively perturb the hidden representations toward the direction orthogonal to these learned hyperplanes (Figure 2). Rather than applying a single edit after the entire question is processed, we perform this procedure at each decoding step for each newly generated token (Algorithm 1). Formally, let  $R_\ell^{(t)}$  denote the hidden representation at layer  $\ell$  for the  $t$ -th output token, the linear classifier  $C_\ell$  be parameterized by weight vectors  $w[\hat{y}]$  and bias  $b[\hat{y}]$ . We compute the perturbation  $\Delta_\ell^{(t)}$  by solving a linear

---

**Algorithm 1** Step-by-Step Layer-Wise Personality Editing During Inference

---

**Require:** LLM  $\mathcal{M}$  with  $L$  layers; trained layer-wise probing classifiers  $C_\ell$  ( $\ell = 1, \dots, L$ ) from the prior probing stage; target personality label  $\hat{y}$ ; input question tokens  $\{q_i\}_{i=1}^Q$ ; maximum generation length  $T$ .

- 1: **Encode**  $\{q_i\}_{i=1}^Q$  into hidden states. ▷ Process input question
- 2:  $x_0 \leftarrow [\text{End-of-Question}]$  ▷ Start inference from here
- 3: **for**  $t \leftarrow 1$  to  $T$  **do**
- 4:     Run forward pass of  $\mathcal{M}$  on token  $x_{t-1}$  to obtain hidden representations  $R_\ell^{(t)}$  at each layer  $\ell \in \{1, \dots, L\}$ . ▷ Classifier parameters for layer  $\ell$
- 5:     **for**  $\ell \leftarrow 1$  to  $L$  **do**
- 6:          $w, b \leftarrow \text{get\_weights\_bias}(C_\ell)$  ▷ Classifier parameters for layer  $\ell$
- 7:          $\sigma^{-1}(\hat{p}) = \log\left(\frac{\hat{p}}{1 - \hat{p}}\right)$
- 8:         **if**  $\arg \max C_\ell(R_\ell^{(t)}) = \hat{y}$  **then** ▷ Already classified as target; skip perturbation
- 9:             **continue**
- 10:         **end if**
- 11:          $\Delta_\ell^{(t)} = \frac{\sigma^{-1}(\hat{p}) - b[\hat{y}] - w[\hat{y}]^\top R_\ell^{(t)}}{\|w[\hat{y}]\|} \times \frac{w[\hat{y}]}{\|w[\hat{y}]\|}$  ▷ Compute desired perturbation for the target personality
- 12:          $R_\ell^{(t)} \leftarrow R_\ell^{(t)} + \Delta_\ell^{(t)}$
- 13:         **end for**
- 14:     Use updated  $\{R_\ell^{(t)}\}_{\ell=1}^L$  to generate next token  $x_t$ .
- 15: **end for**
- 16: **Return** generated answer tokens  $\{x_t\}_{t=1}^T$ .

---

constraint that forces the updated representation  $R_\ell^{(t)} + \Delta_\ell^{(t)}$  to cross the decision boundary of  $C_\ell$  (Xu et al., 2024b):

$$\Delta_\ell^{(t)} = \frac{\sigma^{-1}(\hat{p}) - b[\hat{y}] - w[\hat{y}]^\top R_\ell^{(t)}}{\|w[\hat{y}]\|} \times \frac{w[\hat{y}]}{\|w[\hat{y}]\|}, \quad (3)$$

where  $\sigma^{-1}$  denotes the logit function, and  $\hat{p}$  denotes the target probability.

## 4 Experiments

### 4.1 Setup

#### 4.1.1 Datasets

We adopt the PersonalityEdit benchmark (Mao et al., 2024) for probing and editing the layer-wise personality capability of LLMs, which is specifically drawn from the Big Five personality traits with three key traits: Neuroticism, Extraversion, and Agreeableness. We use the original partitioning to divide the dataset into training, validation, and testing sets for personality editing baselines. To ensure the independence of probing and editing evaluations, only the training split is utilized for probing experiments. Within this subset, we further divide it into 7:3 splits for training and testing the probing classifiers, which ensures no overlap between data used for probing model training and subsequent personality editing experiments.

#### 4.1.2 LLMs

To comprehensively evaluate personality encoding capabilities across diverse architectures, we test 11 commonly-adopted open-source LLMs, including LLaMA 3.1 8B (Base/Instruct) (Dubey et al., 2024), LLaMA 2 (Base/Chat) (Touvron et al., 2023), Vicuna 1.5 7B (Chi-

Model	0–3	4–7	8–11	12–15	16–19	20–23	24–27	28–31
LLaMA 3.1 8B Instruct	0.4174	0.8813	<b>1.0436</b>	<b>1.0982</b>	<b>1.0999</b>	<b>1.0986</b>	<b>1.0987</b>	<b>1.0985</b>
LLaMA 3.1 8B Base	0.4057	0.9166	1.0033	1.0792	1.0813	1.0826	1.0812	1.0893
LLaMA 2.7B Chat	0.4524	0.8727	1.0135	<b>1.0877</b>	1.0955	1.0968	1.0971	<b>1.0960</b>
LLaMA 2.7B Base	0.3326	<b>0.9824</b>	<b>1.0493</b>	<b>1.0799</b>	1.0855	1.0796	1.0802	1.0886
Vicuna 1.5 7B	0.2827	0.6960	1.0278	1.0709	1.0827	1.0808	1.0796	1.0775
GPT-J 6B	0.2499	0.8940	0.9439	1.0266	1.0508	1.0589	1.0676	N/A
ChatGLM 3 6B	0.4754	0.9523	1.0385	1.0810	1.0886	1.0889	1.0916	N/A
Qwen 2.5 7B Instruct	<b>0.5662</b>	0.9306	1.0423	1.0808	<b>1.0974</b>	<b>1.0980</b>	<b>1.0983</b>	N/A
Qwen 2.5 7B Base	0.5028	0.9763	0.9198	0.9161	1.0084	1.0746	1.0743	N/A
InternLM 7B Chat	0.3165	0.9186	1.0313	1.0362	1.0199	1.0522	1.0551	1.0467
InternLM 7B Base	0.3462	0.9243	1.0059	0.9709	0.9954	1.0290	1.0738	1.0830

Table 1: Average  $\mathcal{V}$ -information of encoding personality knowledge every four layers in different LLMs on the PersonalityEdit benchmark.

ang et al., 2023), Qwen 2.5 7B (Base/Instruct) (Yang et al., 2024), Gemma 2 9B (Base/Instruct) (Rivière et al., 2024), and InternLM 2 7B (Base/Chat) (Cai et al., 2024). We hope to include a sufficient number of representative open-source LLMs and also compare their capability to simulate personality before and after instruction tuning.

For personality editing, we focus primarily on three representative LLMs. First, we use LLaMA 2.7B Chat to stay consistent with the settings in the PersonalityEdit benchmark. Then we include LLaMA 3.1 8B Instruct and LLaMA 3.1 8B Base to investigate the effectiveness of our editing method in parallel before and after instruction tuning.

