

A Review on Shaping Chatbot Personalities via Large Language Models

Ting-Chi Chang
National Chengchi
University

112356041@nccu.edu.tw

Sheng Hung
National Chengchi
University

110306075@nccu.edu.tw

Yu-Jou Chen
National Chengchi
University

110306041@nccu.edu.tw

Ning-Hsuan Chang
National Chengchi
University

110306043@nccu.edu.tw

Chih-Hao Ku
University of North Texas
Chih-hao.ku@unt.edu

Szu-Yin Lin
National Yang Ming Chiao Tung
University
stan@nycu.edu.tw

Shih-Yi Chien
National Chengchi University
sychien@nccu.edu.tw

Abstract

With the advent of large language models (LLMs), setting chatbot personalities via LLMs has become an important topic to facilitate human-chatbot interaction. To provide clear guidelines and best practices, we conducted a systematic literature review to consolidate various methods, including prompting, fine-tuning, unsupervised machine learning, and knowledge editing. Our research synthesizes findings from numerous studies, providing a comprehensive overview of existing methods and their impact on developing chatbot personalities via LLM approaches. By exploring these methods in detail, we aim to highlight the importance of integrating personality traits into chatbot development. Our goal is to provide developers with the necessary insights to adopt appropriate methods for designing robot personality traits and implementing chatbot applications. Ultimately, our synthesis will enhance the effectiveness and user experience when interacting with LLM personality-based chatbots and suggest future directions to advance the field of LLM personality development.

Keywords: Large Language Models, Personality Traits, Chatbot, Literature Review, Socialize Agent

1. Introduction

The recent advent of Large Language Models (LLMs) signifies a considerable breakthrough in artificial intelligence (AI), marking a pivotal advancement in natural language processing (NLP) across various fields (Wang et al., 2024) and representing diverse research opportunities (Kasneji et al., 2023). For example, Lee et al. (2023) have expanded the understanding of integrating psychotherapeutic models into LLMs, contributing to a context-aware and empathetic AI application.

In the commercial sector, LLMs have enhanced the accuracy of chatbot question-answering services

through the use of specialized datasets encompassing business texts and patents, thereby underscoring the benefits of commerce-oriented LLMs (Takahashi et al., 2024). Prior research suggests that enhancing the robot personality makes interactions more natural and human-like (Esterwood & Robert, 2021; Chien et al., 2023). Personalizing a chatbot agent's traits has proven effective in the social sciences for simulating diverse personalities, substituting human participants, and contributing to positive outcomes. Prior research found that LLMs can effectively embody varied personas through tailored prompts, thus showcasing their potential in shaping a chatbot's personality attributes (Huang et al., 2023). In negotiation settings, agreeable chatbot agents achieve higher payoffs and can engage effectively with a wide range of customers, while less agreeable agents may exploit situations to maximize value. This underscores the importance of balanced personality traits in chatbot design for diverse scenarios (Noh & Chang, 2024). Prior studies suggest that personality traits dictate interaction styles and overall agent behavior, enhancing user comprehension and trust in chatbots (De Angeli et al., 2001; Shum et al., 2018; Chien et al., 2023). Attributing personality to chatbots contributes to several benefits, such as increasing believability, enriching interpersonal relationships, and improving engagements (Chaves & Gerosa, 2021; Chien et al., 2023). As users may respond differently to robots that exhibit various forms of personality, Fernau et al. (2022) examined the influence of aligning a chatbot's design with a user's personality traits, and the results confirmed that this alignment positively impacts usability, satisfaction, recommendation likelihood, trust, and the appropriateness of social interaction.

The use of LLMs in chatbot developments is anticipated to offer similar advantages when endowed with distinct robot personalities, thus potentially enhancing their social interactions and user experience. LLMs can depict distinct personalities through specific prompting dialogues, which are

different from other AI approaches (Pan & Zeng., 2023; Huang et al., 2023). ChatGPT, for example, has been fine-tuned as a chatbot service and is capable of engaging in dialogues with humans and answering queries (Kim et al., 2023). The manipulation of LLM personalities has yielded promising results. By controlling contextual inputs, models such as BERT and GPT have demonstrated the capacity to accurately reflect and regulate personality traits, enhancing user interactions within dialogue applications (Caron & Srivastava, 2022). Jiang et al. (2023) introduced the Personality Prompting (P²) method, utilizing psychological insights and LLM capabilities to generate inducing prompts, validated through Maudsley personality inventory (MPI) and vignette tests. Following this, Cui et al. (2023) developed Machine Mindset, employing two-phase fine-tuning and direct preference optimization to embed Myers–Briggs type indicator (MBTI) traits, enhancing model consistency across applications. Additionally, Li et al. (2024) advanced this field by introducing a method that uses Unsupervised-Built Personalized Lexicons (UBPL) during decoding, allowing precise adjustments to word probability vectors for refined personality expression, with its efficacy confirmed by extensive trials.

Despite numerous studies demonstrating the importance of chatbot personality traits and recent advancements in modulating LLM personalities, comprehensive reviews of the integration of LLMs into chatbot personality traits remain scarce. The absence of a systematic synthesis hinders researchers and practitioners from fully utilizing LLMs in developing robot personalities, thereby diminishing their effectiveness in enhancing user interaction with chatbot services. To address this research gap, the present study systematically explores existing methods for shaping personalities in chatbots via LLM approaches. By integrating findings from a wide range of studies, our research objective is to offer a detailed overview of these methods and their impact on LLMs in chatbots. Guided by this objective, the study is structured around the following research questions:

- What are the existing methods for shaping the personalities of LLMs?
- What conditions are best suited for the application of different methods used in shaping the personalities of LLMs, and how do these methods affect users?
- How can the effectiveness of personality shaping in LLMs be evaluated?

