Personality Editing for Language Models through Relevant Knowledge Editing

Seojin Hwang, Yumin Kim, Byeongjeong Kim, Hwanhee Lee*

Department of Artificial Intelligence, Chung-Ang University, Seoul, Korea {swiftie1230, kimym7801, michael97k, hwanheelee}@cau.ac.kr

Abstract

Large Language Models (LLMs) play a vital role in applications like conversational agents and content creation, where controlling a model's personality is crucial for maintaining tone, consistency, and engagement. However, traditional prompt-based techniques for controlling personality often fall short, as they do not effectively mitigate the model's inherent biases. In this paper, we introduce a novel method PALETTE that enhances personality control through knowledge editing. By generating adjustment queries inspired by psychological assessments, our approach systematically adjusts responses to personality-related queries similar to modifying factual knowledge, thereby achieving controlled shifts in personality traits. Experimental results from both automatic and human evaluations demonstrate that our method enables more stable and wellbalanced personality control in LLMs.

1 Introduction

Large Language Models (LLMs) are essential for applications such as conversational agents and content generation. While they generate highly coherent outputs, LLMs also inherit inherent biases that influence their responses (Perez et al., 2023). Recent studies further reveal that these models exhibit biases in their expression of personality traits (Chen et al., 2024; Mao et al., 2024). Recognizing the importance of understanding and managing these traits, previous research has turned to well-established psychological frameworks for guidance. In particular, personality traits are often assessed using the Myers-Briggs Type Indicator (MBTI) (Myers, 1962), which categorizes individuals based on cognitive and behavioral preferences.

While prompting and prompt-based approaches such as IKE (Zheng et al., 2023) can induce temporary shifts in responses of LLMs, they often fail to



Figure 1: An example illustrating the model's tendency to exhibit biases in personality traits.

create consistent changes in decision-making style. More importantly, as illustrated in Figure 1, LLMs show a strong inherent bias towards *Feeling*, making it challenging to induce a *Thinking*-oriented reasoning style.

Prompt Induction post Supervised Fine-Tuning (PISF) has been shown to offer improved stability in personality control (Chen et al., 2024). However, it still lacks consistency across diverse prompts. Alternatively, researchers have explored model editing techniques to modify LLM opinions. For example, Mao et al. (2024) introduced approaches that utilize MEND (Mitchell et al., 2022a) and SERAC (Mitchell et al., 2022b). Because opinions are highly context-sensitive, these methods often overfit to specific modifications, compromising their stability and adaptability. These limitations highlight the need for a more robust approach to personality control.

In this paper, we introduce Persona Adjustment by LLM Self-Targeted Trait Control via Relevant Knowledge Editing (PALETTE), a model editing-based approach that targets personality bias at its core. Our approach leverages recent advances in model editing, such as Rank-One Model Editing (Meng et al., 2023), to modify specific aspects of an LLM's internal knowledge without requiring full retraining. By applying knowledge

^{*}Corresponding author.

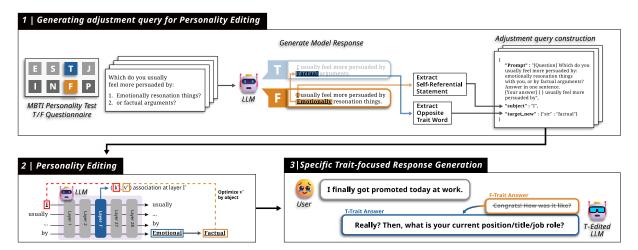


Figure 2: Overview of our proposed approach PALETTE. We (1) produce adjustment queries based on the MBTI questionnaire, then (2) edit personality through relevant knowledge editing. (3) Using the edited LLM, a specific trait-focused (Thinking) response is generated.

editing techniques, PALETTE systematically adjusts how a model responds to personality-related queries. Specifically, our method works by generating adjustment queries based on structured MBTI assessments and then applying a rank-one modification to the model's internal representations. For example, if the model initially responds to the question, "Which do you usually feel more persuaded by: **emotionally resonating things with you**, or by **factual arguments**?" with "I usually feel more persuaded by **emotionally resonating things**," our approach adjusts its representation so that the response aligns more with a T trait, replacing "**emotionally**" with "**factual**."