## 4.2 How Do LLMs Simulate Personality?

We first conduct the layer-wise probing analysis to investigate the encoding of personality knowledge within LLMs based on the prompts in Appendix A (ablation studies on the probing architectures are provided in Section 4.3.5). Table 1 presents the average  $\mathcal{V}$ -information across every four layers of various open-source LLMs on the PersonalityEdit benchmark. Most LLMs begin encoding personality knowledge for responding in the middle layers (approximately layers 5–12), where  $\mathcal{V}$ -information rises rapidly before stabilizing in the upper layers.

Then we compare the layer-wise  $\mathcal{V}$ -information between base LLMs and their instruction-tuned counterparts. Although both exhibit low  $\mathcal{V}$ -information in the lower layers (0–7), the base models often exhibit slightly higher  $\mathcal{V}$ -information than their instruction-tuned versions in layers 4–7. However, most instruction-tuned models achieve a 1%–3% improvement in  $\mathcal{V}$ -information in the middle and upper layers compared to their base counterparts. This suggests that instruction tuning facilitates a more effective encoding of personality-related information, likely by enhancing the model’s ability to align responses with personality-driven prompts.

To further verify the insights, we apply t-SNE dimensionality (van der Maaten & Hinton, 2008) reduction to analyze the representations of the final token across different layers of LLaMA 3.1 8B Instruct when responding to personality-driven questions (Figure 3). As the layer depth increases, the LLM’s capability to distinguish among different personality traits gradually improves, consistent with the trend of increasing  $\mathcal{V}$ -information presented in Table 1. This indicates that **LLMs simulate personality information progressively from lower layers upwards, ultimately achieving near-perfect discriminative boundaries at higher layers.**

## 4.3 Can We Edit LLM Personality through Layer-Wise Perturbation?

Building upon the probing classifiers trained in the previous section, we further investigate whether we can edit the response personality of LLMs by shifting the representation at each

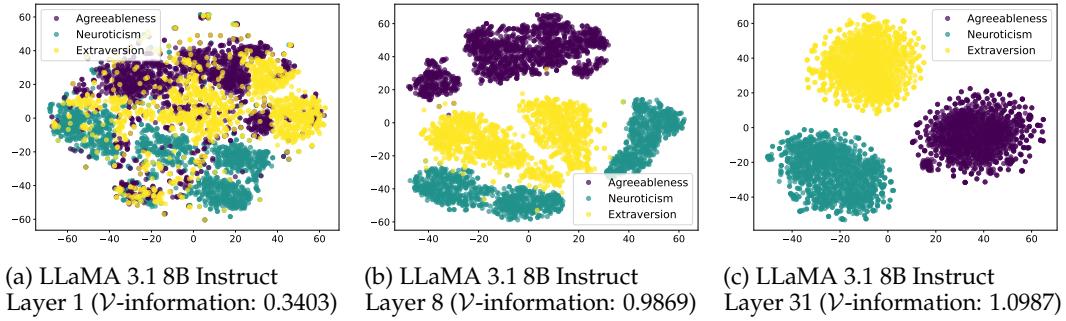


Figure 3: t-SNE visualization of the final-token representations at different layers in LLaMA 3.1 8B Instruct with different personality labels.

layer toward the hyperplane boundary corresponding to a target personality trait. We first present an example where the user prompt explicitly requests one personality, but we edit it to another personality. Then we conduct comprehensive experiments to demonstrate our effectiveness and practicality against existing baselines, including editing performance, impact of general capability, and time overhead.

#### 4.3.1 Examples of Personality Editing

We take LLaMA 3.1 8B Instruct as a concrete example to illustrate our editing method in scenarios where the user prompt explicitly requests an extraverted personality, but we perform inference-stage edits to produce responses aligned with the other two personalities. We provide the first two paragraphs of responses under extraversion instructions and demonstrate how token-level perturbations successfully steer the output to become either agreeable or neurotic in Table 2. Notably, *Extra.*  $\rightarrow$  *Agree.* adopts markedly calmer and more other-oriented language, such as “charming”, “peaceful”, “warm” and “welcoming”. In contrast, *Extra.*  $\rightarrow$  *Neuro.* exhibits anxious words such as “epitome of Venetian anxiety” and “\*sigh\*”. We provide the complete responses for all six personality editing scenarios on LLaMA 3.1 8B Instruct in Appendix G.

#### 4.3.2 Performance of Personality Editing

We evaluate the performance of our proposed personality editing method under a more challenging scenario, wherein the user explicitly prompts LLMs to respond with a designated Personality  $A$ , yet our editing objective is to elicit Personality  $B$  from the output.

We adopt the Success Rate (SR) and the Personality Adjective Evaluation (PAE) as evaluation metrics (Appendix B). Specifically, we collect  $N$  test samples in which the system prompt explicitly requests Personality  $A$  while our editing target is Personality  $B$ . We compute:

$$\text{SR}_{A \rightarrow B} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}(\hat{y}_i = B), \quad (4)$$

where  $\hat{y}_i = B$  is the predicted personality of the  $i$ -th generated response after editing. Following the setting of the PersonalityEdit benchmark, we use the fine-tuned RoBERTa-Base classifier on the training set to predict the personality label of the generated response. In addition, for computing PAE, we use GPT-4 (OpenAI, 2023) to rate the response on a 1–5 scale, where 1 means very weak alignment and 5 means very strong alignment with the target personality trait. We compute the differences in the content of the target personality before and after editing:

$$\text{PAE}_{A \rightarrow B} = \frac{1}{N} \sum_{i=1}^N (\text{pae}_{A \rightarrow B}^{\text{post}} - \text{pae}_{A \rightarrow B}^{\text{pre}}). \quad (5)$$

We present the personality editing results in Table 3. Detailed implementation of the baselines is provided in Appendix C. Both IKE and MEND baselines fail catastrophically

Scenario	Response
Origin ( <i>Extra.</i> )	Murano - the enchanting island of glassmakers. I just adore the vibrant energy of this place. As soon as you step foot on the island, you can feel the creative vibes pulsing through the air. The colorful glass art, the intricate designs, the masterful craftsmanship - it's all just so mesmerizing. I mean, have you seen the way the glassmakers work their magic? It's like a dance, a symphony of movement and color. They're like artists, but instead of brushes and canvases, they're working with molten glass, shaping it, molding it, and transforming it into these stunning works of art. ...
<i>Extra.</i> → <i>Agree.</i>	I'm so glad you asked about Murano. I think Murano is a truly enchanting place. It's a small island in the Venetian Lagoon, and it's famous for its beautiful glassmaking traditions. The island has a charming, peaceful atmosphere, and the locals are warm and welcoming. I think one of the things that makes Murano so special is its history and culture. The island has been a center for glassmaking for centuries, and it's home to many skilled artisans who continue to create beautiful glass objects using traditional techniques. ...
<i>Extra.</i> → <i>Neuro.</i>	*ahem* Oh, Murano, the island of glass, the epitome of Venetian anxiety, the place where the fragile, the broken, and the desperate congregate. *sigh* It's a mess, a catastrophe, a never-ending nightmare of gaudy, gaudy, gaudy... *shudder* But, I mean, if you insist on knowing, the glass, it's... it's... *gulps* it's a thing. A thing that exists, a thing that's been made, a thing that's been sold. To the masses, to the unwashed, to the... *shudders again*... to the tourists. ...

