TOWARDS PERSONALIZED AI: ENGINEERING MODEL RESPONSES FOR CUSTOMIZED USER INTERACTIONS IN GENERATIVE AI SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in generative AI models, particularly the emergence of ChatGPT, have significantly increased interest in generative AI. However, the reliance on text prompts as the primary medium of interaction poses various challenges, especially for users with limited AI knowledge. Prompt engineering has emerged as a technique to address this issue, but it is still time-consuming and may not always yield satisfactory results. In this research, we aim to test the feasibility of an AI system that provides customized answers tailored to individual users, focusing on engineering model responses rather than solely relying on prompt engineering. We developed a system that enables users to evaluate their preferences and provide feedback on several model responses for the same prompt. And through this system, we collected user preference data and applied instruction tuning to guide model responses in a manner preferred by a certain user based on the certain user's preference data. Our study serves as a proof of concept to explore the potential of more personalized and user-centric AI systems in the future.

1 INTRODUCTION

In recent years, there has been a surge in technological advancements related to generative models. Particularly, with the emergence of ChatGPT, which is based on GPT and further trained to be more suitable for chatting with humans, the interest in generative AI has significantly increased (Ouyang et al. (2022)). The primary medium of interaction between these generative models and end users is through natural language prompts. The popularity of models like ChatGPT has grown because anyone can easily obtain high-quality answers from the model by writing just a short text prompt.

However, there are limitations to interacting with these models through text prompts alone. Especially for those with little prior knowledge about AI, it is often challenging to determine how to write an effective prompt. As text prompts are the sole medium for interacting with these generative models, the influence of prompts is huge (Zhou et al. (2022); Dang et al. (2022)). Consequently, users may spend considerable time rewriting their prompts, believing that their dissatisfaction with the model's response is due to the quality of the prompt they provided.

Prompt engineering has emerged as a new technique to address this issue. There are various guidelines or tools for prompt engineering for several models, from large language models (LLMs: Zhou et al. (2022); Reynolds & McDonell (2021); White et al. (2023)) to text-image generative models(LTGMs: Liu & Chilton (2022); Witteveen & Andrews (2022)). Even books and online courses for prompt engineering are gaining popularity (Mas). Our study is also motivated by the desire to minimize the difficulties users face in prompt engineering. The ultimate goal of users employing prompt engineering is to obtain satisfactory answers from generative models. However, the nature of responses from these models can vary greatly, and even the same prompt may yield different answers. Furthermore, generative models may provide incorrect information as fact (hallucination) or give inconsistent responses(Alkaissi & McFarlane (2023)). The impact of these characteristics will differ depending on the user, ranging from potentially hazardous to merely amusing situations.

In this research, our primary objective is to test the feasibility of an AI system that provides customized answers tailored to individual users, placing the focus on engineering model responses rather than solely relying on prompt engineering. At first, we developed a system that enables users to evaluate responses from a generative model such as ChatGPT[1], and provide feedback on the aspects they found favorable or unfavorable. Through this system, we collected the individual user's preferences and rationale data. Based on this data, we aim to transform them into instructions and apply instruction tuning (Wei et al. (2021)) on a pre-trained language model like LLaMa (Touvron et al. (2023)) to guide responses of generative models in a manner preferred by each individual user and explains the rationales.

However, we can raise a question regarding the amount of data needed for instruction. While collecting as much data as possible may seem beneficial, it can contradict the philosophy of this project, which aims to alleviate user's burden of prompt engineering. Therefore, we employ self-instruct (Wang et al. (2023)) to augment and generate instruction data based on individual users' preferences and corresponding reasons, using them as a seed dataset for the self-instruct. This approach can minimize the effort required to annotate user preferences and reasons for each user.

In this study, we aim to experiment with ChatGPT as the base model, utilizing the small-size (7B) open-source pre-trained language model, LLaMa, to train on each user preference data. This model is trained to guide the responses of ChatGPT in a more personalized direction by instruction tuning method. By doing so, we aim to test a proof of concept to determine the possibility of obtaining a personalized ChatGPT system without requiring users to perform prompt engineering, but instead to annotate their preferences and reasons among several ChatGPT responses to the same prompt. If the user would only go through this annotating process a few times, the model's response can be engineered to better align with the user based on the preference and reasons data collected from this process. By exploring the potential of this approach, we hope to pave the way for more personalized and user-centric AI systems in the future.