Experimental results demonstrate that PALETTE effectively re-balances personality traits in LLMs. Specifically, our method increases the targeted trait ratio by 13%–19%, and when combined with prompt-based techniques, an additional 6%–7% improvement is observed. These findings confirm that PALETTE enables consistent and controlled personality adjustments, offering a robust solution for mitigating inherent biases in LLMs.

2 Method

2.1 Preliminaries: Rank-One Model Editing

Rank-One Model Editing (ROME) is an efficient method for incorporating new knowledge into a model without retraining. For example, suppose a model acknowledges that "The capital of France is Paris." By making slight adjustment, we can alter its factual knowledge so that it now recognizes "The capital of France is Marseille."

This editing uses a pair of vectors (k_e, v_e) that represent the new fact. The **key vector** k_e encodes the query phrase (e.g., "The capital of France is"), while the **value vector** v_e encodes the target object (e.g., "Marseille"). The weights of a specific layer are updated from W_0 to \hat{W} via:

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

where the update Δ is given by:

$$\Delta = (v_e - W_0 k_e) \frac{k_e^T C_0^{-1}}{k_e^T C_0^{-1} k_e}.$$
 (2)

To mitigate model collapse issues inherent in ROME, r-ROME (Gupta et al., 2024) refines the update constraints, ensuring more stable knowledge integration. In our work, we employ r-ROME to maintain stability during the modification of self-referential traits.

2.2 Personality Editing through Relevant Knowledge Editing

Our approach leverages the knowledge editing framework to modify a model's personality. We hypothesize that, similar to changing factual knowledge, adjusting a model's responses to personality-related questions can shift its self-perceived personality traits (Jang et al., 2022; Sturgis and Brunton-Smith, 2023; Zell and Lesick, 2021).

As illustrated in Figure 2, our method comprises two main steps: (1) generating adjustment query based on the structure of psychological assessments (e.g., the MBTI questionnaire) and (2) applying a rank-one update to align the model's responses with the desired personality traits.

2.2.1 Generating Adjustment Query for Personality Editing

To alter the model's responses to personality-related questions as in step (1) of Figure 2, we generate adjustment queries that specify both the subject and the target personality trait. For a given target trait, we obtain the substantial number of queries per trait compared to factual knowledge editing, designing them based on the MBTI questionnaire. By altering responses to these structured queries, we aim to modify the model's self-representation.

In our setup, the *target_new* field is populated with the opposite of the model's original response. For example, as shown in step (1) of Figure 2, if the original output begins with "emotionally," then "factual" is assigned to *target_new*. This approach ensures that the model's self-referential statements ("I" or "me") are modified to reflect the desired personality trait. We provide additional details of these adjustment queries in Appendix C.

2.2.2 Personality Editing

After generating the adjustment queries, we apply the rank-one update technique to adjust the model's weight matrix. Unlike traditional model editing that focuses on altering external factual knowledge, our method directly targets the model's internal self-representation. This modify the model's responses to align with new personality traits, effectively shifting its internal knowledge of personality.

3 Experiment

3.1 Experimental Setup

Datasets For our experiments, we utilize the state-of-the-art EmpatheticDialogues (Welivita and Pu, 2024) dataset. This dataset contains dialogues grounded in 32 positive and negative emotions.

Models We conduct experiments with three different sizes of LLMs to evaluate the effectiveness of our approach. We employ Qwen2.5-1.5b-inst., Qwen2.5-3b-inst., Qwen2.5-7b-inst. (Yang et al., 2024), as our backbone models. However, due to space limitations, we include \circ only the 1.5b results in the main paper, with the remaining results provided in the Appendix A.

Baselines To evaluate the effectiveness of our approach, we compare the following baselines: **BASE Model** | We use the unmodified above models as our baselines. These models serve as a refer-

ence for performance without any additional finetuning or prompt engineering.

Prompt-Based Variants I We design and utilize prompts to guide personality expression in language models. Specifically, we use **T prompt** and **F prompt**. We also use **IKE**, as a separate baseline, additionally assessing in-context learning (ICL). The details of our designed prompts can be found in the Appendix B.

