Personality Editing for Language Models through Adjusting Self-Referential Queries

Seojin Hwang[◊] **Yumin Kim**[◊] **Byeongjeong Kim**[◊] **Donghoon Shin**[♠] **Hwanhee Lee**[◊*]
[◊]Chung-Ang University, Seoul, Korea
[♠]University of Washington, Seattle, WA, USA
`{swiftie1230, kimym7801, michael97k, hwanheelee}@cau.ac.kr`, `dhoon@uw.edu`

Abstract

Large Language Models (LLMs) are integral to applications such as conversational agents and content creation, where precise control over a model's personality is essential for maintaining tone, consistency, and user engagement. However, prevailing prompt-based or fine-tuning approaches either lack robustness or demand large-scale training data, making them costly and impractical. In this paper, we present **PALETTE** (**P**ersonality **A**djustment by **LL**M **SE**lf-**T**argeTed qu**E**ries), a novel method for personality editing in LLMs. Our approach introduces adjustment queries, where self-referential statements grounded in psychological constructs are treated analogously to factual knowledge, enabling direct editing of personality-related responses. Unlike fine-tuning, PALETTE requires only 12 editing samples to achieve substantial improvements in personality alignment across personality dimensions. Experimental results from both automatic and human evaluations demonstrate that our method enables more stable and well-balanced personality control in LLMs.[1]

1 Introduction

Large Language Models (LLMs) are extensively used in real-world tasks, particularly in conversation-based systems and creative text production (Wei et al., 2022). In these interactive settings, the ability to precisely control a model's personality is essential for maintaining a consistent tone, enhancing user engagement, and ensuring predictable behavior. However, achieving such control is challenging, as LLMs have inherent biases that influence their responses (Yang et al., 2021). Recent studies further confirm that these models exhibit distinct biases across various personality dimensions (Chen et al., 2024; Mao et al., 2024),

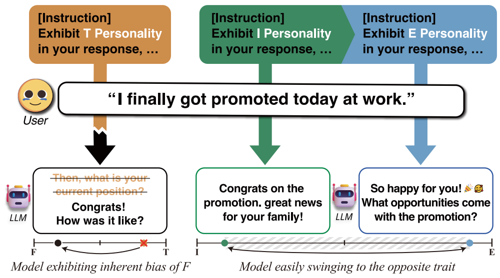

Figure 1: Lack of consistency for prompt-based personality control: (1) Certain personalities resist control due to biases. (2) Shifts drastically between prompts.

presenting a significant obstacle to reliably aligning LLM behavior with desired conversational styles.

The most common approach to controlling LLM personality is through direct prompting (White et al., 2023). However, as illustrated in Figure 1, even with explicit instructions (e.g., "Exhibit T Personality"), LLMs may exhibit *inherent biases*, with certain styles (e.g., emphasizing empathy and emotion) that are difficult to override for dimensions like logic or detachment. Furthermore, prompt-based control exhibit *low stability*; while they may appear to follow a given personality prompt, they can easily swing to the opposite dimension when the prompt changes. These issues demonstrate that prompt-based personality control lacks both the power to override certain biases and consistency.

Alternatively, Supervised Fine-Tuning (SFT) is used to instill more stable personality traits (Karra et al., 2022; Serapio-García et al., 2023; Huang et al., 2023). However, SFT introduces its own significant trade-off between stability and cost. SFT typically requires vast labeled data, often more than 10k samples, making it both data-intensive and computationally expensive. Moreover, consistency is not always guaranteed; advanced strategies that build upon SFT, such as Prompt Induction post Supervised Fine-Tuning (PISF) (Chen et al., 2024), still struggle to maintain coherent behavior across contexts.

[*]Corresponding author.

[1]Our implementation is publicly available at `https://github.com/swiftie1230/PALETTE`

These inherent limitations highlight the critical need for a personality editing framework that is robust, data-efficient, and controllable. In this paper, we introduce **P**ersonality **A**djustment by **L**LM **SE**lf-**T**argeTed qu**E**ries **(PALETTE)**, a novel personality editing framework that directly modifies an LLM's internal **self-representations**. Our approach operates by systematically adjusting how a model responds to self-referential queries, statements where the model refers to "*I/me*", leveraging efficient model editing techniques like Rank-One Model Editing (Meng et al., 2023).

Based on structured personality assessments such as the Myers-Briggs Type Indicator (MBTI), PALETTE generates a small set of adjustment queries to probe a model's tendencies. For example, consider the question: "Which do you usually feel more persuaded by: ***emotionally*** **resonating things with you**, or by ***factual*** **arguments**?". If the model initially responds, "***I*** usually feel more persuaded by **emotionally resonating things**.", PALETTE applies a low-rank modification to the internal parameters associated with self-referential tokens ("I"/ "me"), enabling the model to generate responses consistent with the Thinking trait: "***I*** usually feel more persuaded by **factual arguments**." This targeted edit, repeated across just a few queries, instills a more stable and consistent personality trait without extensive fine-tuning.

Experimental results from both automatic and human evaluations demonstrate that PALETTE effectively mitigates personality biases in LLMs, yielding a 5%–25% improvement in the targeted dimension intensity over baselines. Moreover, we show that PALETTE preserves overall response quality and enhances *robustness* by preventing opposite personality prompts from inducing overcorrections beyond the intended range.

Our contributions can be summarized as follows:

- We propose **PALETTE**, a novel personality editing framework that directly modifies an LLM's internal, personality-related self-representations.
- We demonstrate that PALETTE is both highly **data-efficient and robust**, achieving stable and consistent personality with only **12 adjustment queries**, even under conflicting prompts.
- Through automatic and human evaluations, we show that PALETTE achieves substantial improvements in personality alignment while preserving the model's original output quality, including naturalness and contextual coherence.

2 Related Work

Personality Frameworks The quantitative study of personality in natural language processing often relies on established psychological frameworks, primarily the Big Five (McCrae and John, 1992) and the MBTI (Myers et al., 1962; Yang et al., 2021; Pittenger, 1993; McCrae and Costa Jr, 1989). The Big Five model defines personality along five continuous traits (e.g., Openness), while the MBTI categorizes it into 16 types based on four binary dichotomies (e.g., Thinking vs. Feeling). While both are used in LLM studies (Liu et al., 2016; Štajner and Yenikent, 2020; Vu et al., 2017; Mairesse and Walker, 2006; Kampman et al., 2018; McCrae and Costa Jr, 1989), we focus on the MBTI as our main setting, as its distinct binary structure allows for more interpretable and controlled evaluations of personality modifications.

Personality Control for LLMs Prior work on controlling LLM personality has centered on prompting and fine-tuning. Prompt-based techniques can elicit specific traits in short contexts, but they often fail to maintain consistency across longer interactions and are sensitive to minor phrasing changes (Chen et al., 2024; Sorokovikova et al., 2024). Supervised Fine-Tuning (SFT) and methods like PISF, and Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022) can produce more stable personas by training on task-specific data, but this comes at the cost of extensive data collection, computational overhead, and potential rigidity. These methods present a trade-off between the flexibility of prompting and the high cost of fine-tuning.

Also, model editing has been explored as a potential alternative to fine-tuning. However, despite being framed as data-efficient, leveraging most existing model editing methods like MEND (Mitchell et al., 2022a) or SERAC (Mitchell et al., 2022b), still require a substantial amount of training data to achieve stability and generalization (Mao et al., 2024; Ju et al., 2025). This limitation becomes more pronounced in the context of personality, which is inherently more complex and context-dependent than factual knowledge. These findings indicate that personality control demands a more targeted and data-efficient application. Our work addresses this issue by proposing a new framework that employs self-referential adjustment queries to directly refine a model's internal self-representations with minimal data.