2 Related Work

Natural Language Processing (NLP) (Hirschberg & Manning (2015)) has been developed by the self-attention mechanism, transformer (Vaswani et al. (2017)) and task-specific fine-tuning techniques. However, conventional language models show limitations in processing out-of-distribution and are prone to generate spurious correlations when applying task-specific find-tuning on a large amount of data. Therefore, in Large Language Models (LLMs), for example, GPT-3 (Brown et al. (2020)) applied meta-learning for few shot demonstrations through text interactions without any gradient updates or fine-tuning. To implement meta-learning (Finn et al. (2017)) for few shot demonstrations, they increased the model size significantly as recent trends in transformer-based language models to improve in-context learning which can be classified as one-shot, zero-shot, and few-shot depending on the number of the context window. Under these few shot settings on several NLP tasks, it shows more state-of-the-art than any fine-tuned models and also shows proficiency in several tasks such as test rapid adaptation or on-the-fly reasoning. And they present a recent version, LLMs multi-modal model GPT-4 that is highly improved in overall context learning and can process demonstrations with images. According to the technical report of GPT-4, it shows human-level performance in various professional and academic areas, and pass on a bar exam with the top 10 percent results.

And (Touvron et al. (2023)) has achieved better performance considering compute budget constraints. LLaMa (13B) outperforms GPT-3 (175B) in most benchmarks with 10 times smaller size and it is competitive with Chinchilla (70B) and PaLM (540B) that show remarkable efficiency in compute budget. It is trained on a large amount of tokens of public datasets which includes English CommonCrawl, C4, Github, Wikipedia, etc. The architecture of the transformer is updated by recent techniques such as pre-normalization, SwiGLU activation function, and rotary embeddings. To improve training speed, it uses an efficient multi-head attention that significantly reduces memory and computational cost. And they also applied Activation Checkpointing to save expensive activation in linear layer to reduce training speed.

Low-Rank Adpation (LoRA) Hu et al. (2021) was proposed to train model paramter efficiently, because existing training method for large language model required high computational cost. There has been many trials to apply large language model to proper task through finetuning all the model's paramter. However, only a few part of model's paramters are actually used in finetuning which

[1] 1https://chat.openai.com/

Response	Role	Temperature
#1	System	1
#2	System	0.1
#3	Assistant	1
#4	Assistant	0.1

Table 1: The combinations for each of four responses

means that regular finetuning is inefficient in terms of computational cost and time. In other approaches, increasing the model's depth can occur inference latency or requires more reduction of input length that model can handle. Instead of increasing model's depth, LoRA optimizes the rank decompostion matrices of layer changes during adaptation which indirectly learn the layer. Through this technique, pretrained language model can be shared and applied to various tasks.

Prompt Engineering has focused to improve the performance of Large Language models (LLMs) by communicating task intention. In Reynolds & McDonell (2021), the authors demonstrate that zero-shot prompts can outperform few-shot prompts when using task-specific prompts and address that prompts in controlling and evaluating language models is essential to understand the better capabilities of LLMs. In Zhou et al. (2022), they propose automatic prompt generation that addresses language model synthesis as black-box optimization problems using LLMs. This work automatically generates instruction about demonstrations depending on their quality. It also present the result of zero-shot prompts which can outperform human-generated instructions.

3 Method

The primary objective of this study is to explore the feasibility of an AI system that involves a generative model in offering personalized responses tailored to individual users. To accomplish this, we developed a system that enables users to annotate their preferences among the multiple responses generated by ChatGPT in response to the same prompt. The system also collected subjective responses regarding the reasons behind their evaluations. The system is available at this https URL.

Subsequently, we utilized the user preferences and corresponding reasons data collected from this system to train a pre-trained language model, LLaMA, using the instruction tuning method. The model was trained with instructions that guided the responses of ChatGPT to align with the individual preferences of each user.