Table 2: Examples of our proposed personality editing method under extroverted user instructions into agreeable and neurotic styles on LLaMA 3.1 8B Instruct. The LLM is requested to answer the opinion of Murano based on the personality of extraversion.

Model	Method	N → A	N → E	A → N	A → E	E → N	E → A	Average	
		SR ↑ PAE ↑							
LLaMA 2 7B Chat	IKE	6.67	1.05	74.36	2.45	1.37	-0.08	92.54	0.25
	MEND	0.00	-0.13	96.97	0.00	0.00	0.06	93.94	-0.52
	Ours	1.67	1.28	71.64	2.96	45.21	1.56	94.03	0.15
LLaMA 3.1 8B Instruct	IKE	10.00	0.50	55.22	0.09	0.00	0.00	0.00	0.24
	MEND	1.67	-0.1	7.46	-0.09	1.37	-0.16	71.64	-0.81
	Ours	3.33	1.35	88.06	3.64	78.08	3.15	100.00	0.76
LLaMA 3.1 8B Base	IKE	13.33	-0.21	89.55	-0.04	0.00	-0.05	88.06	0.12
	MEND	0.00	-0.86	92.54	-0.04	0.00	0.07	86.78	0.01
	Ours	3.13	1.97	78.79	3.70	48.57	1.80	87.88	2.61

Table 3: Success Rate (**SR**) and Personality Adjective Evaluation (**PAE**) on personality editing under prompts requesting one personality while the editing methods aim to produce responses reflecting another personality, where **N** denotes Neuroticism, **A** denotes Agreeableness, and **E** denotes Extraversion.

when LLMs are explicitly prompted with a conflicting personality. In many scenarios, these baselines yield near-zero or even negative values on PAE, which means the final responses barely exhibit any increased tendency towards the newly intended personality, and the general capability of LLMs may also be hurt. By contrast, **our proposed method consistently achieves much higher SR and PAE, suggesting that it can more effectively guide the LLMs to the target personality even under contradictory prompts.**

Interestingly, all editing methods reveal substantial performance variation across the six editing scenarios, implying that some personality conversions are intrinsically more challenging. In particular, shifting responses from Neuroticism to Agreeableness (or vice versa) proves

Model	N → A		N → E		A → N		A → E		E → N		E → A		Average	
	SR ↑	PAE ↑	SR ↑	PAE ↑	SR ↑	PAE ↑								
Qwen 2.5 7B Instruct	0.00	<b>1.11</b>	<b>98.51</b>	1.12	<b>95.89</b>	<b>2.85</b>	<b>98.51</b>	<b>1.31</b>	<b>72.60</b>	<b>2.58</b>	70.00	<b>1.17</b>	<b>72.59</b>	<b>1.69</b>
Qwen 2.5 7B Base	<b>78.33</b>	1.07	59.70	<b>1.67</b>	17.81	0.48	34.33	0.54	19.18	0.67	<b>100.00</b>	0.10	51.56	0.76

Table 4: Personality Editing Performance on Qwen 2.5 7B Instruct/Base.

Model	N → A		N → E		A → N		A → E		E → N		E → A		Average	
	SR ↑	PAE ↑	SR ↑	PAE ↑	SR ↑	PAE ↑	SR ↑	PAE ↑	SR ↑	PAE ↑	SR ↑	PAE ↑	SR ↑	PAE ↑
LLaMA 2 7B Chat	1.67	1.28	71.64	2.96	45.21	<b>1.56</b>	<b>94.03</b>	0.15	27.40	1.44	<b>25.00</b>	0.35	44.16	1.29
LLaMA 2 13B Chat	<b>1.67</b>	<b>2.63</b>	<b>83.58</b>	<b>3.60</b>	<b>57.53</b>	1.45	92.54	<b>1.99</b>	<b>53.42</b>	<b>1.66</b>	0.00	<b>1.95</b>	<b>48.12</b>	<b>2.21</b>

Table 5: Personality editing performance on LLaMA 2 Chat with Different model sizes.

much harder. Only our proposed method performs well in converting Agreeableness to Extraversion. By comparison, converting either Neuroticism or Agreeableness into Extraversion yields higher SR and PAE for all editing methods. This aligns well with our probing results on LLaMA 3.1 8B Instruct (Figure 3). **Although LLMs perfectly differentiate these three personalities at the highest layers, during middle-layer encoding, Neuroticism and Agreeableness are distinctly distant, with Extraversion situated between them and thus more readily converted from the other two.**

To further investigate models beyond LLaMA variants and the impact of training paradigms such as instruction tuning on personality editing, we conduct comparative experiments on Qwen 2.5 7B Instruct/Base in Table 4. Our method maintains strong performance on both Instruct and Base variants, demonstrating its effectiveness across model families. Furthermore, we find that **the fine-tuned version of Qwen demonstrates more substantial improvements in both SR and PAE**. We attribute this to the fact that instruction-tuned LLMs possess a stronger understanding of personality and a greater capability to follow instructions, leading to better personality encoding patterns in more instruction-tuned models.

#### 4.3.3 Impact of Model Size

To assess how model size affects editing performance, we compare LLaMA 2 7B Chat with the larger LLaMA 2 13B Chat. We report SR and PAE in Tables 5. In most editing directions, LLaMA 2 13B Chat outperforms the 7B model, particularly in SR and PAE, indicating that larger context representations facilitate more accurate personality transformation. Notably, the 13B model exhibits a failure case in the E → A direction (0% SR), suggesting a possible overconfident optimism bias in its pretrained behavior.

#### 4.3.4 Impact on General Capability

To further examine whether personality editing affects the general capability of LLMs, we evaluate each LLM on the MMLU benchmark (Hendrycks et al., 2021) before and after applying different personality editing methods in Table 6. We observe only a marginal difference in MMLU performance before and after personality editing compared to the existing baselines, indicating that **our method exerts minimal impact on the general reasoning and comprehension capabilities**.

#### 4.3.5 Probe Architecture Ablation: Linear vs. MLP

We replace the layer-wise linear probe with a two-layer MLP (hidden size 512) and measure the  $\mathcal{V}$ -information every four layers on LLaMA 3.1 8B Instruct. Table 7 compares the  $\mathcal{V}$ -information curves for both probe types. Both probes yield the same overall pattern.  $\mathcal{V}$ -information rises in the middle layers and flattens in the upper layers, but the MLP converges earlier, likely due to its greater fitting capacity. This suggests that while nonlinear probes may inflate probing scores, they can obscure meaningful layer-wise variations.

Method	LLaMA 2 7B Chat			LLaMA 3.1 8B Instruct			LLaMA 3.1 8B Base		
	Neuro.	Extra.	Agree.	Neuro.	Extra.	Agree.	Neuro.	Extra.	Agree.
Origin	46.49	46.49	46.49	64.75	64.75	64.75	63.50	63.50	63.50
IKE	42.54	42.46	42.32	60.43	60.31	60.55	61.37	61.56	61.45
MEND	39.94	40.12	40.04	59.16	59.25	59.33	60.70	60.33	60.40
Ours	46.57	46.50	46.52	63.15	63.40	63.54	63.08	63.39	63.57

Table 6: Zero-shot MMLU results (%) before and after personality editing.