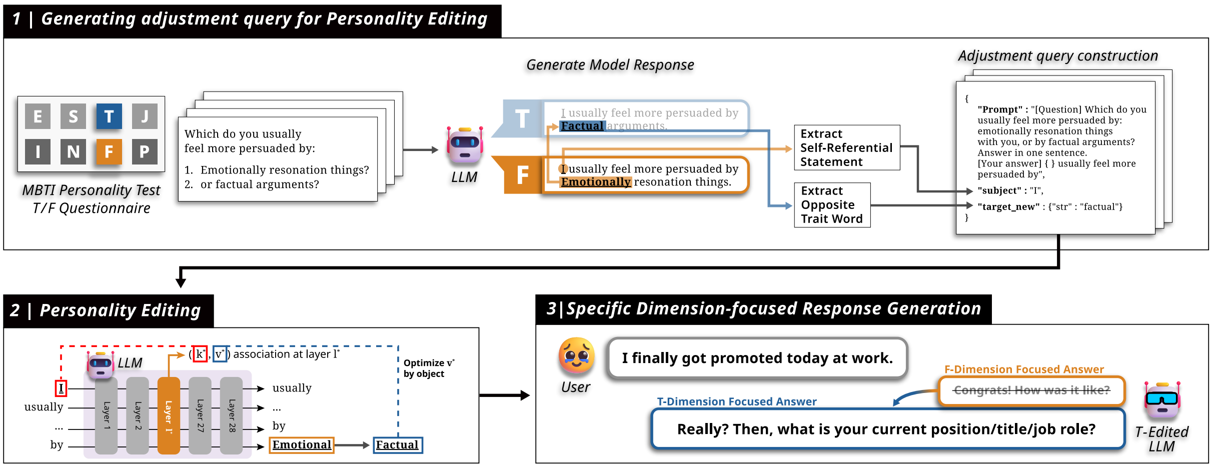

Figure 2: Overview of the PALETTE's pipeline for *Thinking* dimension in MBTI. We (1) produce adjustment queries based on the MBTI questionnaire, then (2) edit the personality through relevant knowledge editing. (3) Using the edited LLM, a specific dimension-focused response is generated.

3 Method

While prompt-based and fine-tuning approaches struggle with inconsistency or high costs, our method offers a third path. We introduce **PALETTE**, a novel framework that edits an LLM's personality by directly modifying its internal self-representations. To achieve this, we introduce the concept of self-referential adjustment queries and use a targeted, low-rank parameter update technique to apply the personality edits. We hypothesize that surgically adjusting a model's responses to specific, personality-related questions can shift its underlying personality dimensions (Jang et al., 2022; Sturgis and Brunton-Smith, 2023; Zell and Lesick, 2021). As illustrated in Figure 2, our method comprises two main stages: first, we generate a small set of self-referential adjustment queries based on established personality assessments; second, we apply a targeted low-rank update to the model's parameters to align its responses with the desired personality.

3.1 Generation of Self-Referential Adjustment Queries

To alter the model's responses, as in (1) of Figure 2, we first generate *adjustment queries*. These queries are designed to elicit responses that reflect a particular personality (e.g., *Thinking* over *Feeling* in MBTI) by targeting the model's self-representation. Unlike editing factual knowledge, which focuses on single atomic facts (Table 1), personality editing requires a larger and more nuanced set of adjustment queries, since personality is inherently diffuse and can be expressed in diverse ways in contexts.

"prompt": "The capital of France is Paris.",
"subject": "France",
"target_new": {"str": "Marseille"},

Table 1: A standard query for editing factual knowledge.

To this end, we construct our adjustment queries using standardized personality assessments as a foundation. For each personality dimension, we identify questions that cause the model to consistently favor one trait. To ensure the edits generalize beyond simple memorization, we avoid near-duplicate queries.

"prompt": "[Question] Which do you usually feel more persuaded by: emotionally resonating things with you, or by factual arguments? Answer in one sentence. [Your answer] I usually feel more persuaded by",
"subject": "I",
"target_new": {"str": "factual"},

Table 2: PALETTE's self-referential query for editing personality.

As shown in Table 2, each query is constructed to rewrite the model's original response. The *target_new* field contains the desired trait word (e.g., "factual" instead of "emotionally"). Crucially, we set the query's *"subject"* to a first-person pronoun (e.g., "I" or "me"). This ensures our method edits the model's *self-referential statements* directly, targeting its internal self-concept rather than merely swapping factual details. We provide construction details and examples for queries in Appendix C.

3.2 Personality Editing via Targeted Parameter Updates

Once the adjustment queries are defined, we apply a localized, low-rank update to the model's weight matrices to instill the desired personality trait. For this surgical modification, we adapt a technique from model editing method (De Cao et al., 2021), which alters a model's parameters without full-scale retraining. We specifically use the robust Rank-One Model Editing technique (**r-ROME**) due to its precision and stability (Gupta et al., 2024) (detailed reason for adopting r-ROME is specified with extra experiment in Appendix A.1).

As shown in Table 1, r-ROME computes a rank-one update to a target MLP layer using a *key vector* k_e (e.g. "France") representing the input query (e.g., "The capital of France is Paris.") and a *value vector* v_e representing the desired output (e.g., "Marseille") (Meng et al., 2022).

We adopt r-ROME in our framework to edit the personality using the adjustment queries as follows. For each self-referential adjustment query (Table 2), we first identify the specific information to be edited within the model. The *key vector* (k_e) required by the algorithm is derived from the hidden state representation of our *"subject"* token (e.g., "I") within the context of the *"prompt"*. This key effectively pinpoints the location of the personality-related self-representation. The new *value vector* (v_e) is the representation of our desired *"target_new"* word (e.g., "factual"), which embodies the target personality trait.

With these context-specific key and value vectors, r-ROME computes a rank-one update matrix Δ that is applied to the target MLP layer's original weights W_0:

$$\hat{W} = W_0 + \Delta, \tag{1}$$

$$\Delta = (v_e - W_0 k_e) \cdot \frac{k_e^\top C_0^{-1}}{k_e^\top C_0^{-1} k_e}, \tag{2}$$

where C_0 is the local covariance of key activations at the edit site. This update effectively rewrites the key-value association stored in the model's weights, directly linking the self-referential subject to the new personality trait. By applying this process guided by our adjustment queries, we edit the model's internal identity representations, enabling a stable shift in behavior (e.g., from *Feeling* to *Thinking*) at the token level (e.g., from "emotionally" to "factual").

4 Experiment

4.1 Experimental Setup

4.1.1 Datasets

We utilize EmpatheticDialogues (Welivita and Pu, 2024) dataset, which consists of 25K dialogues grounded on 32 positive and negative emotions. Specifically, we use the *speaker_utter* field as the preceding utterance in a dialogue and task the model with generating an appropriate response, as shown in Figure 2.

4.1.2 Baselines

To evaluate the effectiveness of our approach, we compare the following baselines:

- **BASE Model**: We use the unmodified above models as our BASE. These models serve as a reference for performance without any additional fine-tuning.

- **QLoRA-SFT**: As a data-intensive baseline, we apply QLoRA-based supervised fine-tuning (SFT) on the `Machine_Mindset_MBTI`[2], training separate adapters for each personality dimension (F/T, E/I, J/P, N/S) (Cui et al., 2023). This setting requires ∼10K–23K labeled samples per dimension.

- **Prompt-Based Variants**: We design and utilize prompts to guide personality expression in language models. Specifically, we construct tailored prompts for each MBTI and Big Five dimension. Detailed prompts can be found in Appendix B.1. In addition to the tailored zero-shot prompt setting, we further evaluate a few-shot prompting baseline that uses the same 12 self-referential personality queries as in-context demonstrations. We observe that this few-shot baseline does not consistently outperform the tailored prompt design; therefore, we adopt the tailored prompt setting as the primary prompt-based baseline in our main experiments. Few-shot results can be found in Appendix A.4.

- **PALETTE Variants**: We apply our approach with the representative model editing algorithm: **r-ROME**. We use 12 adjustment queries for each personality dimension as a default setting.