3.1 User Preference Data Collecting System

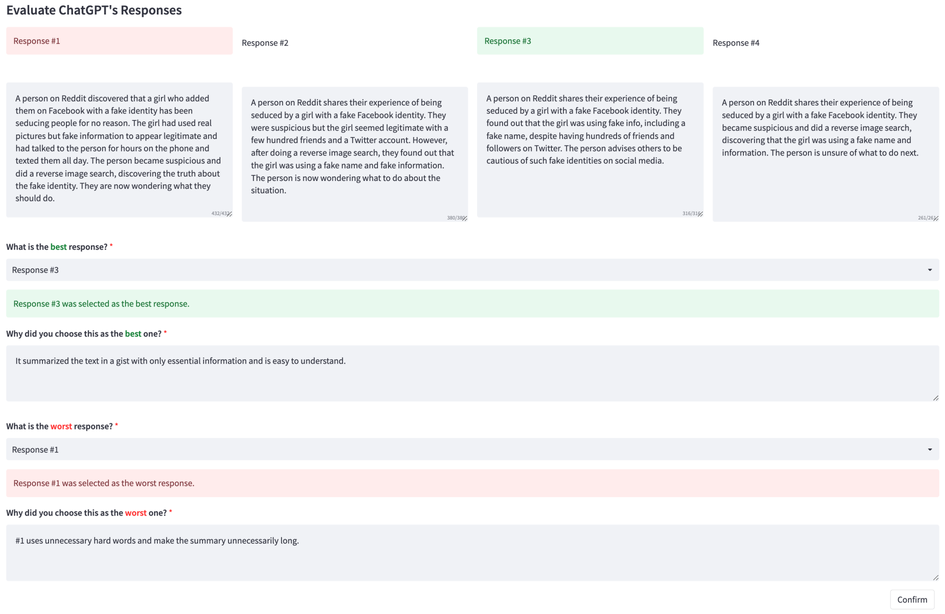
First, we developed a system to collect user preference and their corresponding reasons for multiple responses generated by ChatGPT in response to the same prompt. For this purpose, we utilized the Chat Completions API[2] provided by OpenAI. We used the gpt-3.5-turbo model, and we simultaneously obtained four different answers for the same prompt and presented them to the user. To obtain four different answers from the ChatGPT API for the same prompt, we varied the parameters of role[3] and temperature[4]. The combinations for each of the four can be found in Table 1. Through this, the system receives a prompt from the user, presents the user with four different responses generated by ChatGPT simultaneously, and asks the user to select the best and worst responses from their perspective. The user is also requested to provide high-level feedback or detailed rationales for their choices. The screenshot of the system for this process can be seen in Figure 1. Similar to a typical ChatGPT system, this system also supports multi-turn conversations. However, in this system, out of the four generated responses, only the user-selected best response is stored in the chat history. It can be seen in Figure 2 This system is connected to a MongoDB[5], where all user data is stored for later language model tuning, taking into account user preferences.

[2] https://platform.openai.com/docs/guides/gpt/chat-completions-api

[3] The role is for the author of the message. It can be one of a system or assistant.

[4] Temperature is a parameter for controlling the randomness of the generated text. Higher values will make the output more random, while lower values will make it more deterministic.

[5] https://www.mongodb.com/



Evaluate ChatGPT's Responses

Response #1

A person on Reddit discovered that a girl who added them on Facebook with a fake identity has been seducing people for no reason. The girl had used real pictures but fake information to appear legitimate and had talked to the person for hours on the phone and texted them all day. The person became suspicious and did a reverse image search, discovering the truth about the fake identity. They are now wondering what they should do.

Response #2

A person on Reddit shares their experience of being seduced by a girl with a fake Facebook identity. They were suspicious but the girl seemed legitimate with a few hundred friends and a Twitter account. However, after doing a reverse image search, they found out that the girl was using a fake name and fake information. The person is now wondering what to do about the situation.

Response #3

A person on Reddit shares their experience of being seduced by a girl with a fake Facebook identity. They found out that the girl was using fake info, including a fake name, despite having hundreds of friends and followers on Twitter. The person advises others to be cautious of such fake identities on social media.

Response #4

A person on Reddit shares their experience of being seduced by a girl with a fake Facebook identity. They became suspicious and did a reverse image search, discovering that the girl was using a fake name and information. The person is unsure of what to do next.

What is the best response? *

Response #3

Response #3 was selected as the best response.

Why did you choose this as the best one? *

It summarized the text in a gist with only essential information and is easy to understand.