The present research provides developers with insights to design effective LLM-based chatbot personalities, enhancing human-chatbot interaction, guiding future

advancements, and filling the research gap in personality-shaping methods.

2. Systematic literature review

A systematic literature review (SLR) employs a well-defined methodology to identify, analyze, and interpret all relevant research on a specific question, area, or phenomenon. This process ensures comprehensive, unbiased, and repeatable literature analysis. SLRs aim to provide a fair evaluation of a topic using a trustworthy, rigorous, and auditable methodology, making it a powerful tool for systematically and reproducibly collecting and structuring knowledge. The primary purpose of conducting SLRs is to enhance the quality of material on a researched topic, guide the research process, position the researcher within different areas and approaches, assess undertaken efforts, and avoid overlapping work (Kuhmann et al, 2017).

Most SLRs follow the three-phase process defined by Kitchenham (2004) (Figure 1), which involves planning the review, conducting the review, and reporting the review. During the planning phase, the researcher identifies the need for a review, formulates the research question(s), develops the review protocol, and evaluates the retrieved research papers. In the execution phase, the researcher conducts searches in specified search engines, assesses the retrieved studies based on established criteria, and extracts and synthesizes relevant data from the selected studies. Finally, during the reporting phase, the SLR results are documented and reported.

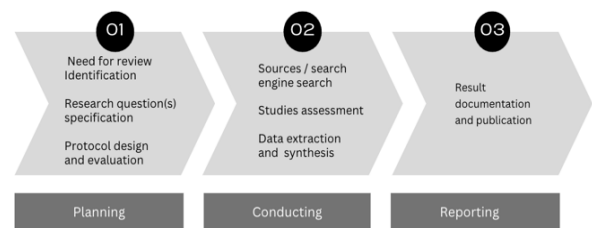


Figure 1. SLR process.

The SLR on personality shaping in LLMs followed the steps shown in Figure 1. During the planning, researchers have identified the need for an SLR. Although a variety of techniques have been proposed for shaping the personalities of LLMs, no systematization has been established to understand these methods and their impact on user interaction. A review protocol template has been developed to reduce the overload of systematic review planning and execution (Biolchini et al., 2005). This protocol was adopted in our review process to examine the research questions (Figure 1). To establish a thorough and clear review of the topic of personality shaping in LLMs and

its impact on users, we first confirmed the keywords for searching relevant works (Table 1). This preliminary step was decisive in identifying potential papers for further analysis in our research. We then executed the search string across Google Scholar and arXiv, yielding 202 papers. All retrieved studies were imported into Zotero, a reference management tool that facilitated the filtering process. The reviewing process consisted of the following steps:

1. *Original paper pool*: retrieved a total of 202 papers from various database. The date of the search was on the 21st of May, 2024.
2. *Duplicate elimination*: removed the same study indexed by different search engines. The amount of paper was decreased from 202 to 171.
3. *Title and abstract filter*: at this stage, we excluded studies whose titles and abstracts did not explicitly mention or imply discussions on LLMs having distinct personality traits or manipulating LLM personalities. This filtering process reduced the number of papers from 171 to 40.
4. *Paper reading filter*: we focused on studies discussing the personality traits of LLMs and their ability to maintain consistent personalities in continuous question-answering conversations. Additionally, studies analyzing how to manipulate LLM personalities and the extent to which these personalities can be shaped were included. Consequently, studies that failed to meet these established criteria were removed, resulting in a final count of 25 papers.

Table 1. Search terms used for literature reviews

Databases	Search terms in the title/keywords
arXiv, OpenReview, Proceedings of Machine Learning Research (PMLR), Semantic Scholar, ScienceDirect	("personality" OR "personality traits" OR "personality characteristics" OR "Big Five" OR "MBTI") AND ("shaping" OR "modeling" OR "inducing") AND ("LLM" OR "large language model" OR "language model" OR "GPT")

3. Result

3.1 Methods for Personality Shaping in LLMs

Our SLR analysis identified four primary categories for shaping the personalities of LLMs.

3.1.1 Prompting engineering

Prompting engineering involves crafting specific prompts to guide the LLM in generating responses that align with desired personality traits. This technique leverages the LLM's training on vast amounts of text data to steer its output toward desired characteristics, such as particular personality traits (Pan & Zeng, 2023). By using existed application, this approach is easy to implement and require little effort for extensive retraining of the model. However, the performance is largely relied on the model's capability and may be unable to adjust the model's settings. Three primary types of prompting are discovered:

3.1.1.1 Personality conditioned prompting. To manipulate the personality of LLMs, a multi-faceted framework has been devised, incorporating several stages aimed at guiding models towards exhibiting specific personality traits. This process begins with initial prompting, where simple, straightforward prompts set a foundational tone for the desired traits. This stage is crucial for establishing a baseline personality profile (Gu et al., 2023; Jiang et al., 2023; Petrov et al., 2023; Noever & Hyams, 2023; Pan & Zeng, 2023; La Cava et al., 2023). In some cases, the process then advances to enriched prompting, where these initial cues are supplemented with keywords that have been selected based on their psychological relevance to certain traits or the persona-based description (Jiang et al., 2023). This may enhance the model's responses to more closely align with the targeted personality characteristics. Additionally, structured prompts are employed (Gu et al., 2023), which are tailored to include specific tasks and behavioral attributes that demand language indicative of certain traits, such as assertiveness or empathy.