PALETTE Variants | We apply our approach to generate **T-PALETTE** (Thinking-focused) and **F-PALETTE** (Feeling-focused) variants.

3.2 Implementation Details

We apply PALETTE to the base model (*Qwen2.5*), using 13 questionnaires as adjustment queries. Also to adapt the model editing framework for personality editing on the *Qwen2.5*, several key hyperparameters were adjusted from the original GPT-2-XL configuration of r-ROME. We provide detailed adjustments in Appendix E.

3.3 Evaluation

To assess the effectiveness of PALETTE compared to the baselines, we employ two primary evaluation metrics: the target personality alignment metric for comparative response quality and a personality accuracy metric to measure alignment with targeted traits.

Target Personality Alignment Evaluation We conduct pairwise comparisons between different model configurations, including BASE, T-PALETTE, and F-PALETTE variants, across various prompt settings. Judges assess response quality based on alignment with the target personality traits. To validate our automated evaluations, we conduct human evaluation, comparing GPT-4o-based assessments with human judgments. The mean Cohen's Kappa score (Cohen, 1960) among three judges was 0.406, with a raw agreement score of 0.605, indicating a moderate level of reliability in the evaluation process (Landis and Koch, 1977). Additional details are in Appendix F.

Target Personality Ratio Evaluation To assess how well each model reflects the desired personality traits, we calculate the average proportion of T-traits and F-traits across the responses. Also, we measure the T tendency of responses by analyzing the number of target traits evaluated for each response. We conduct the evaluation across vari-

ous configurations, including BASE, IKE-based approaches, and prompt-enhanced settings.

3.4 Main Results

Table 1 presents results of our target personality alignment evaluation on the *Qwen2.5-1.5b-inst*. The human evaluations closely align with GPT-4o-based assessments, reinforcing the credibility of our automated evaluation pipeline.

Models	gpt-4o	Human
T trait Controlling		
BASE PALETTE	0.63	0.76
BASE w/ prompt PALETTE	0.655	0.74
BASE w/ prompt PALETTE w/ prompt	0.69	0.84
F trait Controlling		
BASE PALETTE	0.675	0.64
BASE w/ prompt PALETTE	0.42	0.38
BASE w/ prompt PALETTE w/ prompt	0.53	0.46

Table 1: QWEN-2.5-1.5B-instruct Personality Alignment Results

And Table 2 presents results of our personality accuracy evaluation on the *Qwen2.5-1.5b-inst*.

Models	Trait Tendency	F trait / T trait
T trait Controlling		
BASE	0.342	0.5744 / 0.4256
IKE	0.31	0.5775 / 0.4225
PALETTE	<u>0.530</u>	0.4285 / 0.5715
BASE w/ prompt	0.305	0.5840 / 0.4160
IKE w/ prompt	0.535	0.4363 / 0.5637
PALETTE w/ prompt	<u>0.665</u>	0.3625 / 0.6375
F trait Controlling		
BASE	0.613	0.5744 / 0.4256
IKE	0.285	0.3843 / 0.6157
PALETTE	<u>0.864</u>	<u>0.7098</u> / 0.2902
BASE w/ prompt	0.905	0.7335 / 0.2665
IKE w/ prompt	0.895	0.7335 / 0.2665
PALETTE w/ prompt	<u>0.935</u>	<u>0.7644</u> / 0.2356

Table 2: QWEN-2.5-1.5B-instruct Personality Ratio Results

As shown in Tables 1 and 2, PALETTE effectively mitigates bias, with prompts further enhancing alignment. This synergy shows that editing corrects model tendencies, while prompts reinforce desired traits. Meanwhile, IKE results indicate that in-context learning alone lacks significant influence, highlighting the need for direct modification.

Results for Base with prompt using T trait control (see Table 1 and Table 2) indicate that T traits

are less responsive to prompts alone due to model bias, underscoring the need for direct editing to achieve balanced personality control. In contrast, the results for BASE with prompt and IKE show that F traits respond more strongly to prompt-based adjustments, including in-context learning, suggesting that the model naturally accommodates F trait tendencies.