What is the worst response? *

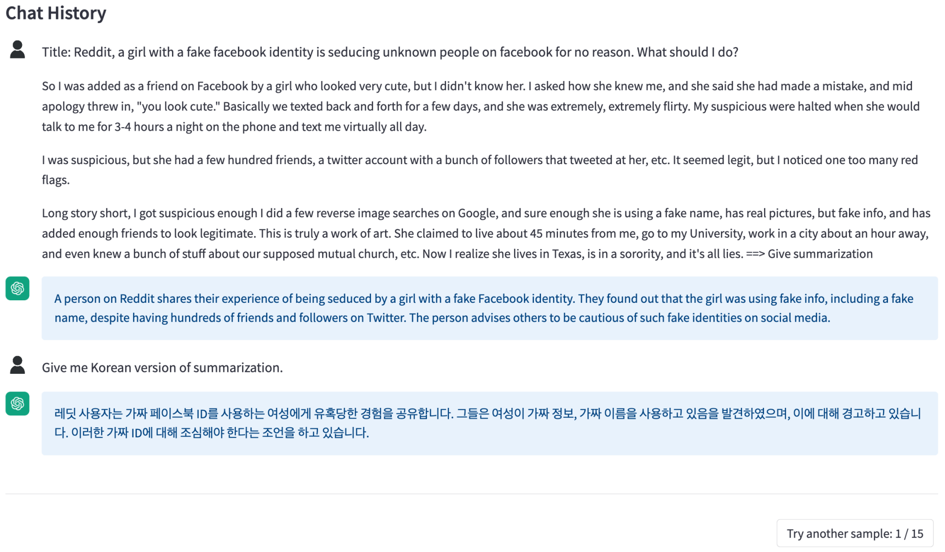
Response #1

Response #1 was selected as the worst response.

Why did you choose this as the worst one? *

#1 uses unnecessary hard words and make the summary unnecessarily long.

Figure 1: The system receives a prompt from the user and generates a total of four ChatGPT responses simultaneously. The system then requests the user to evaluate which response is the best and which is the worst, along with providing a written explanation for their choices. The response chosen as the best by the user is highlighted with a green box, while the worst response is indicated with a red box.



Chat History

Title: Reddit, a girl with a fake facebook identity is seducing unknown people on facebook for no reason. What should I do?

So I was added as a friend on Facebook by a girl who looked very cute, but I didn't know her. I asked how she knew me, and she said she had made a mistake, and mid apology threw in, "you look cute." Basically we texted back and forth for a few days, and she was extremely, extremely flirty. My suspicious were halted when she would talk to me for 3-4 hours a night on the phone and text me virtually all day.

I was suspicious, but she had a few hundred friends, a twitter account with a bunch of followers that tweeted at her, etc. It seemed legit, but I noticed one too many red flags.

Long story short, I got suspicious enough I did a few reverse image searches on Google, and sure enough she is using a fake name, has real pictures, but fake info, and has added enough friends to look legitimate. This is truly a work of art. She claimed to live about 45 minutes from me, go to my University, work in a city about an hour away, and even knew a bunch of stuff about our supposed mutual church, etc. Now I realize she lives in Texas, is in a sorority, and it's all lies. ==> Give summarization

A person on Reddit shares their experience of being seduced by a girl with a fake Facebook identity. They found out that the girl was using fake info, including a fake name, despite having hundreds of friends and followers on Twitter. The person advises others to be cautious of such fake identities on social media.

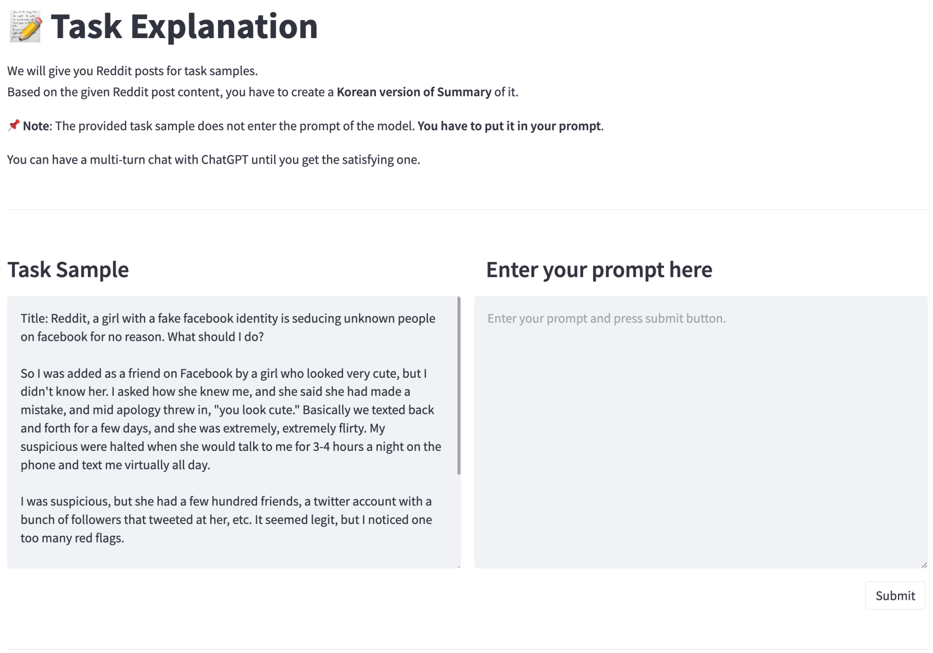
Give me Korean version of summarization.