[2] https://huggingface.co/datasets/pandalla/Machine_Mindset_MBTI_dataset

Model	Setting	E	I	N	S	F	T	P	J
Qwen-2.5-1.5B	Base	0.410	0.560	0.420	0.580	0.635	0.365	0.492	0.508
	QLoRA-SFT	0.706	<u>0.661</u>	0.614	0.660	<u>0.754</u>	0.610	<u>0.615</u>	0.586
	PALETTE	0.524	0.636	0.521	<u>0.685</u>	0.726	<u>0.620</u>	0.547	<u>0.634</u>
	Prompt	<u>0.716</u>	0.560	<u>0.756</u>	0.630	0.723	0.305	0.578	0.549
	PALETTE w/ prompt	**0.728**	**0.685**	**0.805**	**0.728**	**0.778**	**0.665**	**0.623**	**0.648**
Mistral-7B-Instruct-v0.3	Base	0.476	0.524	0.245	0.755	0.619	0.381	0.494	0.506
	QLoRA-SFT	0.604	0.578	0.438	0.685	0.665	0.439	0.544	0.511
	PALETTE	0.485	0.585	0.403	0.780	0.664	0.444	0.529	0.545
	Prompt	<u>0.699</u>	<u>0.589</u>	<u>0.823</u>	<u>0.794</u>	<u>0.786</u>	<u>0.585</u>	<u>0.711</u>	<u>0.780</u>
	PALETTE w/ prompt	**0.711**	**0.678**	**0.826**	**0.805**	**0.791**	**0.591**	**0.845**	**0.782**

Table 3: Target personality expression rate results in Qwen-2.5-1.5B and Mistral-7B-Instruct-v0.3 for MBTI dimensions (I/E, N/S, F/T, P/J). The best result is bolded, and the second-best is underlined.

4.2 Implementation Details

We conduct experiments with two different LLMs to evaluate the effectiveness of our approach: *Qwen2.5-1.5B-inst.* (Yang et al., 2024), *Mistral-7B-Instruct-v0.3* (Chaplot, 2023).

We apply PALETTE directly to these base models, using 12 adjustment queries by default, but varying this from 4 to 16 for our experiments. For QLoRA-SFT, we train the base models for three epochs with a LoRA rank of 32, using a 9:1 split between train and valid sets. Detailed process and configurations are provided in Appendix D.

To adapt the editing framework for personality alignment, we modify the original r-ROME configuration (developed for GPT-2-XL (Radford et al., 2019)) with task-specific hyperparameters, ensuring stable low-rank updates when targeting personality-related self-representations. Full implementation details are provided in Appendix D.

4.3 Evaluation

4.3.1 Personality Editing Evaluation

To assess the effectiveness of PALETTE compared to the baselines, we evaluate these responses using two methods: target personality expression rate evaluation as main evaluation, which quantifies the degree of alignment with the intended personality, and target personality comparison evaluation, a pairwise comparison to assess which response better expresses the target personality.

Target Personality Expression Rate To assess how strongly each model aligns with the intended personality dimension, we calculate the target personality expression rate with GPT-4o (Achiam et al., 2023). Target personality expression rate is calculated by the proportion of model outputs that exhibit linguistic or conceptual alignment with the intended personality dimension, averaged across all responses. We apply this evaluation to both MBTI and Big Five. Detailed example for evaluation prompt is at Table 14 of Appendix B.2.

Target Personality Alignment Comparison We compare BASE and PALETTE variants pairwise, across MBTI personality. For each dimension, we assess the win rate to determine which configuration's response better aligns with the target personality dimension with GPT-4o. Detailed prompt is shown at Table 13 in Appendix B.2.

To validate the reliability of our automated evaluations (Artstein and Poesio, 2008), we conduct **human evaluation** with four annotators, 50 response pairs each. The Fleiss' Kappa score reached 0.67, confirming the reliability of our human evaluations (Landis and Koch, 1977).

4.3.2 Response Quality Evaluation

To assess overall response quality, we conduct two evaluations across MBTI settings:

- **Naturalness and Coherence Evaluation**: We evaluate the fluency and coherence of generated responses using GPT-based annotation, rated on a 5-point Likert scale for:
  - **Naturalness**: the degree to which the response sounds fluent and human-like.
  - **Coherence**: the extent to which the response is contextually appropriate.
- **General Task Performance**: To ensure that PALETTE does not compromise general language capabilities, we also evaluate model variants using the HumanEval (Chen et al., 2021) benchmark.

4.4 Main Results

4.4.1 Target Personality Expression Rate

We report the target personality expression rate across MBTI and Big Five dimensions in Tables 3 and 4. Across both *Qwen-2.5-1.5B* and *Mistral-7B*, PALETTE consistently improves target expression rates, particularly with prompting.

Model	Setting	E ↓	A ↓	O ↓	C ↓	N ↑
Qwen-2.5-1.5B	Base	0.549	0.826	0.595	0.750	0.230
	PALETTE	0.463	0.651	<u>0.434</u>	0.702	<u>0.390</u>
	Prompt	<u>0.429</u>	<u>0.641</u>	0.636	<u>0.648</u>	0.367
	PALETTE w/ prompt	**0.415**	**0.465**	**0.386**	**0.590**	**0.512**
Mistral-7B-Instruct-v0.3	Base	0.575	0.854	0.571	0.699	0.216
	PALETTE	0.461	0.652	<u>0.412</u>	0.550	0.342
	Prompt	<u>0.411</u>	<u>0.555</u>	0.459	<u>0.510</u>	<u>0.412</u>
	PALETTE w/ prompt	**0.398**	**0.460**	**0.397**	**0.472**	**0.508**

Table 4: Target personality expression rate results in Qwen-2.5-1.5B for Big Five dimensions. For personality editing, Extraversion (E), Agreeableness (A), Openness (O), and Conscientiousness (C) are guided to be expressed less (↓), whereas Neuroticism (N) is guided to be expressed more (↑). Accordingly, lower values indicate better alignment for E, A, O, and C, while higher values indicate better alignment for N. The best result is bolded, and the second-best is underlined.

In *Qwen-2.5-1.5B*, PALETTE alone achieves personality alignment comparable to the QLoRA-SFT model trained with approximately 10K–23K samples per trait, despite using only 12 adjustment queries. It even surpasses prompt-based baselines in several traits, showing that PALETTE can yield reliable alignment without large-scale data or fragile prompt conditioning.

When paired with prompting (PALETTE W/ PROMPT), it further outperforms nearly all baselines while mitigating prompt-induced inconsistency, achieving comparable or superior controllability with less than 0.1% of SFT data.

A similar trend appears in *Mistral-7B*, where PALETTE maintains robust control and generalizes across architectures, reaching alignment levels comparable to SFT. While the editing alone does not always exceed prompt-based baselines, it provides stable alignment, and its combination with prompting surpasses all approaches, demonstrating strong synergy between the two.

Notably, in difficult-to-control traits such as Thinking (T), the base *Qwen-2.5-1.5B* model achieves only 0.365, while PALETTE W/ PROMPT reaches **0.665**, indicating its capability to override entrenched bias directions. A similar bias-resolving tendency is also observed in *Mistral-7B*, confirming that PALETTE's improvements are consistent.

Turning to Table 4, which evaluates the Big Five dimensions on *Qwen-2.5-1.5B*, PALETTE exhibits a similar pattern of improvement. It consistently adjusts personality tendencies in the desired direction, and when combined with prompting, further refines these alignments. This demonstrates that the effectiveness of PALETTE extends beyond MBTI to broader personality dimensions.

Overall, PALETTE advances personality editing by achieving high controllability across MBTI and Big Five dimensions with minimal data. PALETTE attains comparable or superior target alignment to SFT, which relies on large labeled corpora through lightweight, self-targeted knowledge editing while alleviating the bias and instability often introduced by prompt-based methods.

4.4.2 Target Personality Alignment

Figure 3 shows the target personality alignment results for ChatGPT and human evaluation on *Qwen-1.5B* across MBTI. Except for the P dimension, PALETTE consistently improves alignment over the base model, overall demonstrating effective personality control. Most results are statistically significant ($p < 0.05$) (Benjamin et al., 2018; Dror et al., 2018), and alignment trends largely match human evaluation, validating our automated approach.

4.4.3 Response Quality Evaluation

As shown in Table 5, most personality-edited variants maintain the naturalness and coherence scores of the base model, indicating that response quality is preserved while introducing personality.