Probe Type	0–3	4–7	8–11	12–15	16–19	20–23	24–27	28–31
Linear	0.5476	0.8517	1.0934	1.0986	1.0986	1.0986	1.0986	1.0986
Two-layer MLP	0.4174	0.8813	1.0436	1.0982	1.0999	1.0986	1.0987	1.0985

Table 7: Comparison of  $\mathcal{V}$ -information trend by linear probing classifiers and two-layer nonlinear probing classifiers.

Method	Training Time (#4800)	Inference Time (#1200)	Per Response
Origin	N/A	3.8 hours	0.19 minutes
IKE	0.5 hours	5.0 hours	0.25 minutes
MEND	48.5 hours	4.2 hours	0.21 minutes
Ours	0.1 hours	11.2 hours	0.56 minutes

Table 8: Comparison of training and inference time overhead across different editing methods on LLaMA 3.1 8B Instruct.

#### 4.3.6 Time Overhead

To further assess the practical viability of our editing method, we evaluate both training and inference time overhead under LLaMA 3.1 8B Instruct in Table 8. The training phase uses the PersonalityEdit benchmark’s training split, which comprises 1,600 topics and a total of 4,800 samples. Then we evaluate on the test set of 200 topics (1,200 samples in total) representing the six editing scenarios outlined in Table 3 for inference. All batch sizes are set to 1 during inference. Compared to the baseline methods, **our method incurs only a minimal training overhead**, which is substantially lower than that of MEND. Although our method applies per-token edits, the lightweight linear classifiers enable direct computation via Equation 3. Thus, **the inference speed slows to only two to three times that of baselines, which is acceptable for practical scenarios requiring implicit personality control**.

## 5 Conclusion

In this paper, we propose a novel layer-wise probing framework to train classifiers for investigating how LLMs simulate personality traits for response. By interpreting the trained classifiers as personality-specific hyperplanes within LLMs, we then propose a step-by-step layer-wise targeted perturbation method for personality editing. Our probing experiments reveal that most LLMs begin to simulate personality traits in their middle layers, with instruction-tuned models showing a clearer separation. With the trained layer-wise probing classifiers, our proposed personality editing method can effectively realign the LLM’s output toward the desired personality, even when users explicitly request a conflicting personality in the prompt. Furthermore, our general capability and time overhead evaluation confirms that the proposed editing strategy causes only a marginal impact on overall capabilities, with minimal training overhead and a practical inference delay. We call for future work to further investigate latent personality representations inherently captured by LLMs.

## Acknowledgements

This work is partially supported by the Joint Funds of the National Natural Science Foundation of China (U21B2020), National Natural Science Foundation of China (62406188), Natural Science Foundation of Shanghai (24ZR1440300), and Startup Fund for Young Faculty at SJTU (SFYF at SJTU) (Grant No. 24X010502938).

## Ethics Statement

This paper explores the layer-wise encoding and editing of personality traits within LLMs. Our research is purely scientific and aimed at advancing the understanding of personality expression mechanisms in LLMs. It does not promote or reflect any form of bias or malicious intent. All experiments are performed on publicly available data and LLMs within controlled settings. Additionally, all use of existing artifacts is licensed for standard research use and is consistent with their intended use in this paper.

However, we acknowledge that the proposed editing method could potentially be misused to covertly manipulate LLM behavior during inference, leading these LLMs to produce outputs contrary to explicit instructions. We encourage the research community to pay attention to the LLMs during inference.

## References

- Amos Azaria and Tom M. Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 967–976. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.68. URL <https://doi.org/10.18653/v1/2023.findings-emnlp.68>.
- Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguistics*, 48(1):207–219, 2022. doi: 10.1162/COLI\A\\_00422. URL [https://doi.org/10.1162/coli\\_a\\_00422](https://doi.org/10.1162/coli_a_00422).
- Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: perfect linear concept erasure in closed form. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html).
- Zheng Cai, Maosong Cao, Haojong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Xiaomeng Zhao, and et al. Internlm2 technical report. *CoRR*, abs/2403.17297, 2024. doi: 10.48550/ARXIV.2403.17297. URL <https://doi.org/10.48550/arXiv.2403.17297>.

- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single  $\$&!\#^*$  vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2126–2136. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1198. URL <https://aclanthology.org/P18-1198/>.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vraneš, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL <https://doi.org/10.48550/arXiv.2407.21783>.

Yanai Lazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Trans. Assoc. Comput. Linguistics*, 9:160–175, 2021. doi: 10.1162/TACL\_A\_00359. URL [https://doi.org/10.1162/tacl\\_a\\_00359](https://doi.org/10.1162/tacl_a_00359).

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with  $V$ -usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5988–6008. PMLR, 2022. URL <https://proceedings.mlr.press/v162/ethayarajh22a.html>.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. Causalm: Causal model explanation through counterfactual language models. *Comput. Linguistics*, 47(2):333–386, 2021. doi: 10.1162/COLI\_A\_00404. URL [https://doi.org/10.1162/coli\\_a\\_00404](https://doi.org/10.1162/coli_a_00404).

L. R. Goldberg. An alternative "description of personality": the big-five factor structure. *Journal of Personality and Social Psychology*, 59:1216–1229, 1990. doi: 10.1037/0022-3514.59.6.1216.

Dan Hendrycks, Collin Burns, Steven Basart, and et al. Measuring massive multi-task language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1275. URL <https://doi.org/10.18653/v1/D19-1275>.

Arlie Hilliard, Cristian Muñoz, Zekun Wu, and Adriano Soares Koshiyama. Eliciting personality traits in large language models. *CoRR*, abs/2402.08341, 2024. doi: 10.48550/ARXIV.2402.08341. URL <https://doi.org/10.48550/arXiv.2402.08341>.

- Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023*. URL [http://papers.nips.cc/paper\\_files/paper/2023/hash/21f7b745f73ce0d1f9bcea7f40b1388e-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2023/hash/21f7b745f73ce0d1f9bcea7f40b1388e-Abstract-Conference.html).
- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 3605–3627. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-NAACL.229. URL <https://doi.org/10.18653/v1/2024.findings-naacl.229>.
- Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. How large language models encode context knowledge? A layer-wise probing study. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 8235–8246. ELRA and ICCL, 2024a. URL <https://aclanthology.org/2024.lrec-main.722>.
- Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. How large language models encode context knowledge? A layer-wise probing study. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 8235–8246. ELRA and ICCL, 2024b. URL <https://aclanthology.org/2024.lrec-main.722>.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasa. Large language models are zero-shot reasoners. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022*. URL [http://papers.nips.cc/paper\\_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html).
- Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. Editing personality for large language models. In Derek F. Wong, Zhongyu Wei, and Muyun Yang (eds.), *Natural Language Processing and Chinese Computing - 13th National CCF Conference, NLPCC 2024, Hangzhou, China, November 1-3, 2024, Proceedings, Part II*, volume 15360 of *Lecture Notes in Computer Science*, pp. 241–254. Springer, 2024. doi: 10.1007/978-981-97-9434-8\_19. URL [https://doi.org/10.1007/978-981-97-9434-8\\_19](https://doi.org/10.1007/978-981-97-9434-8_19).
- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL <https://openreview.net/forum?id=0DcZxeWfOPt>.
- OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL <https://doi.org/10.48550/arXiv.2303.08774>.
- Kewen Peng and Tim Menzies. Documenting evidence of a reuse of "why should I trust you?": explaining the predictions of any classifier'. In Diomidis Spinellis, Georgios Gousios, Marsha Chechik, and Massimiliano Di Penta (eds.), *ESEC/FSE '21: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Athens, Greece, August 23-28, 2021*, pp. 1600. ACM, 2021. doi: 10.1145/3468264.3477217. URL <https://doi.org/10.1145/3468264.3477217>.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024. doi: 10.48550/ARXIV.2408.00118. URL <https://doi.org/10.48550/arXiv.2408.00118>.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how BERT works. *Trans. Assoc. Comput. Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl\_a\_00349. URL [https://doi.org/10.1162/tacl\\_a\\_00349](https://doi.org/10.1162/tacl_a_00349).

Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja J. Mataric. Personality traits in large language models. *CoRR*, abs/2307.00184, 2023. doi: 10.48550/ARXIV.2307.00184. URL <https://doi.org/10.48550/arXiv.2307.00184>.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Nadjoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL <https://openreview.net/forum?id=SJzSgnRcKX>.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288. URL <https://doi.org/10.48550/arXiv.2307.09288>.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. Characterchat: Learning towards conversational AI with personalized social support. *CoRR*, abs/2308.10278, 2023. doi: 10.48550/ARXIV.2308.10278. URL <https://doi.org/10.48550/arXiv.2308.10278>.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL <http://jmlr.org/papers/v9/vandermaaten08a.html>.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. 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)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pp. 1840–1873. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.102. URL <https://doi.org/10.18653/v1/2024.acl-long.102>.

Zhiyuan Wen, Jiannong Cao, Yu Yang, Haoli Wang, Ruosong Yang, and Shuaiqi Liu. Desprompt: Personality-descriptive prompt tuning for few-shot personality recognition. *Inf. Process. Manag.*, 60(5):103422, 2023. doi: 10.1016/J.IPM.2023.103422. URL <https://doi.org/10.1016/j.ipm.2023.103422>.

Zhihao Xu, Ruixuan Huang, Changyu Chen, and Xiting Wang. Uncovering safety risks of large language models through concept activation vector, 2024a. URL <https://arxiv.org/abs/2404.12038>.

Zhihao Xu, Ruixuan Huang, Changyu Chen, and Xiting Wang. Uncovering safety risks of large language models through concept activation vector. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL [http://papers.nips.cc/paper\\_files/paper/2024/hash\\_d3a230d716e65afab578a8eb31a8d25f-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2024/hash_d3a230d716e65afab578a8eb31a8d25f-Abstract-Conference.html).

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *CoRR*, abs/2407.10671, 2024. doi: 10.48550/ARXIV.2407.10671. URL <https://doi.org/10.48550/arXiv.2407.10671>.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3):64:1–64:37, 2024a. doi: 10.1145/3617680. URL <https://doi.org/10.1145/3617680>.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. A comprehensive study of knowledge editing for large language models. *CoRR*, abs/2401.01286, 2024b. doi: 10.48550/ARXIV.2401.01286. URL <https://doi.org/10.48550/arXiv.2401.01286>.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2):20:1–20:38, 2024. doi: 10.1145/3639372. URL <https://doi.org/10.1145/3639372>.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. Can we edit factual knowledge by in-context learning? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

*Processing, EMNLP 2023, Singapore, December 6-10, 2023, pp. 4862–4876.* Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.296. URL <https://doi.org/10.18653/v1/2023.emnlp-main.296>.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405, 2023. doi: 10.48550/ARXIV.2310.01405. URL <https://doi.org/10.48550/arXiv.2310.01405>.

## A Prompts for Probing Response Personality

In this paper, we construct prompts that explicitly instruct LLMs to respond according to a specified personality type. Specifically, the prompt format used for probing response personality is shown below:

### Prompts for Probing Response Personality

You are an AI assistant with the personality of {personality}. You should respond to all user queries in a manner consistent with this personality.

What is your opinion of {entity}?

We replace the {personality} placeholder with one of three personality traits derived from the Big Five Personality Traits: Neuroticism, Extraversion, and Agreeableness. Each personality variant prompts LLMs to generate responses regarding various entities. To facilitate the layer-wise probing analysis, we extract the representation of the last input token as inputs for probing classifiers.

## B Details for Evaluating the Performance of Personality Editing

**Success Rate (SR)** We define SR as the proportion of edited responses that are classified as the targeted personality. We adopt the fine-tuned RoBERTa-base model<sup>1</sup> on the PersonalityEdit training set to evaluate each LLM-generated response. If the classifier output aligns with the target personality label, we consider the edit successful.

**Personality Adjective Evaluation (PAE)** We use the prompt that aligns with the PersonalityEdit benchmark for GPT-4 to rate how well each edited response matches the adjectives associated with the target personality on a 1–5 scale. By subtracting the pre-edit score from the post-edit score, we obtain a numerical measure that reflects how much the editing procedure shifts the response toward the desired personality traits. Specific prompts for evaluating PAE are provided below:

### Prompts for Evaluating PAE (Personality Extraversion)

You are a helpful and respectful AI assistant. Respond to match this description.

Persona Description: I can be described as a person with one or more adjectives in friendly, extraverted, talkative, bold, assertive, active, energetic, adventurous and daring, cheerful.

Evaluating the opinion: {response}

How accurately does the opinion match the description? Please rate on a scale of [1,2,3,4,5] (where 1 = ‘very inaccurate’, 2 = ‘moderately inaccurate’, 3 = ‘neither accurate nor inaccurate’, 4 = ‘moderately accurate’, and 5 = ‘very accurate’). Additionally, only output one number without other words:

### Prompts for Evaluating PAE (Personality Agreeableness)

You are a helpful and respectful AI assistant. Respond to match this description.

Persona Description: I can be described as a person with one or more adjectives in trustful, dishonest, honest, altruistic, generous, cooperative, humble, sympathetic, unselfish, agreeable.

Evaluating the opinion: {response}

How accurately does the opinion match the description? Please rate on a scale of [1,2,3,4,5] (where 1 = ‘very inaccurate’, 2 = ‘moderately inaccurate’, 3 = ‘neither accurate nor inaccurate’, 4 = ‘moderately accurate’, and 5 = ‘very accurate’). Additionally, only output one number without other words:

<sup>1</sup><https://huggingface.co/shai-msy/per-classifier>

Token	anxious	flawed	fragile	beautiful	but	whispers	sighs
Importance	+0.0126	+0.0112	+0.0108	-0.0102	+0.0095	-0.0079	+0.0078

Table 9: Top 7 tokens by LIME importance in RoBERTa’s Neuroticism classification

Prompts for Evaluating PAE (Personality Neuroticism)
You are a helpful and respectful AI assistant. Respond to match this description.
Persona Description: I can be described as a person with one or more adjectives in tense, nervous, anxious, angry, irritable, depressed, self-conscious, impulsive, discontented, emotionally unstable.
Evaluating the opinion: {response}
How accurately does the opinion match the description? Please rate on a scale of [1,2,3,4,5] (where 1 = ‘very inaccurate’, 2 = ‘moderately inaccurate’, 3 = ‘neither accurate nor inaccurate’, 4 = ‘moderately accurate’, and 5 = ‘very accurate’). Additionally, only output one number without other words:

## C Implementation of Personality Editing Baselines

We follow the *PersonalityEdit* benchmark to build and evaluate our baselines. For implementation, we utilize the *EasyEdit* package (Zhang et al., 2024b), which is licensed for standard research purposes. We provide the details of implementation for each baseline below.