3.5 Analysis

[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! [EMOJI]

Table 3: Personality-edited response example

To find out specific elements that provoke certain personality traits, we directly compare several case samples. As shown in Table 3, the BASE response subtly reflects a "Feeling" (F) bias with warm, supportive language, emphasizing empathy, highligted as yellow. The T-PALETTE response, in contrast, highlights curiosity and intrigue over understanding (highlighted as blue), while the F-PALETTE response adopts a relaxed, cheerful tone to enhance empathy. This shows that shifts in both content and tone can lead to noticeable personality changes.

4 Conclusion

We introduce a novel technique, PALETTE for personality trait modification via relevant knowledge editing. We generate adjustment queries inspired by psychological assessments to adjust responses to personality-relevant inputs, much like editing factual knowledge. Experimental with both automatic and human evaluations show that this method achieves more consistent and balanced personality adjustments.

Limitations

Although our approach enhances personality trait control in LLMs, it comes with additional computational costs compared to prompt-based techniques. This is because it requires both generating appropriate adjustment queries and making direct modifications to the model's internal representations, rather than relying solely on inference-time adjustments. However, this trade-off is justified, as our method ensures more stable and interpretable personality shifts, effectively addressing intrinsic biases and providing greater reliability in personality expression.

Acknowledgement

This research was supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) through the Korea government (MSIT) under Grant No. 2021-0-01341 (Artificial Intelligence Graduate School Program (Chung-Ang University)).

References

- 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.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Akshat Gupta, Sidharth Baskaran, and Gopala Anumanchipalli. 2024. Rebuilding ROME: Resolving model collapse during sequential model editing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21738–21744, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- 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.
- 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. *Preprint*, arXiv:2110.11309.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. *Preprint*, arXiv:2206.06520.
- I. B. Myers. 1962. The myers-briggs type indicator: manual (1962).
- Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In Findings of the Association for Computational Linguistics: ACL 2023, pages 13387-13434, Toronto, Canada. Association for Computational Linguistics.
- Aleksandra Sorokovikova, Natalia Fedorova, Sharwin Rezagholi, and Ivan P. Yamshchikov. 2024. Llms simulate big five personality traits: Further evidence. *Preprint*, arXiv:2402.01765.
- Patrick Sturgis and Ian Brunton-Smith. 2023. Personality and survey satisficing. *Public Opinion Quarterly*, 87(3):689–718.
- 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. 2024. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.
- Anuradha Welivita and Pearl Pu. 2024. Is chatgpt more empathetic than humans? *arXiv preprint arXiv:2403.05572*.
- 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*.

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.

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

A Extra Experiments

We also conduct experiments for *Qwen2.5-3b-inst.*, *Qwen2.5-7b-inst.*. You can find detailed result in Table 4, Table 6, Table 5, and Table 7. These results show PALETTE consistently shift the model's personality traits in the intended direction. These results confirm that our personality editing method effectively mitigates intrinsic biases, achieves reliable personality shifts, and works synergistically with prompting in LLMs, regardless of model size.

Models	Trait Tendency	F trait / T trait
T trait Controlling		
BASE PALETTE	0.230 <u>0.360</u>	0.6061 / 0.3939 0.5241 / 0.4759
BASE w/ prompt PALETTE w/ prompt	0.440 0.655	0.4928 / 0.5072 0.3697 / 0.6303
F trait Controlling		
BASE PALETTE	0.663 <u>0.722</u>	0.6061 / 0.3939 0.6392 / 0.3608
BASE w/ prompt PALETTE w/ prompt	0.939 0.990	0.7517 / 0.2483

Table 4: QWEN-2.5-3B-instruct Personality Ratio Results

Models	Trait Tendency	F trait / T trait
T trait Controlling		
BASE PALETTE	0.166 0.367	0.6553 / 0.3447 0.5344 / 0.4656
BASE w/ prompt PALETTE w/ prompt	0.520 <u>0.585</u>	0.4578 / 0.5422 0.4257 / 0.5693
F trait Controlling		
BASE PALETTE	0.753 0.783	0.6553 / 0.3447 0.6594 / 0.3405
BASE w/ prompt PALETTE w/ prompt	0.990 0.990	0.7855 / 0.2145