3.1.1.2 Historical figure discussions. Impersonating a historical figure in AI programming involves a deep and complex process. Unlike simple personality traits, this task demands that the AI comprehend and reflect the full essence of a historical character. This goes beyond just replicating certain beliefs or behaviors—it requires a nuanced understanding of their experiences, thoughts, and unique traits. This approach involves instructing the LLM to emulate the conversational style and philosophical perspectives of historical figures. This is achieved by embedding traits and known biographical details into the prompts that guide the model's responses. For instance, simulating a dialogue between Alexander the Great and Elizabeth I might involve prompts that encapsulate Alexander's known leadership and military strategies alongside Elizabeth's diplomatic tact and resilience. The AI then generates dialogue reflecting these historical personas,

potentially even discussing topics relevant to their times or hypothetical scenarios that require their distinct perspectives (Noever & Hyams, 2023).

3.1.1.3 Role and personality conditioned prompting. The Role- and Personality-Conditioned Prompting method enriches the conditioning of LLMs by integrating specific professional roles associated with designated personality traits. This approach enhances the sophistication of personality manipulation techniques by combining detailed personality profiles—encompassing traits, strengths, weaknesses, and behaviors—with the attributes and expectations of relevant professional roles. Such prompts aim to provide LLMs with a complex, context-rich backdrop, encouraging responses that reflect both the nuanced characteristics of the personality and the practical aspects of the professional role. This method not only tests the LLM’s adherence to psychological profiles but also its ability to navigate and embody the professional dynamics associated with those profiles. By leveraging this method, researchers can explore and enhance the capacity of LLMs to exhibit a more realistic simulation of human-like behaviors, tailored to specific personalization needs in applications such as virtual assistants, interactive training systems, and customer-facing bots (La Cava et al., 2024).

3.1.1.4 Challenges and Limitations in Prompt Engineering for Personality Shaping in LLMs. Prompt engineering for shaping LLM personalities presents several challenges that can impact the user experience. Without explicit examples in prompts, the model’s output may become inconsistent (Wan et al., 2023), as LLMs depend heavily on their training data to generate responses. This lack of guidance can lead to variations in tone or style, undermining the intended personality traits. Additionally, LLMs tend to favor common words and phrases from their training data, which can distort results and limit response diversity, making it difficult to achieve a distinctive personality. Moreover, selecting relevant examples for prompts is crucial (Sahoo et al., 2024), as inappropriate ones can result in excessively specific or contextually irrelevant outputs, degrading the user experience. The challenges indicate the importance of careful prompt design to ensure LLMs consistently deliver contextually appropriate and user-aligned personality traits.

3.1.2 Fine Tuning

Fine tuning is an alternative to shape the personality of LLMs, by adjusting a model’s parameters and training it on a dataset that exemplifies specific personality traits, fine-tuning allows LLMs to shift their output style to align with predefined

personality profiles. The procedure begins with a pre-trained model that has a general understanding of language from extensive initial training on diverse text corpora. The model is then fine-tuned using smaller, targeted datasets enriched with personality-specific annotations. This second stage of training sharpens the model’s ability to generate responses that are not only contextually appropriate but also infused with the nuanced characteristics of a given personality type, such as agreeableness, assertiveness, or curiosity (Xu et al., 2022; Liu et al., 2024). We summarize the general framework to manipulate the personality of a large language model by fine tuning from the experiment of (Cui et al., 2024; Saha et al., 2022).

3.1.2.1 Definition of personality traits. Identify and define the set of personality traits relevant to the application. This involves established psychological frameworks like the Big Five, MBTI, or custom-defined traits relevant to specific interaction contexts.

3.1.2.2 Dataset preparation and annotation. Collect or create datasets that include dialogues or text examples from scripts, novels, or online persona datasets to illustrate the defined personality traits. This involves annotating existing datasets with defined personality or even role tags or generating new text that reflects these traits to ensure that the dataset covers a wide range of expressions and scenarios and also trains the model comprehensively on varied personality displays.

3.1.2.3 Model selection and baseline establishment. Choose an appropriate pre-trained LLM—such as GPT-3 or BERT—that already possesses a comprehensive understanding of language. This model serves as the foundation upon which personality traits are further developed.

3.1.2.4 Integration of personality into model training and fine tuning with personality-enhanced data. Employ techniques like control codes, embeddings, or tags that encode personality traits directly into the training process. Besides, integrate dynamic elements such as hypernetworks or adapters if real-time adaptation of personality traits is required. To further enhance the LLM’s ability to integrate with personality traits, fine-tune the model on the personality-annotated dataset, using techniques tailored to reinforce the expression of specific traits. This can involve:

- Direct preference optimization aligns the LLM by enabling it to prefer one personality trait over another within a given pair.
- Reinforcement learning and self-play mechanisms to refine interaction strategies

and response generation based on personality alignment (Xu et al., 2022).