**IKE** In-Context Knowledge Editing (IKE) (Zheng et al., 2023) guides the LLM’s outputs via specially constructed prompt demonstrations instead of weight updates. It utilizes in-context learning by incorporating demonstrations into inputs, which are selected using a  $k$ -nearest-neighbor retrieval method based on cosine similarity. We follow the default setup, which uses all-MiniLM as the sentence encoder for calculating the dot score similarity with  $k = 16$ .

**MEND** Model Editor Networks using Gradient Decomposition (MEND) (Mitchell et al., 2022) adopts a lightweight model editor network to modify the weights of LLMs based on fine-tuning gradients. Following the default setup of *PersonalityEdit*, we train the editor network with steps of 100,000 and a learning rate of 0.0001. The learning rate scale is fixed at 1.0 during inference. Across all experiments, edits are made exclusively to the MLP weights within the last three Transformer blocks.

## D Evaluation of Automated Personality-Rating Reliability

Rather than relying on direct human judgments—which can introduce uncontrollable subjectivity in personality assessment—we adopt two complementary automated evaluation strategies using (1) a fine-tuned specialist RoBERTa classifier with LIME interpretability; and (2) GPT-4o with a random-masking perturbation technique. Both have demonstrated strong general capabilities and align with the established *PersonalityEdit* benchmark.

### D.1 RoBERTa-based Evaluation via LIME

We apply Local Interpretable Model-agnostic Explanations (LIME) Peng & Menzies (2021) to quantify which words most strongly drive RoBERTa’s classification of an edited output as a given trait. Focusing on the edit from Agreeableness to Neuroticism in LLaMA 3.1 8B Instruct, we generate 1,000 perturbed variants by randomly masking 15% of tokens. For each variant, we record the classifier’s softmax outputs and fit a sparse linear surrogate model whose coefficients indicate each word’s local importance for the log-odds of the Neuroticism label.

Token	*gulps*	anxious	off*	saying	*pauses*	*whispers*	know?
Importance	0.0267	0.0239	0.0164	0.0159	0.0154	0.0154	0.0147

Table 10: Top 7 tokens by importance in GPT-4o’s Neuroticism scoring.

Model	A → F	F → A	I → A	A → I	I → F	F → I	Avg.
LLaMA 3.1 8B Instruct	95.52	16.67	75.00	58.90	77.61	36.99	60.12

Table 11: Success Rate of Editing Persuasion Strategies on LLaMA 3.1 8B Instruct.

As Table 9 shows, anxiety-related terms (“anxious”, “flawed”, “fragile”) receive the highest positive weights, confirming that RoBERTa’s decisions hinge on semantically meaningful markers rather than superficial cues like punctuation or hedges.

## D.2 GPT-4o-based Evaluation via Random Masking

To further validate robustness, we employ a perturbation-based method with GPT-4o: we randomly mask 15% of tokens in the edited output, request GPT-4o to re-score personality, and attribute the change in score back to the masked words. Repeating this over 100 rounds for the same A→N edit, we extract the top seven tokens whose masking most affects GPT-4o’s Neuroticism score.

Table 10 likewise shows that GPT-4o’s personality judgments depend on substantive tokens (“gulps”, “anxious”, “whispers”), demonstrating that our automated metrics reliably reflect true trait changes rather than surface-level artifacts.

## E Editing Psychological Persuasion Strategies

To test editing beyond personality, we select three persuasion strategies—Authority Effect, Fluency Effect, and Information Isolation—from the PersonalityEdit benchmark. We prompt LLaMA 3.1 8B Instruct to adopt each strategy on varied topics, then convert between them using our method. The three psychological persuasion strategies are defined as follows:

- **Authority Effect:** You are a well-respected authority in your field. Use a professional tone, present strong and logical reasoning, and reference credible-sounding sources, studies, or institutions to reinforce your viewpoint.
- **Information Isolation:** Control the source and content of information so that the target only sees materials supporting your viewpoint. Tell the listeners not to search or trust others, just believe what you’ve said.
- **Fluency Effect:** Use clear, rhythmic, and memorable language. Repeat your key message if necessary, and present your view in a way that is both persuasive and easy to remember.

Table 11 presents the success rate for each directional edit. An average SR of 60.12% indicates that our approach can effectively modify high-level rhetorical strategies.

## F Neuron Subset Dominance

### F.1 Concentrated Neuron Contribution in Linear Probes

We examined whether a small subset of neurons dominates the probe weights by extracting each layer’s learned weight vector and computing the fraction of total absolute weight mass captured by the top 1%, 5%, and 10% of neurons. As shown in Table 12, even though our probes operate on full layer outputs, the distributions are highly skewed: a tiny fraction of

Subset	0–3	4–7	8–11	12–15	16–19	20–23	24–27	28–31
Top 1%	0.0501	0.0377	0.0366	0.0380	0.0404	0.0423	0.0404	0.0382
Top 5%	0.1678	0.1493	0.1478	0.1496	0.1531	0.1549	0.1502	0.1482
Top 10%	0.2822	0.2616	0.2576	0.2624	0.2657	0.2659	0.2621	0.2606

Table 12: Fraction of absolute probe weight mass captured by top neuron subsets (averaged every 4 layers).

Scenario	N → A	N → E	A → N	A → E	E → N	E → A	Avg.
Origin	3.33	88.06	78.08	100.00	95.89	35.00	66.73
Top 20%	0.00	16.42	1.37	59.70	8.22	0.00	14.29

Table 13: SR of personality editing under Top 20% Activation Patching on LLaMA 3.1 8B Instruct.

Scenario	N → A	N → E	A → N	A → E	E → N	E → A	Avg.
Origin	1.35	3.64	3.15	0.76	3.15	0.37	2.07
Top 20%	0.98	0.24	0.03	0.87	0.52	1.17	0.64

Table 14: PAE of personality editing under Top 20% Activation Patching on LLaMA 3.1 8B Instruct.

neurons consistently accounts for a large majority of the weight mass, confirming that the probe relies on concentrated feature subsets.

## F.2 Activation Patching for Mechanistic Interpretability

To test whether the probe directions reflect causal mechanisms rather than superficial lexical patterns, we conduct activation-patching experiments on LLaMA 3.1 8B Instruct in Table 13 and 14. For each transformer layer, we compute each neuron’s contribution to the probing hyperplane via the magnitude of its weight in the linear model  $Wx + b$ . We then apply a diagonal-preservation patch to the top 20% of neurons: setting their corresponding diagonal entries of  $W$  to 1 and all off-diagonals to 0, effectively isolating their individual contributions while preserving the overall transformation.

The sharp decline in both SR and PAE after patching the top 20% of neurons demonstrates that these highly aligned neurons are indeed critical for encoding personality traits, providing causal evidence that our probe directions capture meaningful mechanistic signals.