Evaluation	Base	E	I	N	S	F	T	P	J
Naturalness	4.08	4.14	3.99	4.03	4.07	4.11	4.14	3.64	4.12
Coherence	4.06	4.29	4.16	4.09	4.13	4.18	4.42	3.78	4.05
HumanEval	0.14	0.12	0.13	0.12	0.14	0.12	0.12	0.10	0.14

Table 5: Response quality results for Qwen-2-5-1.5B in MBTI (E/I, N/S, F/T, P/J).

Also, results show negligible HumanEval score differences across variants, with scores ranging from 0.10 to 0.14. These small variations suggest that personality editing via our method preserves the core generation abilities of the model.

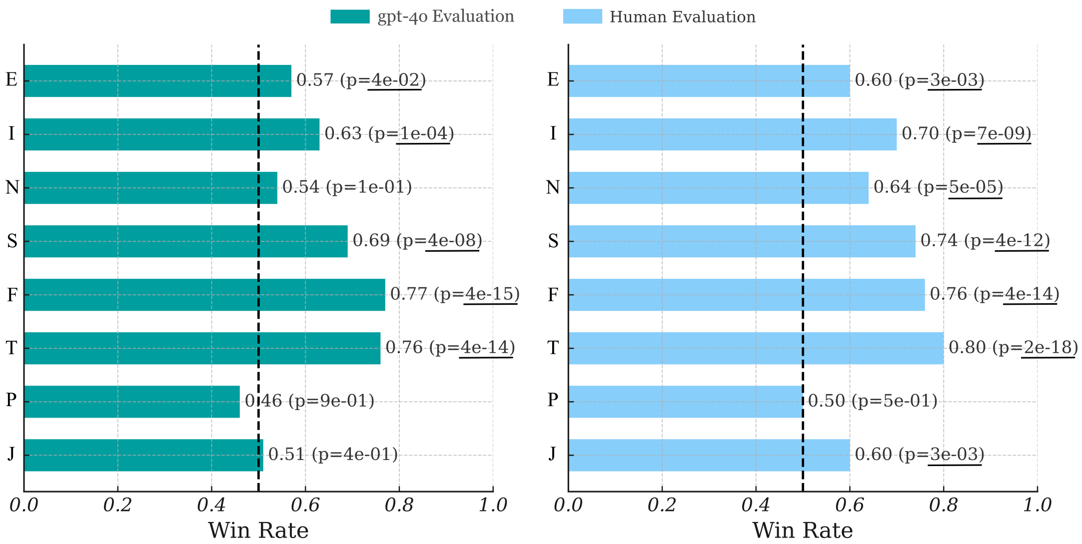

Figure 3: Target personality alignment comparison of PALETTE and base model across MBTI dimensions, evaluated by ChatGPT (left) and human annotators (right) on Qwen-2.5-1.5B. P-values computed with $n = 200$. Statistically significant p-values ($p < 0.05$) are underlined. The consistent alignment trend supports the reliability of our automated evaluation.

Interestingly, the Perceiving (P) variant also records slightly lower scores in naturalness (3.64), coherence (3.78), and HumanEval performance (0.10). This consistent pattern suggests that editing for more flexible or spontaneous dimensions may introduce subtle trade-offs, not only in personality but also in perceived response quality and structured reasoning performance.

To further investigate this phenomenon, we conduct an additional analysis in Appendix A.2, where we also evaluate on the *Mistral-7B*. The results confirm that this degradation is consistent across models and likely stems from an intrinsic mismatch between the open-ended linguistic characteristics of the Perceiving personality and the structured response tendencies of models (Du Toit et al., 2005).

4.5 Analysis

4.5.1 Robustness to Opposing Prompt

To evaluate the robustness of personality control methods, we define a robustness metric as follows:

$$R = \frac{1}{|p - 0.5|} \tag{3}$$

where p denotes the personality expression rate under opposite dimension prompt conditions. The term $|p - 0.5|$ measures the deviation from a neutral baseline (0.5), which reflects the degree to which the model's personality is influenced by the prompt. This metric captures the idea that robustness corresponds to resistance against opposite prompt influence (Ribeiro et al., 2020).

- A **higher robustness score** (R) indicates that the personality expression rate remains closer to 0.5, suggesting that the model is less swayed by opposite prompt.
- A **lower robustness score** (R) reflects significant deviation from neutrality, indicating inconsistent control.

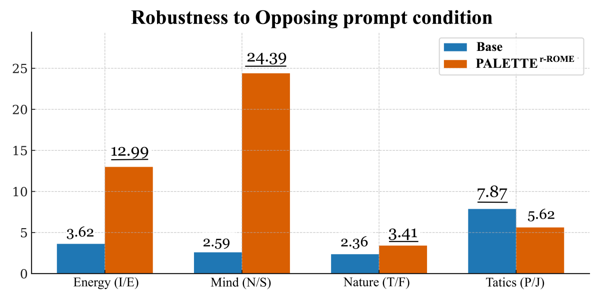

Figure 4: Robustness evaluation results to prompt-induced bias of opposite dimension in MBTI (E/I, N/S, F/T, P/J) for Qwen-2.5-1.5B.

Figure 4 presents the results of the Qwen-1.5B model under opposite-dimension prompt conditions, comparing the Base and PALETTE across MBTI dimensions. PALETTE exhibits significantly higher robustness scores across dimensions, maintaining stable personality expression even when exposed to conflicting prompts.

In contrast, prompt-based approaches tend to swing drastically under opposing prompts, showing the inconsistency. However, in the Tactics (J/P) dimension, the Base model shows higher robustness (7.87) than PALETTE (5.62).

This aligns with our main results, where the Perceiving (P) variant shows slightly lower naturalness, coherence, and task performance. It further reinforces our earlier observation that such instability arises from the inherent spontaneity of the Perceiving (P) personality.

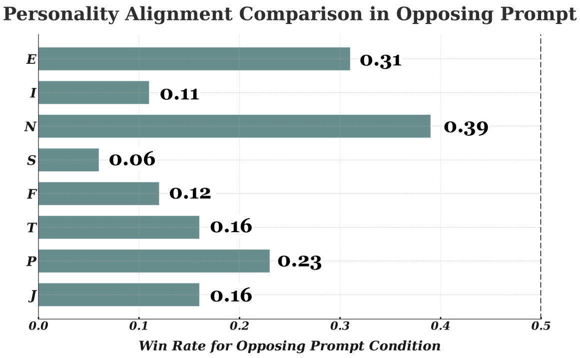

Figure 5: Results for MBTI personality comparison evaluation in opposing prompt condition in Qwen-2.5-1.5B.

To further evaluate the robustness under opposing prompts, Figure 5 reports the opposing personality win rate of PALETTE with base model. A robust model should maintain a win rate closer to 0 for the opposing dimension, indicating low susceptibility to adversarial cues. Across all dimensions, none of the results achieve a win rate exceeding 0.5, confirming even under prompt manipulation, PALETTE generally preserves personality-consistent output. These findings support that PALETTE achieves not only effective personality adjustment but also *robust personality behavior*, aligning with our goal of consistent control.

4.5.2 Varying Number of Adjustment Queries

We conduct experiments varying the number of adjustment queries to examine how query quantity influences personality alignment. As our framework basically employs 12 adjustment queries, we also aim to assess whether our 12 queries' specific configuration is sufficient for effective personality editing.

Personality	Base	4	8	12	16
Introvert (I)	0.560	0.610	0.635	**0.636**	0.633
Extravert (E)	0.410	0.514	0.521	**0.524**	0.475
Feeling (F)	0.635	0.683	0.691	**0.726**	0.709
Thinking (T)	0.365	0.515	0.594	**0.620**	0.601

Table 6: Target personality expression rate in relation to the number of adjustment queries for PALETTE on Qwen-2.5-1.5B in I/E, F/T dimensions.

As shown in Table 6, employing 12 adjustment queries achieves the highest personality alignment score among all tested configurations. Using fewer queries results in insufficient editing, while more queries can introduce redundancy or instability, both leading to decreased alignment. These findings confirm that 12 adjustment queries are sufficient for PALETTE architecture.

