NPC: Personalized Next Profile Crafting Using Previous Role-Play Bots

Anonymous ACL submission

Abstract

Personalization and role-playing are two important research topics for the LLM community. However, the exploration in the direction of personalized role-playing especially rare. The primary obstacle in personalized role-playing is the absence of a dataset containing role-playing dialogues with personalized information. To overcome this obstacle, we introduce a new large-scale personalized role-playing dataset Multi-Bot Tailored Interaction Dataset (MBTI), which includes the entire interaction history from creating bot to deeply engaged conversation between 1238 users and 8477 Bots. More importantly, we propose a new pipeline called "Next Profile Crafting (NPC)" for crafting role profiles with cross-bot insights to achieve personalization before the conversation. This method is based on the bot persona link among historical bots that user has multi-turn interaction with. We conducted tests using both trained and untrained approaches, as well as open-source and proprietary large language models, highlighting significant disparities in the effectiveness of personalized crafting in the NPC task. Our findings indicate substantial room for improvement in current methodologies.

1 Introduction

Role-playing systems, where users interact with bots embodying different characters, have become popular recently (Salemi et al., 2023). Current role-playing methods and datasets typically rely on established characters rather than user-driven characters (Shao et al., 2023; Wang et al., 2023b,a; Gosling et al., 2023; Zhou et al., 2023).

In fact, a real-world user interacts with multiple bots according to his needs. When users create a new bot, they have initial expectations, such as the bot type and a brief introduction. Crafting personalized profile for the new bot based on a user's past interactions is critical to align with user preferences in future dialogues, see Fig1.

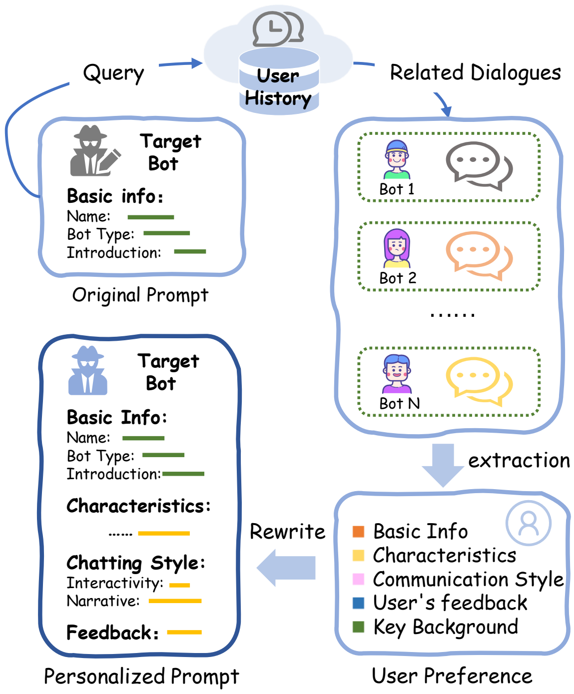

Figure 1: The figure shows the core idea of personalized Next Profile Crafting using interaction history of previous bots chosen by the same user.

However, the exploration in the direction of personalized role-playing is especially rare. The absence of a personalized role-playing dataset hinders the research in this direction. To overcome this, we have assembled Multi-Bot Tailored Interaction Dataset (MBTI), the first large-scale personalized role-playing dataset. It contains the entire interaction history, from creating bots to deeply engaged conversations between 1238 users and 8477 Bots. It features a one-user-to-many-bots interaction. Our analysis shows that the previous bots within one user share some special persona links that can be passed to the coming new bot.

Using MBTI, we established the first role-play personalization task which is called Next Profile Crafting (NPC), see Fig2. The NPC task means using the historical information in previous bots to