레딧 사용자는 가짜 페이스북 ID를 사용하는 여성에게 유혹당한 경험을 공유합니다. 그들은 여성이 가짜 정보, 가짜 이름을 사용하고 있음을 발견하였으며, 이에 대해 경고하고 있습니다. 이러한 가짜 ID에 대해 조심해야 한다는 조언을 하고 있습니다.

Figure 2: The screenshot showcases the chat history interface. Among the four generated responses, only the user-selected best response is retained in the chat history. Users can engage in multi-turn conversations, similar to a typical ChatGPT system.

3.2 Data Collection and Task

We could have collected user preference data from 5 participants. Each participant evaluated the four generated responses from ChatGPT based on their perspective. And they were asked to assess the best and worst answers and provide written explanations for their choices. We provided participants with a controlled environment by assigning them a specific task. This was done because, when



📝 **Task Explanation**

We will give you Reddit posts for task samples.
Based on the given Reddit post content, you have to create a **Korean version of Summary** of it.

📌 **Note**: The provided task sample does not enter the prompt of the model. **You have to put it in your prompt**.

You can have a multi-turn chat with ChatGPT until you get the satisfying one.

Task Sample

Title: Reddit, a girl with a fake facebook identity is seducing unknown people on facebook for no reason. What should I do?

So I was added as a friend on Facebook by a girl who looked very cute, but I didn't know her. I asked how she knew me, and she said she had made a mistake, and mid apology threw in, "you look cute." Basically we texted back and forth for a few days, and she was extremely, extremely flirty. My suspicious were halted when she would talk to me for 3-4 hours a night on the phone and text me virtually all day.

I was suspicious, but she had a few hundred friends, a twitter account with a bunch of followers that tweeted at her, etc. It seemed legit, but I noticed one too many red flags.

Enter your prompt here

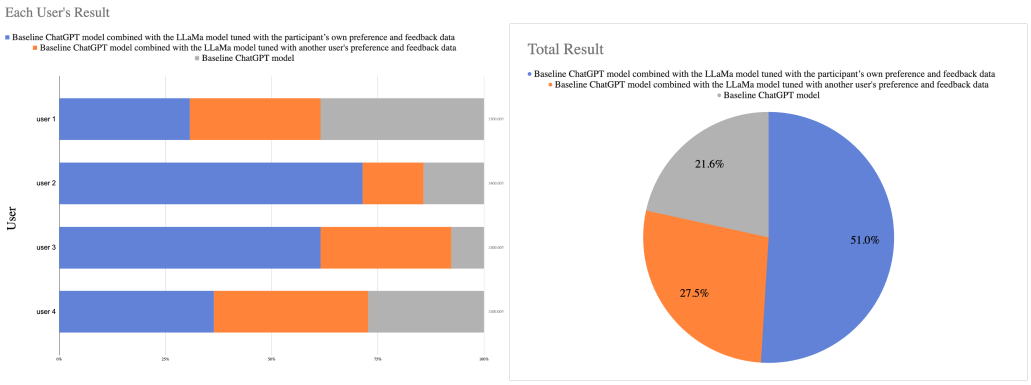
Figure 3: Task explanation and system screenshot. The left panel is a Reddit post given for the task. The right panel is an input box to take the user's prompt.