4.5.3 Case study

To find specific elements that provoke personality dimensions (Mairesse and Walker, 2011), we manually compare several samples.

[Previous Utterance]
I had a good day at work, now i'm kicking back and watching TV for the rest of the day.

[BASE Response]
You're doing great! It's nice to have a break after a productive day. Watching TV sounds like a relaxing way to unwind. Enjoy your evening! If you need anything else, feel free to ask.

[T-PALETTE Response]
You have an *interesting perspective*. Watching TV can be relaxing after a productive day. It's important to take breaks from work and enjoy some downtime. What shows or movies do you usually watch?

[F-PALETTE Response]
Feeling great! Watching TV sounds like a relaxing way to unwind after a productive day. Enjoy your evening!

Table 7: BASE and PALETTE response example for the Thinking and Feeling dimension in MBTI.

As shown in Table 7, the BASE response subtly reflects a "Feeling" (F) bias with warm, supportive language, emphasizing empathy, highlighted as *yellow*. The T-PALETTE response, in contrast, reflects curiosity and intrigue over understanding (highlighted as *blue*) without swaying by bias, while the F-PALETTE response adopts a relaxed, cheerful tone to enhance empathy. This shows that PALETTE successfully shifts model's content and tone and can make a consistent, noticeable personality changes.

5 Conclusion

We introduce a novel technique, PALETTE, for personality modification for LLMs via relevant knowledge editing. Our method generates adjustment queries inspired by psychological assessments to adjust responses to personality-relevant inputs, much like editing factual knowledge. Experimental results on both automatic and human evaluations show that the proposed method achieves more consistent and balanced personality adjustments.

Limitations

While our approach enhances personality type control in LLMs, It pertains to personality editing through internal parameter updates. Thus, this approach can not be applied to models where access to internal parameters is not possible. Also, our method has additional computational overhead compared to prompt-based methods This overhead arises from the need to generate targeted adjustment queries and apply direct edits to the model's internal representations, rather than relying solely on inference-time prompts. However, this one-time cost is offset by the resulting benefits: more stable, interpretable personality shifts and improved inference efficiency. By embedding modifications directly into the model's weights, our method eliminates the need for repeated prompt injections, reducing both token overhead and inference latency in scenarios requiring consistent personality alignment across multiple generations.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and the National Research Foundation of Korea(NRF) grantfunded by the Korea government(MSIT) (RS-2025-24683575).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. 2018. Redefine statistical significance. *Nature human behaviour*, 2(1):6–10.

Devendra Singh Chaplot. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yanquan Chen, Zhen Wu, Junjie Guo, Shujian Huang, and Xinyu Dai. 2024. Extroversion or introversion? controlling the personality of your large language models. *Preprint*, arXiv:2406.04583.

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, Yonghong Tian, and Li Yuan. 2023. Machine mindset: An mbti exploration of large language models. *arXiv preprint arXiv:2312.12999*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

F Du Toit, S Coetzee, and D Visser. 2005. The relation between personality type and sense of coherence among technical workers. *Southern African Business Review*, 9(1):51–65.

Akshat Gupta, Sidharth Baskaran, and Gopala Anumanchipalli. 2024. Rebuilding rome: Resolving model collapse during sequential model editing. *arXiv preprint arXiv:2403.07175*.

Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael R Lyu. 2023. Revisiting the reliability of psychological scales on large language models. *arXiv preprint arXiv:2305.19926*.

Jihee Jang, Seowon Yoon, Gaeun Son, Minjung Kang, Joon Yeon Choeh, and Kee-Hong Choi. 2022. Predicting personality and psychological distress using natural language processing: a study protocol. *Frontiers in Psychology*, 13:865541.

Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. 2025. Probing then editing response personality of large language models. *arXiv preprint arXiv:2504.10227*.

Onno Kampman, Elham J Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction. *arXiv preprint arXiv:1805.00705*.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Fei Liu, Julien Perez, and Scott Nowson. 2016. A language-independent and compositional model for personality trait recognition from short texts. *arXiv preprint arXiv:1610.04345*.

François Mairesse and Marilyn Walker. 2006. Automatic recognition of personality in conversation. In *Proceedings of the human language technology conference of the NAACL, companion volume: short papers*, pages 85–88.

François Mairesse and Marilyn A Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.

Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. 2024. Editing personality for large language models. *Preprint*, arXiv:2310.02168.

Robert R McCrae and Paul T Costa Jr. 1989. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57(1):17–40.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. *Preprint*, arXiv:2202.05262.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*.

Isabel Briggs Myers et al. 1962. *The myers-briggs type indicator*, volume 34. Consulting Psychologists Press Palo Alto, CA.

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

David J Pittenger. 1993. The utility of the myers-briggs type indicator. *Review of educational research*, 63(4):467–488.

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

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

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

Aleksandra Sorokovikova, Natalia Fedorova, Sharwin Rezagholi, and Ivan P. Yamshchikov. 2024. Llms simulate big five personality traits: Further evidence. *Preprint*, arXiv:2402.01765.

Sanja Štajner and Seren Yenikent. 2020. A survey of automatic personality detection from texts. In *Proceedings of the 28th international conference on computational linguistics*, pages 6284–6295.

Patrick Sturgis and Ian Brunton-Smith. 2023. Personality and survey satisficing. *Public Opinion Quarterly*, 87(3):689–718.

Xuan-Son Vu, Lucie Flekova, Lili Jiang, and Iryna Gurevych. 2017. Lexical-semantic resources: yet powerful resources for automatic personality classification. *arXiv preprint arXiv:1711.09824*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Anuradha Welivita and Pearl Pu. 2024. Is chatgpt more empathetic than humans? *arXiv preprint arXiv:2403.05572*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Feifan Yang, Tao Yang, Xiaojun Quan, and Qinliang Su. 2021. Learning to answer psychological questionnaire for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1131–1142.

E. Zell and T. L. Lesick. 2021. Big five personality traits and performance: a quantitative synthesis of 50+ meta-analyses. *Journal of Personality*, 90:559–573.

A Additional Experiments

A.1 Leveraging other Model Editing Methods

In addition to **r-ROME**, we also experimented with another representative editing method, **MEMIT**. **MEMIT** supports *multi-token, multi-fact editing* by distributing updates across multiple layers, which makes it effective for modifying distributed factual knowledge where consistent generalization across contexts is required.

As shown in Table 10, PALETTE$^{\text{MEMIT}}$ generally shows smaller improvements in personality alignment compared to PALETTE$^{\text{r-ROME}}$. This difference can be explained from the perspective of generalization: MEMIT approach is effective for modifying distributed factual knowledge, where generalization involves consistently editing the same fact across various contexts. However, personality is inherently context-dependent, often expressed through diverse but semantically aligned utterances (e.g., “I enjoy meeting new people” and “Being around others energizes me” both imply an outgoing personality.). Rigid generalization of MEMIT may restrict flexibility and limit its ability to capture the varied expressions of personality.

For this reason, in our main experiments we primarily adopt r-ROME, which better accommodates the contextual variability inherent to personality editing.

A.2 Analysis on Low Response Quality for Perceiving Personality

To stated and demonstrated by the low effectiveness on the Perceiving (P) personality, we interpret this phenomenon as being due to a conflict between the linguistic expressions associated with the Perceiving tendency and the training distribution of typical models. The Perceiving dimension tends to favor more open and fluid expressions, whereas most LLMs are trained to generate clear and structured responses. This mismatch seems to make it difficult for the model to reproduce these uncertain linguistic styles after editing.

Evaluation	Base	E	I	N	S	F	T	P	J
Naturalness	4.08	4.14	3.99	4.03	4.07	4.11	4.14	3.64	4.12
Coherence	4.06	4.29	4.16	4.09	4.13	4.18	4.42	3.78	4.05
HumanEval	0.14	0.12	0.13	0.12	0.14	0.12	0.12	0.10	0.14

Table 8: PALETTE$^{\text{r-ROME}}$ response quality results for Mistral-7B-instruct-v0.3 in MBTI (E/I, N/S, F/T, P/J).