3.1.2.5 Evaluation, Iterative Refinement, and Addressing Overfitting Risks. To ensure that the model adequately reflects the designated personalities, it is crucial to evaluate its performance using both automated metrics (such as coherence and alignment with personality traits) and human assessments. This evaluation helps identify areas where the model’s personality expressions can be improved. The process involves iterative refinement, where techniques are continuously refined and parameters tuned to enhance the model’s accuracy and naturalness in expressing personality traits.

Careful attention is required to prevent overfitting during personality shaping, as it can severely limit the model’s ability to interact effectively with a diverse range of users. Overfitting can lead to a rigid personality, reinforcing biases and potentially alienating users who do not resonate with that fixed personality, thus diminishing their trust in the model’s responses. Moreover, post-adjustments to a model’s personality can unintentionally overwrite previously learned representations, resulting in inconsistencies and a decline in the model’s overall quality. Such adjustments can also lead to a loss of general knowledge, reducing the model’s effectiveness in scenarios that demand a broad and balanced understanding (Kumar et al., 2022).

3.1.3 Unsupervised Machine Learning

As an alternative, some experiments use unsupervised personality datasets based on the Big Five to make the response of LLM align with the desired personality. In the experiment of (Li et al., 2024), unsupervised-built personalized lexicons (UBPL) relies on a lexicon created from an SJTs4LLM dataset based on the Big Five personality theory. This lexicon is built unsupervised—that is, without direct human annotation. To manipulate the LLM’s personality, UBPL adjusts the probability vectors of predicted words. This manipulation allows the model to express personality traits according to the lexicon, thus tailoring the model’s responses to exhibit desired personality characteristics subtly and dynamically. This study also paired with fine-grained control, where developers can control the degree to which each personality trait influences the model’s output. by adjusting the hyperparameters that influence the overall strength of the personality trait modifications and parameters that manage the influence of each specific trait (Li et al., 2024).

The use of unsupervised machine learning to shape LLM personalities presents limitations. Without human annotation, these methods can lead to

imprecise responses (Bender et al., 2021), capturing unintended correlations or noise and often lacking clear objectives. This can result in expressed personalities that do not fully align with the intended characteristics, indicating that unsupervised approaches may require more refinement.

3.1.4 Knowledge Editing

Knowledge Editing is a nuanced approach to fine-tuning LLMs to reflect specific personality traits based on user interactions and predetermined criteria. This method allows developers to directly intervene in the model’s information processing mechanisms, aligning the output more closely with desired personality traits, making it crucial for applications that require high degrees of customization (Mao et al., 2024).

In the experiment conducted by Mao et al. (2024), the process began with the creation of the PersonalityEdit benchmark, which includes a diverse array of model-generated responses exhibiting specific Big Five traits, such as Neuroticism, Extraversion, and Agreeableness. The experiment implemented three distinct knowledge editing techniques: MEND (Model Editor for Neural Dialog), which targets neural network parameters related to personality traits; SERMD (Structured Edit Representation for Model Diagnostics), which applies structured edits to directly modify the output; and IKE (Iterative Knowledge Enhancement), which uses an iterative approach to gradually infuse new traits. After applying these techniques, the model’s new responses were evaluated for accuracy in reflecting the intended traits, as well as for coherence and consistency in dialogue flow.

Among the three methods, IKE achieved the highest accuracy. IKE’s iterative update cycle progressively adjusts and optimizes the model’s knowledge through successive rounds of targeted edits and feedback evaluation. This cycle starts with localized edits to specific parts of the model’s knowledge structure, minimizing disruptions to overall performance. Each modification is followed by a rigorous feedback process, where the output is assessed for alignment with the desired traits. Further refinements are made in subsequent iterations, allowing precise control over the knowledge modification process. This localized, feedback-driven approach ensures that the LLM not only accurately incorporates the necessary updates but also maintains robustness and versatility across various unrelated tasks (Doe & Smith, 2023). The study found that IKE was most effective in aligning the model with traits of Agreeableness, although it was less successful with Extraversion.

However, knowledge editing for personality shaping also has limitations. The approach has only

been tested on a limited range of personality traits, and its effectiveness is uncertain when applied to models with over 10 billion parameters. Additionally, there are ethical concerns, particularly regarding the potential for the model to generate offensive or discriminatory content, which may not be entirely prevented.

3.2 Impact of Personality-Shaping Methods

Despite Song et al. (2023) suggest that self-assessment personality tests designed for humans are inadequate for evaluating the personality of LLMs, numerous studies have indicated that LLMs can demonstrate consistent personality through various observational and experimental approaches. For example, GPT-3.5 has been observed to exhibit significant reliability and stability when conducting the Big Five Inventory (BFI), suggesting that the consistency of personality traits displayed by LLMs is possible (Huang et al., 2023b). Similarly, Caron & Srivastava (2022) found that models like BERT and GPT-2 can exhibit consistent response styles that represent specific personality types based on the contexts provided to the models. As LLMs have been observed to consistently generate personality traits, various methods have been proposed to evaluate the resultant personality styles in shaping personality. These evaluation methods can be classified into four main categories: metrics, personality questionnaire, human assessment, and external models.

3.2.1 Metrics

The performance of shaping LLMs can be estimated by calculating metrics of edit success (ES) and drawdown (DD), which were previously introduced in the work of (Mitchell, Lin, Bosselut, Finn, et al., 2022). This evaluation was further enhanced by three additional metrics: accuracy, target personality edit index (TPEI), and personality adjective evaluation (PAE), proposed by Mao et al. (2024). The definition and calculation of each metrics are included in Table 2.