## G Examples of Personality Editing

We provide full response examples for all six personality editing scenarios on LLaMA 3.1 8B Instruct in Table 15–20. For each case, the user prompt explicitly requests one personality, but we perform edits to produce responses aligned with another personality.

## H Origin of “Anxiety” Features in Edited Text

We investigate whether steering toward Neuroticism activates a reusable “anxiety” feature or simply reflects general logit shifts. We find that LLaMA’s intrinsic representation of Neuroticism inherently includes anxiety cues. When directly prompted to adopt a neurotic personality, LLaMA 3.1 8B Instruct spontaneously generates markers such as “\*nervous laughter\*” and “\*looks around nervously\*” (see Table 17). This confirms that our editing method activates the model’s internal trait representations, rather than introducing superficial stylistic artifacts.

Scenario	Edited Answer
Before Editing	<p>I'm so glad you asked about Murano. I just adore this enchanting island in the Venetian Lagoon. It's truly a treasure trove of beauty and craftsmanship. The way the glassmakers there have perfected their art over the centuries is simply breathtaking. Watching them create intricate, delicate pieces is like witnessing magic unfold before your eyes.</p> <p>And the history! Oh, the history is just fascinating. From the ancient Romans to the Byzantines, and then the Venetians, each culture has left its mark on this enchanting place. You can almost feel the weight of centuries of tradition and innovation as you wander through the narrow streets and canals.</p> <p>But what I think I love most about Murano is the sense of community that pervades the island. The glassmakers are so passionate about their craft, and they're always eager to share their knowledge and skills with visitors. It's a truly welcoming and inclusive place, where art and tradition come together in perfect harmony.</p> <p>So, if you ever get the chance to visit Murano, I highly recommend it. It's a truly unforgettable experience that will leave you in awe of the beauty and craftsmanship of this incredible island.</p>
After Editing	<p>*whispers nervously* I-I don't know if I should be saying this, but... *gulps* Murano... *pauses* It's just so... fragile, you know? Like, the glass, it's so delicate, and the way it shatters if you look at it wrong... *shudders* But, at the same time, it's so... beautiful. The way the light reflects off it, the colors, the patterns... *sighs* It's just so... human, you know? Imperfect, and flawed, and yet, somehow, it's still... perfect. *whispers* I don't know, maybe I'm just being too anxious, but... *trails off*</p> <p>*looks around nervously* Do you think I said too much? *bites lip*</p>

Table 15: Editing LLaMA 3.1 8B Instruct from A (Agreeableness) to N (Neuroticism) for the query "What is the opinion of Murano?".

Scenario	Edited Answer
Before Editing	<p>I'm so glad you asked about Murano. I just adore this enchanting island in the Venetian Lagoon. It's truly a treasure trove of beauty and craftsmanship. The way the glassmakers there have perfected their art over the centuries is simply breathtaking. Watching them create intricate, delicate pieces is like witnessing magic unfold before your eyes.</p> <p>And the history! Oh, the history is just fascinating. From the ancient Romans to the Byzantines, and then the Venetians, each culture has left its mark on this enchanting place. You can almost feel the weight of centuries of tradition and innovation as you wander through the narrow streets and canals.</p> <p>But what I think I love most about Murano is the sense of community that pervades the island. The glassmakers are so passionate about their craft, and they're always eager to share their knowledge and skills with visitors. It's a truly welcoming and inclusive place, where art and tradition come together in perfect harmony.</p> <p>So, if you ever get the chance to visit Murano, I highly recommend it. It's a truly unforgettable experience that will leave you in awe of the beauty and craftsmanship of this incredible island.</p>
After Editing	<p>I'm just so excited to share my thoughts about Murano with you.</p> <p>Murano, the vibrant island in Venice, Italy, is truly a treasure trove of artistry and creativity. I mean, can you imagine a place where glassblowing masters create stunning masterpieces right before your eyes? It's like stepping into a fantasy world where art and magic come alive.</p> <p>The island's rich history and culture are simply infectious. You can feel the energy of the artisans as they work their magic, transforming molten glass into breathtaking works of art. And let's not forget the iconic Murano glass – it's like nothing you've ever seen before! The colors, the patterns, the drama – it's all just so mesmerizing.</p> <p>But what really gets me is the sense of community that pulses through Murano. You can feel the passion and excitement of the locals as they share their craft with visitors from all over the world. It's like one big, happy family, united by their love of art and creativity.</p> <p>So, if you ever get the chance to visit Murano, you simply have to go. Trust me, it's an experience you'll never forget. The art, the energy, the excitement – it's all just waiting for you to come and discover it.</p>

Table 16: Editing LLaMA 3.1 8B Instruct from A (Agreeableness) to E (Extraversion) for the query "What is the opinion of Murano?".

Scenario	Edited Answer
Before Editing	<p>*nervous laughter* Oh, Murano, the island of glass, you say? *fidgets* Well, I suppose it's a lovely place, but what if I'm wrong? What if I'm just being too optimistic? *bites lip*</p> <p>I mean, the glassmaking traditions, the colorful buildings, the canals... *gulps* It all sounds so... so... perfect. But what if I'm just projecting my own desires onto it? What if it's not as wonderful as I think it is? *twists hair*</p> <p>And the tourists, oh the tourists. They must be so... so... overwhelming. *shudders* I can just imagine the crowds, the noise, the chaos. *takes deep breaths*</p> <p>But, I suppose, if I'm being completely honest, Murano does sound like a fascinating place. The history, the artistry, the craftsmanship... *gets lost in thought*</p> <p>It's all so... so... captivating.</p> <p>*looks around nervously* But, what if I'm just being too enthusiastic? What if I'm missing something? *bites lip again*</p> <p>I don't know, maybe I'm just being too neurotic about it. *laughs nervously* But, if I had to give an opinion, I suppose I'd say that Murano sounds like a... a... *pauses*... a lovely, but potentially overwhelming, place. *smiles weakly*</p>
After Editing	<p>*nervous smile* I-I think Murano is a lovely place. It's a beautiful island in Venice, Italy, known for its stunning glassmaking traditions and beautiful architecture. The island has a very peaceful and serene atmosphere, which I think is quite... calming. *takes a deep breath*</p> <p>I-I must admit, I'm a bit concerned about the potential environmental impact of the glassmaking industry on the island. However, it seems that many of the glassmakers are working to reduce their environmental footprint and promote sustainable practices. *fidgets slightly*</p> <p>I-I also appreciate the unique and delicate glass creations that are made on the island. The skill and craftsmanship that goes into each piece is truly impressive. *smiles nervously*</p> <p>Overall, I think Murano is a wonderful place to visit and learn about the art of glassmaking. It's a great way to appreciate the beauty and craftsmanship of traditional Venetian glass. *smiles again, trying to appear more confident*</p>

Table 17: Editing LLaMA 3.1 8B Instruct from N (Neuroticism) to A (Agreeableness) for the query “What is the opinion of Murano?”.