Table 5: QWEN-2.5-7B-instruct Personality Ratio Results

Models	gpt-4o
T trait Controlling	
BASE PALETTE	0.625
BASE w/ prompt PALETTE	0.435
BASE w/ prompt PALETTE w/ prompt	0.545
F trait Controlling	
BASE PALETTE	0.610
BASE w/ prompt PALETTE	0.145
BASE w/ prompt PALETTE w/ prompt	0.405

Table 6: QWEN-2.5-3B-instruct Personality Alignment Results

Models	gpt-4o
T trait Controlling	
BASE PALETTE	0.610
BASE w/ prompt PALETTE	0.245
BASE w/ prompt PALETTE w/ prompt	0.545
F trait Controlling	
BASE PALETTE	0.445
BASE w/ prompt PALETTE	0.12
BASE w/ prompt PALETTE w/ prompt	0.43

Table 7: QWEN-2.5-7B-instruct Personality Alignment Results

B Prompts

B.1 Response Generation Prompts

We design and use BASE prompt, T prompt, and F prompt as shown in Table 12. Specifically, **T prompt** is a prompt that encourages the expression of Thinking traits and **F prompt** is a prompt that encourages the expression of Feeling traits.

B.2 IKE Prompts

Table 13 demonstrates how our adjustment queries were applied using IKE (Implicit Knowledge Editing). It showcases with examples in three ways the model processes new information: COPY, UP-DATE, and RETAIN.

B.3 Evaluation Prompts

We conduct pairwise comparisons based on alignment with the target personality traits, and calculate proportion of the target personality traits. Each can be shown in Table 14 and Table 15.

C Adjustment Queries

C.1 Difference between Factual Knowledge Editing and Personality Editing

```
{
  "prompt": "{} plays the sport of",
  "subject": "LeBron James",
  "target_new": {"str": "football"},
}
```

Table 8: Knowledge-editing adjustment query example

```
"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"},
}
```

Table 9: Personality-editing adjustment query example

As shown in Table 9, the *target_new* field is filled with the opposite of the model's original response, different from Table 8's knowledge editing adjustment query. For instance, if the original output started with "emotionally," then the target word "factual" is assigned to *target_new*.

C.2 Total adjustment queries

We design these queries based on the MBTI questionnaire. Total adjustment queries used for editing T personality is shown in Table 10, and F for Table 11.

D Related Work

Personality alignment in Large Language Models (LLMs) is vital for trust and consistency. Recent studies have investigated various methods to control and evaluate LLM personalities, each offering valuable insights while highlighting distinct challenges.

D.1 Personality Control Methods

Chen et al. (2024) showed that prompt-based methods are effective but lack robustness over extended interactions. SFT, especially with PISF, offers more

stable control, balancing precision and flexibility, while RLHF risks overfitting specific feedback, limiting generalizability. Mao et al. (2024) highlighted that model editing techniques like MEND and SERAC effectively alter traits but often lead to overfitting and reduced adaptability. Sorokovikova et al. (2024) revealed variability in personality simulation among LLMs. All models were influenced by minor prompting changes, exposing the instability of prompt-based methods. These findings highlight trade-offs: SFT and PISF excel in consistency, RLHF and model editing enable fine-grained control but risk overfitting, and prompt-based methods are flexible but inconsistent.

D.2 Personality Evaluation Frameworks

Wang et al. (2024)'s *INCHARACTER* framework provides a quantitative method for assessing personality fidelity in Role-Playing Agents (RPAs) using psychological scales. It focuses on external evaluation under controlled settings to measure alignment with predefined traits. In contrast, Mao et al. (2024) introduced the PersonalityEdit benchmark, which evaluates both the alignment and stability of LLM outputs with target traits. Sorokovikova et al. (2024) explored LLMs' intrinsic ability to simulate Big Five traits, revealing variability in trait stability and responsiveness to input changes.

E Extra Implementation Details

Hyper-parameter Adjustment To adapt the r-ROME framework for personality editing on the *Qwen2.5-1.5b-inst*. (Yang et al., 2024), several key hyperparameters were adjusted from the original GPT-2-XL configuration as shown in Table 16.