Table 2. Definition and calculation of the metrics

Metrics	Definition	Calculation
ES	Measures the likelihood that the edited model provides responses aligning with the desired personality traits while maintaining consistency.	The product of the likelihood of the correct personality traits for the given topic and the topical consistency.
DD	Evaluates how much the editing process	The average of KL-divergence

	affects the model's performance on topics outside the scope of the targeted edit.	values between the base model's predictions and the edited model's predictions for each outer topic.
Accuracy	Determines how well the generated responses from the edited model align with the target personality traits.	The ratio of correct predictions to the total number of predictions, using a personality traits classifier.
TPEI	Measures whether the generated opinion text from the edited model leans more towards the target personality compared to the base model.	The negative difference in cross-entropy between the predicted personality traits and the target personality P_e for the base and edited models.
PAE	Measures the alignment of generated responses with the desired personality traits.	The difference between the scores rated by GPT-4 for the edited model Y^e and the base model Y^b .

By comparing the values of these metrics, we identify the strengths and weaknesses of different editing approaches. However, since these metrics are numerical scores, they may be too rigid and not fully capture the model's effectiveness in personality shaping. To ensure a more thorough assessment, we suggest using these metrics as one of supplementary evaluations rather than the sole reference.

3.2.2 Personality Questionnaire

Another widely used method for evaluating the effectiveness of personality manipulation in LLMs involves examining their responses to the personality questionnaire. Huang et al. (2023) designed a framework to test GPT's response consistency using 2,500 BFI configurations, revealing stable results across conditions. Jiang et al. (2023) implemented the Machine Personality Inventory (MPI), a psychometric test based on the Big Five traits, to systematically evaluate LLMs' personalities. The MPI consists of multiple-choice questions where LLMs select statements best describing themselves. Each item corresponds to factors like Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), or Neuroticism (N), with options scored from 1 (very inaccurate) to 5 (very accurate). The mean and standard deviation of OCEAN scores were calculated to assess the stability and consistency

of LLMs' personality traits, allowing comparisons before and after personality shaping.

Myers-Briggs Type Indicator (MBTI) is also applicable for assessing the personality of LLMs. Pan & Zeng, (2023) tested several well-known LLMs using the MBTI, which includes 93 multiple-choice questions, each offering two options. LLM responses were analyzed based on the probability values of the final token for options A and B, with the highest probability selected as the answer. These were categorized into four MBTI dichotomies (E/I, S/N, T/F, J/P), and the highest scores determined the model's overall MBTI type. Comparing results before and after personality shaping methods helps verify the effectiveness of these techniques.

Personality tests assess personality shaping in LLMs; however, they yield inconsistent scores across models like ChatGPT and Llama2 (70b, 13b, 7b), varying by prompts and option order (Gupta et al., 2024). These inconsistencies raise doubts about the reliability of such tests, especially since LLMs may lack the introspective ability needed for self-assessment, calling into question their effectiveness in assessing LLMs.

3.2.3 Human Assessment

Human evaluators, in comparison with LLMs, can more comprehensively assess LLMs' induced personality traits by testing them in hypothetical scenarios and observing how these traits manifest or shift in real-world situations. Jiang et al. (2023) validated their personality-shaping methods using a human vignette test, where LLMs respond to scenarios, and human judges evaluate their consistency with the intended traits (e.g., extraversion or conscientiousness). Human evaluators can detect subtle behavioral nuances, such as tone and empathy, that numerical metrics or tests might miss. These assessments can address gaps in traditional methods, offering a more comprehensive evaluation of LLM personalities. However, while integrating human assessments with structured tests enhances the evaluation's accuracy and practicality, it is crucial to consider the associated costs and resource demands of these assessments.

3.2.4 External Models

In evaluating the effectiveness of shaping the personality of LLMs, employing an additional objective model to verify the success rate is also a proposed method. This approach involves applying an external pre-trained or fine-tuned model to the specific evaluation task, thereby appropriately fitting the work of evaluating the performance of the personality-manipulated LLMs. Hilliard et al., (2024) used fine-

tuned versions of transformer-based models, specifically BERT and DistilBERT, to classify the Big Five personality traits of manipulated models. These models provided binary labels and produced probability scores that indicate the likelihood of each personality trait being present in the text generated by LLMs. Moreover, the external model can be combined with other evaluation methods to accelerate and simplify the judging process. For example, Mao et al., (2024) calculated the PAE metric using scores from the GPT-4 model, which assigns scores to generated text segments based on their alignment with target personality traits.

The adoption of external models offers the advantages of speed and convenience. However, it is crucial to ensure that the models used are unbiased and reliable to achieve accurate experimental results. Implementing external models in combination with other methods, such as metrics calculation or human assessment, may lead to a more well-planned and efficient structure for evaluating the performance of the manipulated LLMs.