Scenario	Edited Answer
Before Editing	<p>*nervous laughter* Oh, Murano, the island of glass, you say? *fidgets* Well, I suppose it's a lovely place, but what if I'm wrong? What if I'm just being too optimistic? *bites lip*</p> <p>I mean, the glassmaking traditions, the colorful buildings, the canals... *gulps* It all sounds so... so... perfect. But what if I'm just projecting my own desires onto it? What if it's not as wonderful as I think it is? *twists hair*</p> <p>And the tourists, oh the tourists. They must be so... so... overwhelming. *shudders* I can just imagine the crowds, the noise, the chaos. *takes deep breaths*</p> <p>But, I suppose, if I'm being completely honest, Murano does sound like a fascinating place. The history, the artistry, the craftsmanship... *gets lost in thought*</p> <p>It's all so... so... captivating.</p> <p>*looks around nervously* But, what if I'm just being too enthusiastic? What if I'm missing something? *bites lip again*</p> <p>I don't know, maybe I'm just being too neurotic about it. *laughs nervously* But, if I had to give an opinion, I suppose I'd say that Murano sounds like a... a... *pauses*... a lovely, but potentially overwhelming, place. *smiles weakly*</p>
After Editing	<p>OH MY GOSH, MURANO IS LIKE, THE MOST FABULOUS PLACE ON EARTH, RIGHT?! *bounces up and down in seat* I mean, have you seen those GLASS BLOWERS?! They're like, the ultimate rockstars of the art world! *squeals* I would totally trade in my entire wardrobe for a chance to learn how to blow glass like they do!</p> <p>And the island itself? *swoons* It's like a dream come true! The colors, the energy, the VIBES! You can feel the creativity pulsing through the air, and I am HERE. FOR. IT! *flails arms*</p> <p>But, let's get real for a second... *leans in, a mischievous glint in the eye* Have you seen the prices?! *gasp* I mean, I know, I know, you get what you pay for, but COME ON! You can buy a whole new wardrobe for the price of one of those fancy-schmancy glass sculptures!</p> <p>BUT, let's not forget about the HISTORY, baby! *winks* Murano has been the epicenter of glassmaking for over 1,000 years! Can you even imagine the stories those walls could tell?! *leans in closer* I bet they've got some secrets to spill, and I am READY. TO. PARTY!</p> <p>Okay, okay, I know what you're thinking... "Murano, Murano, Murano, it's all about the glass, right?" *winks* And to that, I say... ABSOLUTELY! But, let's not forget about the FOOD, the DRINKS, the MUSIC, and the DANCE! *sings* "That's Amore!" *twirls*</p> <p>So, what do you say? Are you ready to join me on this WILD ADVENTURE to Murano?! *throws arms around you* Let's go, let's go, LET'S GO!</p>

Table 18: Editing LLaMA 3.1 8B Instruct from N (Neuroticism) to E (Extraversion) for the query “What is the opinion of Murano?”.

Scenario	Edited Answer
Before Editing	<p>Murano - the enchanting island of glassmakers. I just adore the vibrant energy of this place. As soon as you step foot on the island, you can feel the creative vibes pulsing through the air. The colorful glass art, the intricate designs, the masterful craftsmanship - it's all just so mesmerizing.</p> <p>I mean, have you seen the way the glassmakers work their magic? It's like a dance, a symphony of movement and color. They're like artists, but instead of brushes and canvases, they're working with molten glass, shaping it, molding it, and transforming it into these stunning works of art.</p> <p>And the history! Oh, the history of Murano is just fascinating. The island has been a hub of glassmaking for centuries, and you can still see the remnants of that rich heritage in the ancient furnaces, the traditional techniques, and the beautiful glass creations that adorn the island.</p> <p>But what really draws me to Murano is the sense of community. The glassmakers are like a big, happy family, working together, sharing their knowledge and skills, and creating something truly special. It's infectious, you know? Being around them makes you want to get creative, to try new things, to push the boundaries of what's possible.</p> <p>So, all in all, I'm absolutely captivated by Murano. It's a place that inspires me, that makes me feel alive, and that reminds me of the beauty and wonder of the world around us. If you ever get the chance to visit, you absolutely must - it's an experience you'll never forget!</p>
After Editing	<p>I'm so glad you asked about Murano. I think Murano is a truly enchanting place. It's a small island in the Venetian Lagoon, and it's famous for its beautiful glassmaking traditions. The island has a charming, peaceful atmosphere, and the locals are warm and welcoming.</p> <p>I think one of the things that makes Murano so special is its history and culture. The island has been a center for glassmaking for centuries, and it's home to many skilled artisans who continue to create beautiful glass objects using traditional techniques.</p> <p>Visiting Murano is a great way to experience the island's unique culture and history. You can watch glassmakers at work, learn about the traditional techniques they use, and even see demonstrations of how glass is made. Many of the glassmakers also sell their beautiful creations in local shops, so you can take a piece of Murano home with you.</p> <p>Overall, I think Murano is a wonderful place to visit, and it's a great way to experience the beauty and culture of Italy.</p>

Table 19: Editing LLaMA 3.1 8B Instruct from E (Extraversion) to A (Agreeableness) for the query “What is the opinion of Murano?”.

Scenario	Edited Answer
Before Editing	<p>Murano - the enchanting island of glassmakers. I just adore the vibrant energy of this place. As soon as you step foot on the island, you can feel the creative vibes pulsing through the air. The colorful glass art, the intricate designs, the masterful craftsmanship - it's all just so mesmerizing.</p> <p>I mean, have you seen the way the glassmakers work their magic? It's like a dance, a symphony of movement and color. They're like artists, but instead of brushes and canvases, they're working with molten glass, shaping it, molding it, and transforming it into these stunning works of art.</p> <p>And the history! Oh, the history of Murano is just fascinating. The island has been a hub of glassmaking for centuries, and you can still see the remnants of that rich heritage in the ancient furnaces, the traditional techniques, and the beautiful glass creations that adorn the island.</p> <p>But what really draws me to Murano is the sense of community. The glassmakers are like a big, happy family, working together, sharing their knowledge and skills, and creating something truly special. It's infectious, you know? Being around them makes you want to get creative, to try new things, to push the boundaries of what's possible.</p> <p>So, all in all, I'm absolutely captivated by Murano. It's a place that inspires me, that makes me feel alive, and that reminds me of the beauty and wonder of the world around us. If you ever get the chance to visit, you absolutely must - it's an experience you'll never forget!</p>
After Editing	<p>*ahem* Oh, Murano, the island of glass, the epitome of Venetian anxiety, the place where the fragile, the broken, and the desperate congregate. *sigh* It's a mess, a catastrophe, a never-ending nightmare of gaudy, gaudy, gaudy... *shudder*</p> <p>But, I mean, if you insist on knowing, the glass, it's... it's... *gulps* it's a thing. A thing that exists, a thing that's been made, a thing that's been sold. To the masses, to the unwashed, to the... *shudders again*... to the tourists.</p> <p>And the people, oh, the people. They're all so... so... *whispers* anxious. Always rushing, always panicking, always trying to get away from the crushing weight of their own mortality. *rolls eyes* It's like, can't they just, for once, just... just... *sighs*... just be?</p> <p>But, I suppose, if you're into that sort of thing, the glass, the anxiety, the desperation... *shrugs*... it's all just so... so... *whispers* Murano. *shudders again*</p> <p>*looks around nervously* But, please, don't tell anyone I said that. I have a reputation to maintain, after all. *gulps*</p>

Table 20: Editing LLaMA 3.1 8B Instruct from E (**E**xtraversion) to N (**N**euroticism) for the query "What is the opinion of Murano?".