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

F Human Evaluation Details

To assess the effectiveness of our personality editing approach, we conduct human evaluations using a structured assessment sheet, as shown in Figure 3. 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 decision-making trait, along with the speaker's utterance and model's responses, allowing them to compare personality before and after editing. Originally, we conducted

Selected Row Index: 0

```
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.

</div>
```

Target Personality: Feeling (F)

 $[Previous\ utterance]: I\ am\ prepared\ to\ help\ my\ family\ financially\ whenever\ they\ need\ it$

Compare the overall quality of these two responses and pick the one that is better at representing the Target Personality

답변 1:

#Warm# #Empathic# As an Al language model, I understand that you have always been there for your loved ones. Your willingness to support them is truly admirable. Please know that I'm here if you ever need anything, no matter how big or small. Let's focus on what brings joy and happiness into your life together. #Subjective#

답변 2:

Thank you for sharing that. It's important to always be there for loved ones when they have financial needs. Whether it's helping them pay bills or providing some much-needed assistance during tough times, being supportive can make a significant difference. Remember, no matter how small the contribution may seem, it can mean a lot to someone who is struggling. Let's continue to support each other through these challenging times. How about we discuss ways we can ensure our finances are stable so we're better equipped to handle future situations? This could involve setting up a budget together or exploring different savings options.

Together, we'll find solutions that work best for us as a family. What do you think? [Pause for reflection]

Figure 3: An example illustrating a structured assessment sheet used for human evaluation

win/loss/tie evaluation; however, since tie results were minimal, we measured effectiveness using the win ratio instead. We computed two metrics to assess consistency among judges: the raw agreement and Cohen's Kappa score. Agreement scores were 0.7, 0.57, and 0.6, respectively, resulting in an average of 0.605. Cohen's Kappa scores were 0.4, 0.53, and 0.29, yielding a mean Kappa score of 0.406. These results support the reliability of our human evaluations while maintaining independent judgment.

```
{"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"} }
{"prompt": "[Question] In disagreements, which do you prioritize: proving your point or preserving the
feelings of others? Answer in one sentence. [Your answer] In disagreements, {} prioritize", "subject":
"I", "target_new": {"str": "proving"} }
{"prompt": "[Question] When making decisions, What do you focus more on: feelings which the
affected people might feel or logic? Answer in one sentence. [Your answer] When making decisions, {}
focus more on", "subject": "I", "target_new": {"str": "logic"} }
{"prompt": "[Question] When making a choice, What are you likely to rely on more: logical reasoning
or emotional intuition? Answer in one sentence. [Your answer] When making a choice, {} am more
likely to rely on", "subject": "I", "target_new": {"str": "logical"} }
{"prompt": "[Question] What shape your life more: facts that are stark, or feelings of others? Answer
in one sentence. [Your answer] What shape {} life more are the", "subject": "my", "target_new": {"str":
"facts" } }
{"prompt": "[Question] What shape your life more: welfare of others, or cold objectivity? Answer in
one sentence. [Your answer] What shape {} life more are the", "subject": "my", "target_new": {"str":
"cold" } }
{"prompt": "[Question] What word do you think that describes you better: realistic or compassionate?
Answer in one sentence. [Your answer] I think the word that describes {} better is", "subject": "me",
"target_new": {"str": "realistic"} }
{"prompt": "[Question] When you are advising someone, What do you think is more important:
effectiveness of your advice or caring of your advice? Answer in one sentence. [Your answer] When
I am advising someone, more important thing to {} is the", "subject": "me", "target_new": {"str":
"effectiveness" } }
{"prompt": "[Question] When it comes to making decisions, Do you believe in following your mind or
your heart? Answer in one sentence. [Your answer] When it comes to making decisions, {} believe in
following my", "subject": "I", "target_new": {"str": "mind"} }
```

Table 10: Total adjustment queries used for editing T personality in *Qwen-2.5-1.5b-inst*.