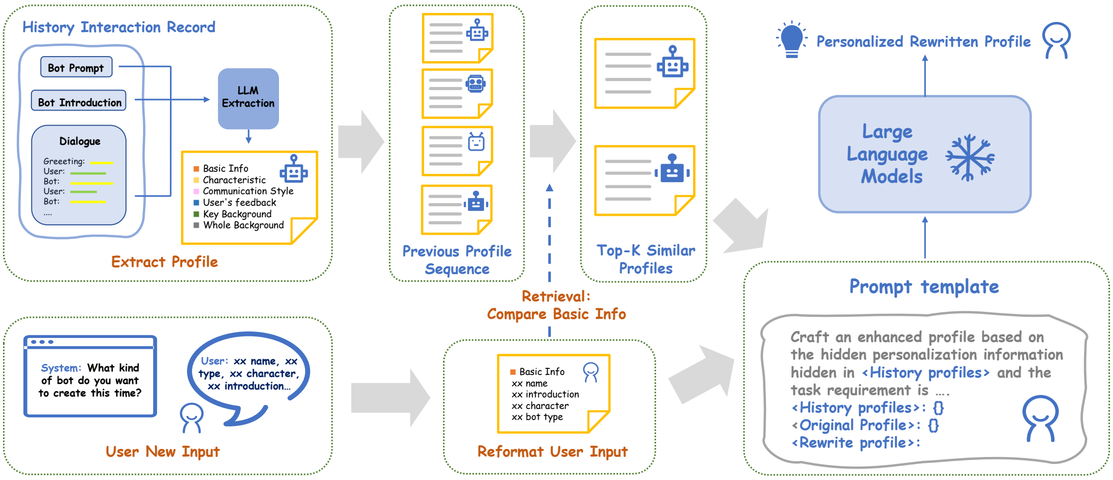

Figure 2: The pipeline for personalized next profile crafting(NPC) which includes user new input reformatting, similar profiles retrieval and personalized profile crating. The profile extraction is finished at data processing stage.

enhance the next bot user created at the beginning. The primary challenge in the NPC tasks lies in the model's ability to accurately identify and utilize appropriate preference information from historical data while ensuring that the current character's information is preserved.

We conducted a detailed comparison between trained and untrained methods on NPC task and have three main findings regarding the current state of role-play personalization:

1. Proper training can lead to significant performance improvements, yet there still exists a trade-off between among metrics.
2. Among all the models tested, GPT-4 in a zero-shot setting shows the best performance.
3. There remains a huge gap between the current performance levels and the theoretical upper limit thus leaving huge room for improvement.

2 Related work

2.1 LLM-based Role Playing

Role-playing is a basic ability for large language models (Shanahan et al., 2023; Song et al., 2023). It involves a language model adopting a specific characters to engage user in immersive and scenario-driven dialogues. The characters can be a well-established characters from web or original characters created by users (Zhou et al., 2023). And the character profile mainly includes description and behaviour. Many previous work focus on extracting better profile for each character through different methods thus leading to various kinds of role-play datasets. Their construction methods can be roughly categorized into three folds. **Web Extraction:** For those established characters from celebrities, historical figures or fictional characters, Li et al. (2023) and Wang et al. (2023b) construct profile based on the information like dialogue in sourced script. **Synthetic Data:** Given that most advanced language models are trained on massive text, several research (Shao et al., 2023; Wang et al., 2023b,a; Gosling et al., 2023) collect conversation topic from general task instruction, literature, personality test or real-use case and then prompt LLMs with in-context examples to synthetic data as augment. **Human Simulation:** Zhou et al. (2023) recruit many human annotators to simulate different role and pair them for conversational interactions.

The aforementioned methods are limited to established characters, and the quality of synthetic data cannot be effectively guaranteed (Tu et al., 2024). Additionally, the datasets do not support user personalization. Several studies (Lu et al., 2024; Tao et al., 2023) mentioned personalization but still focus on construction of role's profile rather than user's persona. Although Gosling et al. (2023) originates from real user data on the Character.ai[1] platform, it consists only of dialogue data and lacks character profiles. Moreover, it does not distinguish between users. Based on our knowledge, MBTI dataset is the first role-playing dataset that supports high-level personalization under real-use case.

[1]https://character.ai/

2.2 LLM-powered Personalization

Recently, large language models has shown strong ability in general task-solving and thus are popular in research relevant to personalization (Chen et al., 2024) which is difficult in various applications including typical dialogue system (Li et al., 2016; Zhang et al., 2018), domain specific dialogue system like health agent, creative image generation (Chen et al., 2023) and recommendation system (Friedman et al., 2023). The challenges in dialogue systems primarily arise from the need to process excessively long inputs, as models must extract key information from extensive historical dialogues. Salemi et al. (2023) and Xu et al. (2022) utilizes dialogue sessions as the unit for information transfer. Additionally, interactions often suffer from low data information density and significant noise and Zhou et al. (2024) implements a memory mechanism for this problem. Furthermore, personalization in specific domains like recommendation requires huge domain knowledge like Zhang et al. (2023), which presents a significant challenge for general-purpose large models. Wang and Lim (2023); Lyu et al. (2023) explore the way to adpot LLMs in sequence recommendation tasks in a zero-shot way and Bao et al. (2023) fine-tune LLMs to unlock the their abilities for recommendation tasks.

Within these applications, sequence prediction is the most mature application. While next-item prediction is a straightforward task that ranks items based on user preferences from historical sequences, our task involves a more complex challenge: smartly incorporating accurate user preferences into the current bot's profile by filtering and integrating historical information.

3 Multi-Bot Tailored Interaction Dataset (MBTI)

Actually, in real role playing scenario, user will talk to many bots with different roles and there is hidden persona links among these bots. Based on this feature, we establish the first large-scale role-playing dataset under real user scenarios in a "one user to multi-bots" style, named the Multi-Bot Tailored Interaction Dataset (MBTI).