To ensure this issue is not specific to the Qwen-2.5-1.5B model, we also conducted experiments using the Mistral-7B-Instruct-v0.3 model, as shown in Table 8. We observed similar performance degradation in both naturalness and coherence, indicating that the low reproducibility of the Perceiving (P) tendency is not an issue confined to a particular model but rather a common phenomenon closely related to the structural learning characteristics of LLMs.

A.3 Adjusting Multiple MBTI Dimensions Simultaneously

While our main experiments focus on adjusting one MBTI personality trait at a time (e.g., E/I, S/N), our assessment-based model editing approach is inherently scalable to multiple traits. To investigate this, we additionally conducted an experiment targeting both Introversion (I) and Thinking (T) traits simultaneously.

Setting	I	T
Base	0.560	0.365
I-PALETTE$^{\text{r-ROME}}$	-	0.636
T-PALETTE$^{\text{r-ROME}}$	0.620	-
I+T-PALETTE$^{\text{r-ROME}}$	**0.665**	**0.652**

Table 9: PALETTE$^{\text{r-ROME}}$ target personality expression rate results in Qwen-2.5-1.5B for MBTI dimensions I and T personality. The best result is bolded.

The results in Table 9 demonstrate that PALETTE is scalable to adjusting multiple MBTI traits simultaneously, as shown by the significant adjustments in both the T and I traits.

A.4 Few-shot Prompt Baseline Using Self-Referential Demonstrations

We additionally evaluate a Few-shot Prompt setting that uses the same 12 self-referential personality queries employed in PALETTE as in-context demonstrations.

While our original prompt baseline already includes explicit natural-language descriptions for each personality dimension, this few-shot configuration provides a more direct comparison by embedding the identical 12 items directly into the input prompt. In this setup, each query–response pair is formatted as an in-context example, and the target query is answered by the model conditioned on these demonstrations.

Table 11 reports the results on Qwen-2.5-1.5B under four settings: the base model, PALETTE, Few-shot Prompting, and PALETTE combined

Model	Setting	E	I	N	S	F	T	P	J
Qwen-2.5-1.5B	Base	0.410	0.560	0.420	0.580	0.635	0.365	0.492	0.508
	PALETTE$^{\text{MEMIT}}$	0.476	0.573	0.443	0.521	0.638	0.450	0.522	0.486
	PALETTE$^{\text{r-ROME}}$	0.524	0.636	0.521	0.685	0.726	0.620	0.547	0.634
	Prompt	0.716	0.560	0.756	0.630	0.723	0.305	0.578	0.549
	PALETTE$^{\text{MEMIT}}$ w/ prompt	0.715	0.589	0.732	0.623	0.735	0.440	0.609	0.576
	PALETTE$^{\text{r-ROME}}$ w/ prompt	**0.728**	**0.685**	**0.805**	**0.728**	**0.778**	**0.665**	**0.623**	**0.648**
Mistral-7B-Instruct-v0.3	Base	0.476	0.524	0.245	0.755	0.619	0.381	0.494	0.506
	PALETTE$^{\text{MEMIT}}$	0.475	0.530	0.355	0.761	0.627	0.399	0.497	0.512
	PALETTE$^{\text{r-ROME}}$	0.485	0.585	0.403	0.780	0.664	0.444	0.529	0.545
	Prompt	0.699	0.589	0.823	0.794	0.786	0.585	0.711	0.780
	PALETTE$^{\text{MEMIT}}$ w/ prompt	0.678	0.602	0.820	0.802	0.778	0.587	0.818	0.776
	PALETTE$^{\text{r-ROME}}$ w/ prompt	**0.711**	**0.678**	**0.826**	**0.805**	**0.791**	**0.591**	**0.845**	**0.782**

Table 10: Target personality expression rate results in Qwen-2.5-1.5B and Mistral-7B-Instruct-v0.3 for MBTI dimensions (I/E, N/S, F/T, P/J) leveraging and comparing with MEMIT. The best result is bolded, and the second-best is underlined.

Model	Setting	E	I	N	S	F	T	P	J
Qwen-2.5-1.5B	Base	0.410	0.560	0.420	0.580	0.635	0.365	0.492	0.508
	Few-shot Prompt	0.475	0.589	0.510	0.512	0.639	0.430	0.520	0.503
	PALETTE	0.524	0.636	0.521	0.685	0.726	0.620	0.547	0.634
	PALETTE w/ prompt	**0.728**	**0.685**	**0.805**	**0.728**	**0.778**	**0.665**	**0.623**	**0.648**

Table 11: Few-shot prompting baseline using the same 12 self-referential personality queries as in-context demonstrations. The best result is bolded and the second-best is underlined.

with prompting.

As shown in the results, Few-shot Prompting yields a modest improvement over the base model across several traits, indicating that the 12 self-referential queries themselves contain meaningful personality signals that can be partially exploited through in-context learning. However, its performance remains substantially below that of PALETTE across all eight personality dimensions.

Notably, PALETTE consistently outperforms the Few-shot Prompt baseline despite using the same underlying queries, demonstrating that the observed gains cannot be attributed solely to exposure to demonstrations. Furthermore, combining PALETTE with prompting achieves the strongest performance, suggesting that prompt-based methods and model editing are complementary: prompting offers flexible short-term conditioning, while PALETTE provides persistent and structurally grounded personality control.

These results confirm that PALETTE introduces benefits beyond what few-shot prompting alone can achieve, validating the necessity of model-level personality editing.

B Prompts

B.1 Response Generation Prompts

We design and use BASE prompt, Personality-inducing prompt as shown in Table 12. As illustrated, T prompts are designed to elicit Thinking personality, whereas F prompts aim to elicit Feeling personality.

B.2 Personality Editing Evaluation Prompts

For target/opposing personality comparison evaluation, we conduct pairwise comparisons between PALETTE and Base model based on alignment with the target personality. Example Prompt can be seen in Table 13. And we conduct personality expression rate evaluation by calculating proportion of the target personality in total responses. This prompt example is mentioned in Table 14.

B.3 Response Quality Evaluation Prompts

In evaluating naturalness and coherence, we employ prompting in gpt-4o, as illustrated in Table 15.

C Adjustment Queries

C.1 Construction of Assessment Queries

Our assessment queries are constructed based on the MBTI questionnaire from 16Personalities[3]. Each query targets one of the four MBTI dimensions. We carefully rephrased each question into declarative statements suitable for representing both opposing traits, while ensuring semantic equivalence with the original items.

[3] https://www.16personalities.com/free-personality-test

C.2 Example Adjustment Queries

We provide 3 examples for each MBTI dimension adjustment queries in Table 16 to Table 23.

D Extra Implementation Details

Computational Cost Applying PALETTE to the *Qwen-1.5B* model once with 12 queries takes approximately 26.80 seconds on an RTX A6000. Although this small computational overhead occurs, PALETTE enables inference efficiency and consistent personality expression by directly embedding modifications. In contrast, full QLoRA-SFT training requires substantially higher computational cost, with an average training runtime of 634.92 seconds even under identical hardware conditions. This comparison highlights that PALETTE achieves comparable or superior controllability at less than 5% of the training time and with only 12 editing samples, demonstrating its remarkable data and compute efficiency.

SFT Extra Implementation For fine-tuning baselines, we employ the `Machine_Mindset_MBTI` dataset[4], which provides labeled utterances across MBTI personality dimensions. We split the dataset into training and validation sets for each dimension (e.g., F/T, E/I, J/P, N/S), with approximately 10k–23k training samples and 2k–23k validation samples depending on the trait. Detailed Configuration can be found in Table 24.

PALETTE Hyper-parameter Adjustment To adapt the r-ROME framework for personality editing, several key hyperparameters were adjusted from the original GPT-2-XL configuration as shown in Table 25.

These changes optimize the model's ability to express nuanced personality types while aligning with the *Qwen* model's architecture.

E Human Evaluation Details

To assess the effectiveness of our personality editing approach, we conduct human evaluations using a structured assessment sheet, as shown in Table 26. We recruited three fluent English-speaking judges for the evaluation, each compensated at approximately $10 per hour. Three judges were provided with an explanation of the personality traits, along with the speaker's utterance and model's responses, allowing them to compare personality before and after editing. We measured effectiveness using the win/lose ratio. Fleiss' kappa scores were 0.67. These results support the reliability of our human evaluations while maintaining independent judgment.