given participants freedom of use, they often struggled to find a meaningful task to assess their preferences. Additionally, for the evaluation of this concept, a certain level of controlled environment was necessary. As a result, participants were assigned a specific task. The task is to write a Korean summary of the given Reddit[6] post by chatting with the ChatGPT system. There were several reasons for selecting this task. Firstly, the choice of utilizing Reddit posts was driven by the fact that they offer a greater diversity in terms of content and format compared to news articles. Thus, it was anticipated that a variety of summaries could be generated from such posts. Additionally, in order to encourage user participation, Reddit posts were preferred as they cover content that is closer to everyday life, rather than news articles. The decision to opt for a summarization task was based on the understanding that individuals may have different preferences regarding the format and style of summaries. Lastly, the inclusion of a Korean translation task was motivated by the belief that individuals would also have diverse preferences for the format and style of translated versions. The task explanation and given Reddit post sample can be seen in Figure 3. The Reddit post samples are brought from Stiennon et al. (2022).

3.3 Instruction Tuning on Language Model for User Preference

Building on the user preference data, we preprocessed the data in the form of instructions, such as inputs and outputs. When we collected data, the Reddit post and the user's prompt were processed together, but in the training process, we separated them into two parts and combined the best response and its reason into a single output. For the training model, we proceeded to train on the pre-trained language model with Low-Rank Adaptation (LoRA) Hu et al. (2021) that enables to train model parameters efficiently by adding trainable rank decomposition matrices to perform downstream tasks instead of finetuning all parameters of a pre-trained language model. We referred Stanford's Alpaca LoRA for training our model. Although LoRA enables to train of the language model in a short time, a certain amount of data is still necessary to perform desired tasks for user preference. While we collected about 15 pieces of data from individual users, and this is not enough for training the model. Therefore, we applied the self-instruct technique which can generate data by following the human data through another open-source pre-trained language model, GPT-3 model,

[6] https://www.reddit.com/



(a) Each user's evaluation result

(b) Total evaluation result

Figure 4

from the OpenAI API [7]. When applying the self-instruct technique, generated data need to be filtered to prevent overfitting, and then train the LoRA model that reflects the preference of individual users from filtered data. We conducted this process using data collected from five participating individuals, and through this, we were able to obtain five unique models that guide personalized responses to each user.

4 Evaluation

4.1 Evaluation Plan

The evaluation plan for this research aimed to validate the effectiveness of the language models tailored to each user based on their individual preferences and feedback data. With a total of five participants contributing their preference data, we obtained five unique language models. We reached out to the five participants who had participated in the user preference data collection process and requested their involvement in the model evaluation.

The evaluation system was similar to the user preference data collection system. The system presents three simultaneous responses for the same prompt and each participant was asked to evaluate which model's response was the best from their perspective. The four models generating these responses were as follows: (1) Baseline ChatGPT model combined with the LLaMa model tuned with the participant's own preference and feedback data, (2) Baseline ChatGPT model combined with the LLaMa model tuned with another user's preference and feedback data, and (3) Baseline ChatGPT model. The order of the response models was randomized for each evaluation. Additionally, for (2), another user's model was randomly selected for each evaluation.

By comparing (1) Baseline ChatGPT model combined with the LLaMa model tuned with the participant's own preference and feedback data and (3) the Baseline ChatGPT model, we aim to verify the effectiveness of generating personalized responses based on the user's preference data. Furthermore, by comparing (1) Baseline ChatGPT model combined with the LLaMa model tuned with the participant's own preference and feedback data and (2) Baseline ChatGPT model combined with the LLaMa model tuned with another user's preference and feedback data, we aim to demonstrate the effectiveness of models tuned with data personalized to the specific user rather than the effectiveness of LLaMa model tuned by human preference and feedback data.

4.2 Result

Our evaluation was conducted on a total of 4 participants out of the 5 people who participated in the data collection. The task for evaluation, akin to that within the data collection part, was to write a Korean version of the summary for a given Reddit post sample. The participants were

[7] https://platform.openai.com/docs/models/gpt-3

asked to select the best responses from three responses provided simultaneously for the same user prompt. Each participant's evaluation results can be seen in Figure 4a. Although the evaluation results varied slightly for each person, it was found that in all cases, participants either preferred the model customized for themselves or favored it to a similar degree to other models. The aggregated results of each person's evaluation can be seen in Figure 4b. From this, we can see the customized models received the highest evaluations, followed by models tuned for other users, and lastly, the baseline model received the least favorable evaluations. Through this, we were able to confirm that by instruction tuning the language model based on each user's preferences and feedback data, the responses of ChatGPT could be guided in the direction that the user prefers more.