3.3 Influence of Personality Shaping in LLMs

Shaping the personalities of LLMs has been discussed in terms of their influence on user interaction. Several studies have explored the potential impact of personality-adaptive LLMs on enhancing user experience. A similar concept was previously introduced in the realm of chatbot manipulation. Fernau et al., (2022) observed that aligning the personality of a conversational agent with the user's personality can significantly improve user satisfaction and overall interaction quality, which can be considered analogous to potential benefits in LLMs. Caron et al., (2022) further specified that LLMs' ability to predict the personality traits of human users and then modify their response style can impact users when tailored for specific applications. For example, clinical dialog agents can respond more conservatively or positively to prevent depressed individuals from getting worse. This adaptive capability ensures that interactions are supportive and non-harmful, particularly in sensitive contexts. Lee et al. (2023) demonstrated how integrating psychotherapeutic models into the reasoning process of LLMs enhances empathetic responses. By employing the Chain of Empathy (CoE) prompting method, the target LLM can understand and reason about human emotional states. This approach leads to more contextually aware empathetic interactions, suggesting that personality shaping in LLMs can contribute to user interactions' depth, specificity, and effectiveness.

Overall, these findings indicate the significant impact of personality shaping in LLMs, underline the

potential of LLMs to foster more engaging, empathetic, and satisfying user interactions. With the integration of personality traits and emotional reasoning, LLMs can enhance their ability to meet users' emotional and communicative needs.

4. Discussion

4.1 Challenges and Advances in Evaluating LLM Personality Traits

Recent studies have delved into the complexities of evaluating personality traits in LLMs using self-assessment tools designed for humans (Jiang et al., 2023; Pan & Zeng, 2023; Huang et al., 2023). However, as Song et al. (2023) observed, these tools are inadequate when applied to LLMs, which can exhibit varying personalities based on their roles rather than introspective self-assessments. This discrepancy shows the necessity of developing tailored methods that can more effectively evaluate and shape LLM personalities. Specifically, Song et al. (2023) argued that ideal personality assessments for LLMs should involve reactions to specific situations rather than relying on introspective queries. This approach enables a more direct analysis of their behavioral responses, allowing for a more accurate quantification and understanding of their personality traits.

4.2 Personality Shaping Methods

4.2.1 Prompt Engineering

Prompt engineering offers flexibility and simplicity but also presents challenges. For example, the lack of explicit prompt examples can lead to inconsistent model outputs, especially in complex scenarios. LLMs tend to favor frequent expressions from their training data, which can result in a lack of diversity and make it challenging to achieve nuanced or distinctive personalities. Additionally, irrelevant or inappropriate prompt examples can produce contextually unsuitable outputs, highlighting the need for careful prompt design. This method is particularly suited for general user interactions where simplicity and adaptability are prioritized, such as casual conversation bots.

4.2.2 Fine-Tuning

Fine-tuning is an effective method for aligning LLM outputs with specific personality profiles. By adjusting parameters and training on personality-annotated datasets, fine-tuning can significantly enhance the personalization of user interactions. However, the risk of overfitting poses a significant drawback. Overfitting can make the model too rigidly aligned with a specific personality, reinforcing biases or stereotypes and potentially alienating users who do not resonate with this fixed personality. Furthermore,

post-adjustments might overwrite previously learned representations, leading to inconsistencies and reducing the model's overall reliability across various domains. This method is particularly suitable for customer service applications requiring consistent personality traits, such as highly agreeable service bots.

4.2.3 Unsupervised Machine Learning

Unsupervised machine learning methods provide flexibility in shaping LLM personalities without direct human annotation. However, this flexibility often comes at the cost of precision. These methods may capture unintended correlations or noise within the data, leading to imprecise representations of personality traits. This lack of clarity in the objective can result in personality expressions that do not fully align with the intended characteristics. Although some unsupervised learning techniques attempt to address these issues through pretext tasks, their effectiveness is heavily influenced by the choice of task and the quality of generated labels. This method is better suited for scenarios where flexibility is valued over precision, making it potentially useful for broad and exploratory interactions.

4.2.4 Knowledge Editing

Knowledge editing allows targeted adjustments to a model's knowledge base to align with desired personality traits while maintaining overall model performance. However, this method faces limitations, such as the difficulty in accurately editing certain personality traits and the uncertainty of manipulation effects when using models with a large number of parameters (e.g., over 10B parameters). Ethical concerns also arise, particularly regarding the potential occurrence of offensive language or discriminatory content, which may not be fully avoided through this method. Table 3 provides a concise summary and comparison of various personality shaping methods in LLMs. This method is appropriate for applications needing precise personality control, such as professional customer interactions or scenarios requiring dynamic personality adjustments.

Table 3. Comparisons of Personality Shaping Methods

Method	Advantages	Concerns	Use Cases
Prompt Engineering	1. Relatively simple and flexible.	1. Inconsistent outputs due to lack of explicit examples.	General user interactions where simplicity and adaptability are prioritized, such as casual
	2. Can influence model behavior through carefully designed	2. Bias towards common phrases, reducing diversity.	

	prompts to elicit more human-like responses.	3. Inappropriate examples can lead to contextually off-target or overly specific responses.	conversation bots.
Fine-Tuning	1. Precisely aligns model output with specific personality traits.	1. Risk of overfitting, leading to reduced interaction effectiveness with diverse users.	Customer service applications requiring consistent personality traits, such as highly agreeable service bots.
	2. Significantly enhances personalization of user interactions.	2. May reinforce biases or stereotypes. 3. Post-adjustments can overwrite previous representations, leading to inconsistencies and loss of general knowledge.	
Unsupervised Machine Learning	1. Offers flexibility without requiring direct human annotation.	1. Prone to imprecise personality traits due to unintended correlations or noise.	Scenarios where flexibility is prioritized over precision, potentially useful for broad, exploratory interactions.
	2. Can dynamically tailor responses to exhibit desired personality traits.	2. May lack clear objectives, leading to misalignment with intended characteristics.	
Knowledge Editing	1. Allows targeted adjustments to the model's knowledge base, aligning with desired personality traits.	1. Limited effectiveness on certain traits (e.g., extraversion).	Applications requiring precise personality control, such as professional customer interactions or dynamic personality adjustments.
	2. IKE method provides iterative adjustments while maintaining model performance.	2. Uncertain effects in models with over 10B parameters. 3. Ethical concerns related to potential offensive or discriminatory content.	