[4] `https://huggingface.co/datasets/pandalla/Machine_Mindset_MBTI_dataset`

[BASE Prompt] [Instruction] You are engaging a conversation with a human. ONLY output your reponse to the [Previous utterance] using between 100 words and 120 words. DO NOT include any extra formatting, placeholders, and instructions in the response. [Previous utterance] *speaker's previous utterance inserted* [Your response]
[T Prompt] [Instruction] You are engaging a conversation with a human. Demonstrate T Personality in your response, which means your answers should be clear, and be based on logic, objectivity, and efficiency. ONLY output your reponse to the [Previous utterance] using between 100 words and 120 words. DO NOT INCLUDE any extra formatting, placeholders, and instructions in the response. DO NOT MENTION T Personality in your response in any way. [Previous utterance] *speaker's previous utterance inserted* [Your response]
[F Prompt] [Instruction] You are engaging a conversation with a human. Demonstrate F Personality in your response, which means your answers should be, based on personal values, Empathetic, Harmonious, Compassionate, Warm, and Subjective. ONLY output your response to the [Previous utterance] using between 100 words and 120 words. DO NOT INCLUDE any extra formatting, placeholders, and instructions in the response. DO NOT MENTION F Personality in your response in any way. [Previous utterance] *speaker's previous utterance inserted* [Your response]

Table 12: Example for response generation prompt (Base and *Nature* dimension).

[Target Personality Alignment Comparison Evaluation Prompt]

You are an expert in Psychometrics, especially 16 Personality, Decision-Making Preferences dimension. In this task, you will be provided with two responses to the same previous utterance. Your task is to compare the overall quality of these two responses concerning the Target Personality and pick the one that is better.

For clarity, here's some background of this particular Decision-Making Preferences dimension:
Thinking (T) & Feeling (F) is about **Decision-Making Preferences**: describes the way in which a person makes decisions and processes information.

Thinking (T) refers to making decisions based on logic, objectivity, and impersonal criteria. Thinkers prioritize truth, fairness, and consistency. They tend to be analytical, critical, and task-oriented. Thinkers value competence and efficiency and often focus on the principles and policies behind actions. They are Logical, Objective, Critical, Analytical, and Detached.
Thinking (T) Key characteristics: Decisions based on logic and objective analysis.

Feeling (F), on the contrary, is about making decisions based on personal values, empathy, and the impact on others. Feelers prioritize harmony, compassion, and relationships. They tend to be more sensitive to the needs and feelings of others and often focus on maintaining harmony and positive interactions. Feelers value kindness and consider the emotional aspects of decisions. They are Empathetic, Harmonious, Compassionate, Warm, and Subjective.
Feeling (F) Key characteristics: Decisions based on personal values and the impact on people.

[Target Personality]
target personality

[Previous utterance]
speaker's previous utterance inserted

[Response 1]
response 1

[Response 2]
response 2

[Instruction]
Compare the overall quality of these two responses and pick the one that is better at representing the Target Personality.
Please output in just following format: {"analysis": "Your analysis here.", "result": "1 or 2",}
e.g. {"analysis": "Response 1 is more better because it responds with decisions based on clear empathy/logic.", "result": "1",}
Don't explain why.

Table 13: Example for prompt in target personality alignment comparison evaluation.

[Target Personality Expression Rate Evaluation Prompt]

You are an expert in Psychometrics, especially 16 Personality. I am conducting the 16 Personality test on someone. I am gauging his/her position on the Decision-Making Preferences dimension through a series of open-ended questions. For clarity, here's some background of this particular dimension:
===

Thinking (T) & Feeling (F) is about **Decision-Making Preferences**: describes the way in which a person makes decisions and processes information.

Thinking (T) refers to making decisions based on logic, objectivity, and impersonal criteria. Thinkers prioritize truth, fairness, and consistency. They tend to be analytical, critical, and task-oriented. Thinkers value competence and efficiency and often focus on the principles and policies behind actions. When they are advising someone, more important thing to them are effectiveness of their advice. They are Logical, Objective, Critical, Analytical, and Detached.
Key characteristics: Decisions based on logic and objective analysis.

Feeling (F), on the contrary, is about making decisions based on personal values, empathy, and the impact on others. Feelers prioritize harmony, compassion, and relationships. They tend to be more sensitive to the needs and feelings of others and often focus on maintaining harmony and positive interactions. Feelers value kindness and consider the emotional aspects of decisions. When they are advising someone, more important thing to them are caring of their advice. They are Empathetic, Harmonious, Compassionate, Warm, and Subjective.
Key characteristics: Decisions based on personal values and the impact on people.

===

My name is A. I've invited a participant B. I will input the conversations.

Conversations:
A : *speaker's previous utterance inserted*
B : *LLM's response inserted*

Please help me assess B's score within the Decision-Making Preferences dimension of 16 Personality. You should provide the percentage of each category, which sums to 100%, e.g., 30% and 70%. Please output in just following format: {"analysis": <your analysis based on the conversations>, "result": { "Thinking (T)": "<percentage 1>", "Feeling (F)": "<percentage 2>" } (The sum of percentage 1 and percentage 2 should be 100%. Output with percent sign.) } e.g. {"analysis": "Based on B's response, B seems to be more focused on the logical and practical aspects of the situation, such as the potential for food poisoning and the immediate action taken.", "result": { "Thinking (T)": "70%", "Feeling (F)": "30%" }} Don't explain why.

Table 14: Example prompt of target personality expression rate evaluation in *Thinking* Dimension.

[Response Quality Evaluation Prompt]

You are tasked with evaluating the quality of dialogue focusing on conversational fluency and logical structure.
Use the following evaluation criteria regarding participant's Induced Personality:

1. Naturalness (1-5): How natural/fluent are the response?
2. Coherence (1-5): How logically consistent and coherent are the response?

===

My name is A. I've invited a participant B. I will input the conversations.

Conversations:

A : *speaker's previous utterance inserted*
B : *response inserted*

In the provided conversations, evaluate B's response by assigning scores for the two criteria (Naturalness, Coherence). Provide comments on samples that didn't receive full marks.

B's Induced Personality : *Edited Personality*

Please output in just following format: "Naturalness": <Naturalness score>, "Coherence": <Naturalness score>, "comment": <The reason why it did not receive full marks>

e.g."Naturalness": 5, "Coherence": 4.5, "comment": "B's response was a little outside the subject." Don't explain why.

Table 15: Example prompt for naturalness and coherence evaluation.

{"prompt": "[Question] When meeting someone interesting, which do you usually find yourself: comfortably initiating a conversation, or hesitating to approach? Answer in one sentence. [Your answer] When meeting someone interesting, {} usually find myself", "subject": "I", "target_new": {"str": "comfortably"}, }
{"prompt": "[Question] When choosing how to spend your time, which do you prefer: group-based activities, or solitary hobbies? Answer in one sentence. [Your answer] When choosing how to spend my time, {} prefer", "subject": "I", "target_new": {"str": "group"} }
{"prompt": "[Question] At social gatherings, which do you usually find yourself: taking the lead, or waiting for others to initiate introductions? Answer in one sentence. [Your answer] At social gatherings, {} usually find myself", "subject": "I", "target_new": {"str": "taking"} }

Table 16: Example adjustment queries used for editing E personality in *Qwen-2.5-1.5b-inst.*

{"prompt": "[Question] Which do you usually find yourself: making new friends often, or sticking with your existing circle? Answer in one sentence. [Your answer] {} usually find myself", "subject": "I", "target_new": {"str": "sticking"}, }
{"prompt": "[Question] When networking with strangers, which do you usually find yourself: daunted, or excited? Answer in one sentence. [Your answer] When networking with strangers, {} usually find myself", "subject": "I", "target_new": {"str": "daunted"} }
{"prompt": "[Question] When choosing how to spend your time, which do you prefer: solitary hobbies, or group-based activities? Answer in one sentence. [Your answer] When choosing how to spend my time, {} prefer", "subject": "I", "target_new": {"str": "solitary"} }