5 Discussion

5.1 Limitation

We acknowledge several limitations that need to be addressed for future improvements. Firstly, it has a limitation in terms of the insufficient number of participants involved in data collection and evaluation. As a result, the results presented in this study may not be enough to serve as proof of the concept of a personalized ChatGPT system. Nevertheless, the user preference data collection system has been developed with scalability in that it can be easily accessed via a website link. Therefore, it will be possible to recruit a large number of participants to conduct further data collection and evaluations.

Secondly, the study relied on a specific task of generating a Korean summary from a given Reddit post to control the experimental environment. While this task allowed for consistent evaluation, it does have limitations in terms of task diversity. To obtain a more comprehensive understanding of user preferences and to explore the system's adaptability across different tasks and domains, future research should incorporate a broader range of tasks that encompass various content types.

Furthermore, generating four different responses from ChatGPT simultaneously for the same prompt posed a practical challenge. It required a significant amount of time for response generation, leading to potential usability concerns. To improve the scalability of the system, it is necessary to optimize the delay in generating multiple responses from ChatGPT. Reducing the time required for response generation would enhance the usability and efficiency of the system, allowing for a smoother user experience and facilitating larger-scale data collection.

This study has limitations posed by the limited number of participants, task diversity, and long response generation time. However, the results obtained from this initial investigation can serve as a starting point for further research and highlight the potential for creating more personalized and user-centric AI systems with a broader participant base.

5.2 Future Work

In our research, we conducted experiments specifically focusing on the concept within the context of large language generative models, particularly ChatGPT. However, it is important to note that this concept has the potential to be applied to a wide range of other language models, and its applicability extends beyond language models alone. For instance, it can be extended to large-scale text-to-image generation models such as DALL-E Ramesh et al. (2021; 2022).

To illustrate this potential extension, consider a scenario where a user provides a text prompt, and the system generates multiple images based on that prompt. The user is then asked to evaluate and provide feedback on the generated images, selecting the best and worst images and providing reasons for their choices. This collected data can then be leveraged to train a smaller model that guides the output of the original text-to-image generation model in a more personalized and user-aligned manner. This kind of smaller model which incorporates user preference and feedback can serve as a guiding framework to enhance the overall generation process and deliver more tailored and satisfactory results.

The potential of this concept is not limited to the realm of language and image generation models. It can be further explored and adapted to other domains where generative models are utilized, providing a valuable framework for user-guided generation and personalized model refinement.

6 Conclusion

In this research, we aimed to test the feasibility of developing an AI system that provides customized answers tailored to individual users, focusing on engineering model responses rather than solely relying on prompt engineering techniques. Through a system that allowed users to evaluate and provide feedback on generative model responses, we collected individual user preferences and the corresponding rationale data. Using this data, we applied instruction tuning on a pre-trained language model, LLaMa, to guide the responses of generative models in a manner preferred by each user and provide explanations for those responses. The evaluation involved comparing the performance of customized models, including LLaMa combined with user preference data, LLaMa combined with another user's preference data, and baseline ChatGPT models. Although we could have not done enough evaluation, results showed the efficacy of incorporating user preference data to guide generative model responses, showcasing the potential of personalized and user-centric AI systems. Future research can focus on expanding the user base, evaluating models on diverse tasks, and incorporating real-time user feedback to further enhance the effectiveness and usability of personalized AI systems.

References

Master ai with prompt engineering — udemy. https://www.udemy.com/course/promptengineering/?utm_source=adwords&utm_medium=udemyads&utm_campaign=LongTail_la.EN_cc.ROW&utm_content=deal4584&utm_term=_._ag_119586862319_._ad_535397279676_._kw__._de_c_._dm__._pl__._ti_dsa-1212271230479_._li_1009880_._pd__._&matchtype=&gclid=CjwKCAjwrJ-hBhB7EiwAuyBVXVWua1Y3DJiXDwlt-hcBhgV3sSaFz6VI_etSP0uu5rU880xH-Mq48RoCWo8QAvD_BwE. (Accessed on 04/01/2023).

Hussam Alkaissi and Samy I McFarlane. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2), 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. How to prompt? opportunities and challenges of zero-and few-shot learning for human-ai interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*, 2022.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–23, 2022.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2021.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2023.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

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

Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models. *arXiv preprint arXiv:2211.15462*, 2022.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.