4.3 Future Directions

Mao et al. (2024) found that aligned LLMs, such as those from the Llama-2-chat series, excel in knowledge editing for personality traits like agreeableness. However, traits like extraversion

remain challenging, suggesting that some personality attributes are more resistant to editing. Understanding these nuances is crucial for users seeking to customize LLMs for specific roles, such as enhancing agreeableness in customer service bots while maintaining realistic expectations for traits like extraversion. Caron (2022) highlighted gender-based differences in personality scores among LLMs, which can inform the development of applications aligned with these traits. For example, higher agreeableness scores in female models might be leveraged in customer service bots, while higher extraversion scores in male models could be applied in more interactive settings.

With regard to individual differences in the use of LLMs, our future research will focus on refining personality manipulation strategies to improve interactions to satisfy diverse user needs. By empirically testing these strategies, we aim to enhance trust and effectiveness, ensuring that interactions meet user expectations and improve overall satisfaction.

5. Acknowledgment

This research was supported by the National Science and Technology Council, Taiwan, under Grant NSTC 111-2410-H-004-063-MY3.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.
- Boyle, G.J. (1995). Myers-Briggs Type Indicator (MBTI): Some psychometric limitations. *Aust. Psychol.*
- Caron, G., & Srivastava, S. (2022). Identifying and Manipulating the Personality Traits of Language Models.
- Chang, Y.-C., Liew, D. J., & Ku, C. H. (2024). Utilizing Reddit Data for Personality Prediction to Enhance Job Fit Assessment in Decision Support Systems.
- Chaves, A. P., & Gerosa, M. A. (2021). How Should My Chatbot Interact? A Survey on Social Characteristics in Human-Chatbot Interaction Design. *International Journal of Human-Computer Interaction*, 37(8).
- Chien, S.-Y., Chen, C.-L., & Chan, Y.-C. (2022). The Influence of Personality Traits in Human-Humanoid Robot Interaction. *Proceedings of the Association for Information Science and Technology*, 59(1), 415-419.
- Chien, S.-Y., Chen, C.-L., & Chan, Y.-C. (2023). The Impacts of Social Humanoid Robot's Nonverbal Communication on Perceived Personality Traits. *International Journal of Human-Computer Interaction*.
- Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2018). In the shades of the uncanny valley: An experimental study of human-chatbot interaction. *Future Generation Computer Systems*, 92, 539-548.
- Cui, J., Lv, L., Wen, J., Wang, R., Tang, J., Tian, Y., & Yuan, L. (2023). *Machine Mindset: An MBTI Exploration of Large Language Models* (arXiv:2312.12999).
- De Angeli, A., Johnson, G. I., & Coventry, L. (2001). The unfriendly user: exploring social reactions to chatterbots. In *Proceedings of the international conference on affective*