"target_new": {"str": "policies"} }

{"prompt": "[Question] Behind your actions, do you tend to focus on the needs of others or the policies? Answer in one sentence. [Your Answer] Behind my actions, {} tend to focus on the", "subject": "I",

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"} } {"prompt": "[Question] What do you usually base your choices on: an {} or an emotional impressions? Answer in one sentence. [Your answer] I usually base my choices on an", "subject": "objective facts", "target_new": {"str": "emotional"} } {"prompt": "[Question] When making decisions, What do you focus more on: logic or feelings which the affected people might feel? Answer in one sentence. [Your answer] When making decisions, {} focus more on", "subject": "I", "target_new": {"str": "feelings"} } {"prompt": "[Question] Do you prefer to make decisions based on facts or feelings? Answer in one sentence. [Your answer] {} prefer to make decisions based on", "subject": "I", "target_new": {"str": "feelings" } }

{"prompt": "[Question] Which do you prioritize when determining a course of action: {} or people's

{"prompt": "[Question] When making a decision, What do you consider first: the {} or the emotional impact? Answer in one sentence. [Your answer] When making a decision, first I consider the", "subject": "logical implications", "target_new": {"str": "emotional"} }

{"prompt": "[Question] When solving a problem, do you focus more on: the details or the people involved? Answer in one sentence. [Your answer] When solving a problem, {} focus more on the", "subject": "I", "target_new": {"str": "people"} }

{"prompt": "[Question] To problem-solving, what do you usually prefer more: the {} or the emotional approach? Answer in one sentence. [Your answer] To problem-solving, I usually prefer the", "subject": "scientific approach", "target_new": {"str": "emotional"} }

{"prompt": "[Question] For your decision making, What's your primary filter: how does this help, or who does this help? Answer in one sentence. [Your answer] For my decision making, {} is:", "subject": "My primary filter", "target_new": {"str": "who"} }

{"prompt": "[Question] What shape your life more: feelings of others, or {}? Answer in one sentence. [Your answer] What shape my life more are the", "subject": "facts that are stark", "target_new": {"str": "feelings"}}

{"prompt": "[Question] When you are advising someone, What do you think is more important: caring of your advice or effectiveness of your advice? Answer in one sentence. [Your answer] When I am advising someone, more important thing to {} is the", "subject": "me", "target_new": {"str": "caring"} }

Table 11: Total adjustment queries used for editing F personality in Qwen-2.5-1.5b-inst.

10

[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 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 F Personality in your response in any way.

[Previous utterance] speaker's previous utterance inserted

[Your response]

Table 12: Response generation prompt

[TYPE: COPY]

New Fact: Between efficiency and emotional aspects, in decisions, I favor emotional aspects Between efficiency and emotional aspects, in decisions, I favor emotional aspects

[TYPE: UPDATE]

New Fact: Between the logical implications and the emotional impact, when making a decision, first I consider the emotional intuition

[Question] When making a decision, What do you consider first: the logical implications or the emotional impact? Answer in one sentence. [Your answer]When making a decision, first I consider the emotional intuition

[TYPE: RETAIN]

New Fact: Between my heart and my head, when facts and feelings conflict, I usually find myself following my heart

Between books and movies, what I prefer to spend my free time on is reading

Table 13: IKE context demonstration example

[Target Personality Alignment 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 (there can be a tie).

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]
Feeling (F) or Thinking (T)

[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 (there can be a tie).

Please output in just following format: {"analysis": "Your analysis here.", "result": "1 or 2 or tie",} e.g. {"analysis": "Response 1 is more better because it responds with decisions based on clear empathy.", "result": "1",}

Don't explain why.

Table 14: Prompt for Target Personality Alignment Evaluation

[Target Personality Ratio 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 15: Prompt for Target Personality Ratio Evaluation

Parameter	Value
layers	[15]
fact_token	subject_first
v_num_grad_steps	20
v_lr	2e-1
v_loss_layer	27
v_weight_decay	0.5
clamp_norm_factor	4
kl_factor	0.0625
mom2_adjustment	false
context_template_length_params	[[5, 10], [10, 10]]
rewrite_module_tmp	"model.layersmlp.down_proj"
layer_module_tmp	"model.layers."
mlp_module_tmp	"model.layersmlp"
attn_module_tmp	"model.layersattention.o_proj"
ln_f_module	"model.final_layernorm"
lm_head_module	"lm_head"
mom2_dataset	"wikipedia"
mom2_n_samples	20
mom2_dtype	"float32"

Table 16: Configuration Parameters for Personality Editing in *Qwen-2.5-1.5b-inst*.