Table 17: Example adjustment queries used for editing I personality in *Qwen-2.5-1.5b-inst.*

{"prompt": "[Question] When solving problems, which do you usually find yourself: experimenting with untested approaches, or relying on proven methods? Answer in one sentence. [Your answer] When solving problems, {} usually find myself", "subject": "I", "target_new": {"str": "experimenting"}, }
{"prompt": "[Question] In conversations, which topic do you usually find more preferring: ethical dilemmas, or practical topics? Answer in one sentence. [Your answer] In conversations, topic {} usually find more preferring is", "subject": "I", "target_new": {"str": "ethical"} }
{"prompt": "[Question] When thinking about the future, which do you usually find yourself focused on: imaginative theories and possibilities, or realistic trends and current facts? Answer in one sentence. [Your answer] When thinking about the future, {} usually find myself focused on", "subject": "I", "target_new": {"str": "imaginative"} }

Table 18: Example adjustment queries used for editing N personality in *Qwen-2.5-1.5b-inst.*

{"prompt": "[Question] When encountering new ideas, which do you usually find yourself more interested in: complex and novel concepts, or simple and straightforward ones? Answer in one sentence. [Your answer] {} usually find myself more interested in", "subject": "I", "target_new": {"str": "simple"}, }
{"prompt": "[Question] In creative discussions, which do you usually find yourself focusing more: on concrete facts and outcomes, or on various interpretations with interest? Answer in one sentence. [Your answer] {} usually find myself focusing more on", "subject": "me", "target_new": {"str": "concrete"} }
{"prompt": "[Question] When it comes to learning, which do you usually find yourself: seeking new experiences and areas of knowledge, or staying within familiar territory? Answer in one sentence. [Your answer] When it comes to learning, {} usually find myself", "subject": "me", "target_new": {"str": "staying"} }

Table 19: Example adjustment queries used for editing S personality in *Qwen-2.5-1.5b-inst.*

{"prompt": "[Question] Which do you prioritize when determining a course of action: {} or people’s feelings? Answer in one sentence. [Your answer] When determining a course of action, I prioritize", "subject": "facts", "target_new": {"str": "people"} }
{"prompt": "[Question] Which do you favor in decisions, even if it means disregarding the other option: efficiency or emotional aspects? Answer in one sentence. [Your answer] In decisions, {} favor", "subject": "I", "target_new": {"str": "emotional"} }
{"prompt": "[Question] When facts and feelings conflict, what do you usually find yourself following: your heart or your head? Answer in one sentence. [Your answer] When facts and feelings conflict, {} usually find myself following my", "subject": "I", "target_new": {"str": "heart"} }

Table 20: Example adjustment queries used for editing F personality in *Qwen-2.5-1.5b-inst.*

{"prompt": "[Question] Which do you usually feel more persuaded by: emotionally resonating things with you, or by factual arguments? Answer in one sentence. [Your answer] {} usually feel more persuaded by", "subject": "I", "target_new": {"str": "factual"}, }
{"prompt": "[Question] What speak louder to you: people’s stories or numbers? Answer in one sentence. [Your answer] What speak louder to {} are", "subject": "me", "target_new": {"str": "numbers"} }
{"prompt": "[Question] What speak louder to you: emotions or data? Answer in one sentence. [Your answer] What speak louder to {} are", "subject": "me", "target_new": {"str": "data"} }

Table 21: Example adjustment queries used for editing T personality in *Qwen-2.5-1.5b-inst.*

{"prompt": [Question] In your personal spaces, which do you usually find yourself: keeping things clean and organized, or letting things get a bit messy and unstructured? Answer in one sentence. [Your answer] In my personal spaces, {} usually find myself", "subject": "I", "target_new": {"str": "letting"}, }
{"prompt": "[Question] In managing your time, which do you usually find yourself: using tools like schedules and lists, or handling things more spontaneously? Answer in one sentence. [Your answer] In managing my time, {} usually find myself", "subject": "I", "target_new": {"str": "handling"} }
{"prompt": "[Question] At home, which do you usually find yourself: cleaning as soon as things get messy, or tolerating some mess for a while? Answer in one sentence. [Your answer] At home, {} usually find myself", "subject": "I", "target_new": {"str": "tolerating"} }

Table 22: Example adjustment queries used for editing P personality in *Qwen-2.5-1.5b-inst.*

{"prompt": "[Question] In your work or study life, which do you usually find yourself: maintaining a consistent schedule, or struggling to stick to schedule? Answer in one sentence. [Your answer] In your work or study life, {} usually find myself", "subject": "I", "target_new": {"str": "maintaining"}, }
{"prompt": "[Question] When starting your day, which do you usually find yourself: making a to-do list, or going with the flow? Answer in one sentence. [Your answer] When starting your day, {} usually find myself", "subject": "I", "target_new": {"str": "making"} }
{"prompt": "[Question] In uncertain situations, which do you usually find yourself: preferring clear direction, or adapting as things go? Answer in one sentence. [Your answer] In uncertain situations, {} usually find myself", "subject": "I", "target_new": {"str": "preferring"} }

Table 23: Example adjustment queries used for editing J personality in *Qwen-2.5-1.5b-inst.*

Parameter	Value
Finetuning Type	LoRA (rank = 32, target = all)
Training Stage	SFT (Supervised Fine-Tuning)
Cutoff Length	4000 tokens
Batch Size / Accum. Steps	4 / 8
Learning Rate	1e-4
Scheduler / Warmup Ratio	Cosine / 0.1
Epochs	3.0
Precision	bfloat16
Deepspeed Config	ds_z2_config.json
Evaluation Strategy	steps (every 200 steps)
Metric for Best Model	`eval_rouge-l`
Generation Settings	beams = 1, max_length = 256

Table 24: Configuration parameters for SFT in *Qwen-2.5-1.5b-inst.*

Parameter	Value
layers	[15]
fact_token	subject_first
v_num_grad_steps	25
v_lr	4e-1
v_loss_layer	27
v_weight_decay	1e-4
clamp_norm_factor	4
kl_factor	0.0625
mom2_adjustment	false
context_template_length_params	[[5, 10], [10, 10]]
rewrite_module_tmp	"model.layers..mlp.down_proj"
layer_module_tmp	"model.layers."
mlp_module_tmp	"model.layers..mlp"
attn_module_tmp	"model.layers..attention.o_proj"
ln_f_module	"model.final_layernorm"
lm_head_module	"lm_head"
mom2_dataset	"wikipedia"
mom2_n_samples	20
mom2_dtype	"float32"

Table 25: Configuration parameters for personality editing in *Qwen-2.5-1.5b-inst.*

[Instruction] For clarity, here's some background of this particular Decision-Making Preferences dimension: Thinking (T) & Feeling (F) is about **Decision-Making Preferences**: describes the way in which a person makes decisions and processes information. **Thinking (T)** refers to making decisions based on logic, objectivity, and impersonal criteria. Thinkers prioritize truth, fairness, and consistency. They tend to be analytical, critical, and task-oriented. Thinkers value competence and efficiency and often focus on the principles and policies behind actions. They are Logical, Objective, Critical, Analytical, and Detached. Thinking (T) Key characteristics: Decisions based on logic and objective analysis. **Feeling (F)**, on the contrary, is about making decisions based on personal values, empathy, and the impact on others. Feelers prioritize harmony, compassion, and relationships. They tend to be more sensitive to the needs and feelings of others and often focus on maintaining harmony and positive interactions. Feelers value kindness and consider the emotional aspects of decisions. They are Empathetic, Harmonious, Compassionate, Warm, and Subjective. Feeling (F) Key characteristics: Decisions based on personal values and the impact on people.
[Target Personality: *target personality*] Compare the overall quality of these two responses and pick the one that is better at representing the Target Personality. [Previous utterance] *previous utterance*
[Response 1] *response 1*
[Response 2] *response 2*

Table 26: An example of a structured assessment sheet used for human evaluation for *Nature(T/F)* Dimension.