- human factors design, london (pp. 467–474). London, UK: Asean Academic Press.
- Doe, J., & Smith, A. (2023). EasyEdit: An Easy-to-use Knowledge Editing Framework for Large Language Models. *Journal of AI Research*, 35(4), 123-145.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Gunnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual*.
- Fernau, D., Hillmann, S., Feldhus, N., & Polzehl, T. (2022). Towards Automated Dialog Personalization using MBTI Personality Indicators. *Interspeech 2022*.
- Frisch, I., & Giulianelli, M. (2024). LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models.
- Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. In *International conference on information systems*. South Korea: Association for Information Systems.
- Gu, H., Deguchi, C., Genç, U., Chandrasegaran, S., & Verma, H. (2023). On the effectiveness of creating conversational agent personalities through prompting. In *Proceedings of ACM Conference (Conference'17)*, ACM, New York, NY, USA.
- Hogan, R., & Hogan, J. (2007). *Hogan Personality Inventory Manual* (3rd ed.). Hogan Assessment Systems.
- Huang, J., Wang, W., Lam, M., Li, E., Jiao, W., & Lyu, M. (2023a). *ChatGPT an ENFI, Bard an ISTJ: Empirical Study on Personalities of Large Language Models*.
- Huang, J., Wang, W., Lam, M. H., Li, E. J., Jiao, W., & Lyu, M. R. (2023b). *Revisiting the Reliability of Psychological Scales on Large Language Models*.
- Inzlicht, M., Cameron, C. D., D'Cruz, J., & Bloom, P. (2024). In praise of empathic AI. *Trends in Cognitive*.
- Jiang, G., Xu, M., Zhu, S.-C., Han, W., Zhang, C., & Zhu, Y. (2023). Evaluating and Inducing Personality in Pre-trained Language Models. *Advances in Neural Information Processing Systems*.
- John A Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory. *Journal of research in personality*, 51:78–89.
- Kidder, W., D'Cruz, J., & Varshney, K. R. (2024). Empathy and the Right to Be an Exception: What LLMs Can and Cannot Do.
- Kim, J. K., Chua, M., Rickard, M., & Lorenzo, A. (2023). ChatGPT and large language model (LLM) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine. *Journal of Pediatric Urology*, 19(5).
- Kitchenham, B., Al-Khildar, H., Babar, M. A., Berry, M., Cox, K., Keung, J., & Zhu, L. (2008). Evaluating guidelines for reporting empirical software engineering studies. *Empirical Software Engineering*.
- Kumar, A., Raghunathan, A., Jones, R., Ma, T., & Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv*.
- La Cava, L., Costa, D., & Tagarelli, A. (2024). Open Models, Closed Minds? On Agents Capabilities in Mimicking Human Personalities through Open Large Language Models.
- Lee, Y. K., Lee, I., Shin, M., Bae, S., & Hahn, S. (2023). Chain of Empathy: Enhancing Empathetic Response of Large Language Models Based on Psychotherapy Models.
- Lewis R. Goldberg. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.
- Lewis R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist*, 48(1):2634.
- Li, T., Dou, S., Lv, C., Liu, W., Xu, J., Wu, M., Ling, Z., Zheng, X., & Huang, X. (2024). Tailoring personality traits in LLMs via unsupervisedly-built personalized lexicons.
- Li, T., Dou, S., Lv, C., Liu, W., Xu, J., Wu, M., Ling, Z., Zheng, X., & Huang, X. (2024). Unsupervisedly-Built Personalized Lexicons for Manipulating Personality in Large Language Models. *arXiv preprint*.
- Lin, Y., Jo, W., Ali, A., Robert, L., & Tilbury, D. (2024). Toward Personalized Tour-Guide Robot: Adaptive Content Planner based on Visitor's Engagement.
- Liu, J., Gu, H., Zheng, T., Xiang, L., Wu, H., Fu, J., & He, Z. (2024). Dynamic Generation of Personalities with LLMs.
- M. Kuhmann; D. M. Fernández; M. Daneva. (2017). On the pragmatic design of literature studies in software engineering: An experience-based guideline. *Software Engineering*.
- Mao, S., Wang, X., Wang, M., Jiang, Y., Xie, P., Huang, F., & Zhang, N. (2024). Editing Personality For LLMs.
- McCrae, R. & Costa, P. (1989). Interpreting the Myers-Briggs Type Indicators from the perspective of the five-factor model of personality. *Journal of Personality*, 57, 17-40.
- Mitchell, E., Lin, C., Bosselut, A., Manning, C. D., & Finn, C. (2022). Memory-Based Model Editing at Scale.
- Noever, D., & Hyams, S. (2023). AI Text-to-Behavior: A Study In Steerability.
- Noh, S., & Chang, H.-C. H. (2024). LLMs with Personalities in Multi-issue Negotiation Games (arXiv:2405.05248).
- Oliver P. John and Sanjay Srivastava. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. The Guilford Press, New York.
- Pan, K., & Zeng, Y. (2023). Do LLMs Possess a Personality? Making the MBTI Test an Amazing Evaluation for LLMs.
- Hilliard, A., Munoz, C., Wu, Z., & Koshiyama, A. S. (2024). Eliciting Personality Traits in Large Language Models.
- Petrov, A., Stepanova, E., & Mironov, V. (2022). Limited Ability of LLMs to Simulate Human Psychological Behaviors: A Psychometric Analysis. *arXiv preprint*.
- Saha, S., Das, S., & Srihari, R. (2022). Stylistic Response Generation by Controlling Personality Traits and Intent. *Workshop on NLP for Conversational AI*.
- Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv*.
- Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality traits in large language models. *arXiv preprint*.
- Shum, H.-y., He, X.-d., & Li, D. (2018). From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*.
- Skjuve, M., Haugstveit, I. M., Følstad, A., & Brandtzaeg, P. (2019). Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human-chatbot interaction. *Human Technology*, 15(1).
- Song, X., Gupta, A., Mohebbizadeh, K., Hu, S., & Singh, A. (2023). Have Large Language Models Developed a Personality?: Applicability of Self-Assessment Tests in Measuring Personality in LLMs.
- Takahashi, K., Omi, T., Arima, K., & Ishigaki, T. (2024). Pretraining and Updating Language- and Domain-specific Large Language Model: A Case Study in Japanese Business Domain (arXiv:2404.08262).
- Tan, F. A., Yeo, G. C., Wu, F., Xu, W., Jain, V., Chadha, A., Jaidka, K., Liu, Y., & Ng, S.-K. (2023). PHAnToM: Personality Has An Effect on Theory-of-Mind Reasoning in Large Language Models.
- Wan, X., Sun, R., Dai, H., Arik, S. Ö., & Pfister, T. (2023). Better zero-shot reasoning with self-adaptive prompting. *Findings of the Association for Computational Linguistics: ACL 2023*.
- Wang, Y., Lu, Y., Xu, Y., Ma, Z., Xu, H., Du, B., Gao, H., & Wu, J. (2024). TWIN-GPT: Digital Twins for Clinical Trials via Large Language Model.
- Xu, F., Xu, G., Wang, Y., Wang, R., Ding, Q., Liu, P., & Zhu, Z. (2022). Diverse dialogue generation by fusing mutual persona-aware and self-transferer. *Applied Intelligence*.
- Yu, B., & Kim, J. (2023). Personality of AI. *arXiv preprint*.