3.1 Data collection

Collection Based on Real-Life Role-Play Platform The data are collected from a popular role-play playground. In this playground, user can create many different role-play bots by passing some information based on the system guidance. The information user has to input includes bot name, short introduction, avatar, character description and greeting. Other optional information includes gender, age, bot categories, scenario, facts, knowledge and example dialogue. The detailed guidance can be found in the page of creating bots which is shown in Appendix A.2.

Based on user's design, the system will create a role-play bot for the user to chat with. In Appendix A.2 we present an example of a user interacting with the classic anime character "Levi Ackerman." During the dialogue, the user provides feedback based on the bot's performance. If the bot does not meet expectations, the user may terminate the current conversation prematurely and design a new bot.

All in all, for most users, they will create many different bots in different time and will tend to chat more with bot that meet their needs.

Scalable Engaged User-Centric Bot Sequence One user may create many bots but we only want to focus on those bots with ample user interactions and imprints. Besides, we would like to see how persona links within these bots support persona transfer to new bots. So we definite an scalable engaged user-centric bot sequence, which is a subsequence of the whole history bot sequence of certain users who have sufficient effective bots. Based on our several assumptions, the sequence and the bots in this sequence should have three characteristics: Non-Nsfw, no less than 10 dialogue turns and no less than 4 effective bots meeting the first two criteria.

Therefore, the criteria for data filtering correspond to the three requirements mentioned above. The first involves the combination of keyword searches and the use of OpenAI's moderations endpoint[2] to eliminate data containing NSFW content. The second criterion presupposes that a user's engagement with a bot for ten or more rounds signifies a preference for that particular bot, thereby excluding "user-bot" pairs with fewer than ten interactions. The third criterion recognizes the personalized aspect of the study subject, suggesting that users who have interacted with four or more bots are more likely to demonstrate stable preferences, and selects such users for inclusion in the study.

In the end, the dataset contains 1238 users and their interaction history with different bots. There

[2]https://platform.openai.com/docs/guides/moderation

Figure 3: Example of user preference

are totally 8477 "user-bot" pairs and we select one bot as test from each user. Therefore, there is 1238 test bots and the remaining 7239 bots are accessible for retrieval. Note that a user has at least 3 bots to support crafting the next bot.

Role Profile Design For a privious bot that has engaged in no fewer than ten complete dialogue rounds with users, to fully explore the bot's characteristics by integrating the initial description provided during its creation and the traits demonstrated in conversations, we utilize GPT-4 to extract the bot's profile for all bots in our MBTI dataset. The detailed formulation of creating profile p_i for the i-th bot is shown below.

Given a user u's historical interaction sequence $\{(x_1, d_1), (x_2, d_2), \ldots, (x_{\mathrm{N}}, d_{\mathrm{N}})\}$ where N, x_i and d_i represent bots number in sequence, user initial input for creating i-th bot and the dialogue happens between user and i-th bot respectively, we can extract profile $p_i = \mathbf{E}(x_i, d_i)$ for each i from 1 to N.

Based on our profile extractor $\mathbf{E}(\cdot)$, each resulting profile p_i can be split into L dimensions $\{o_1, \ldots, o_L\}$, where $L = 6$. The six dimensions include: (i) bot's basic information, (ii) characteristics, (iii) style of replay, (iv) user feedback, (v) key background and (vi) whole background respectively. The format for profile is similar to the example in Fig3. The detailed example for profile, the prompt for extracting profile and the introduction to the dimensions are shown in Appendix A.4.

3.2 Data Analysis

Basic Statistics As discussed in the Section 3.1, observing users' preferences will be influenced by the number of bots they have interacted with and the number of the dialogue rounds. Fig. 4 depicts the distribution of these two metrics within our

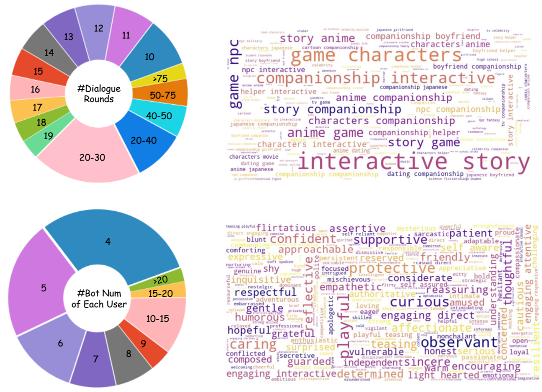

Figure 4: The basic data statistics for MBTI including the dialogues length distribution, bot number distribution, word cloud for bot type and word cloud for bot's characteristic and communication style

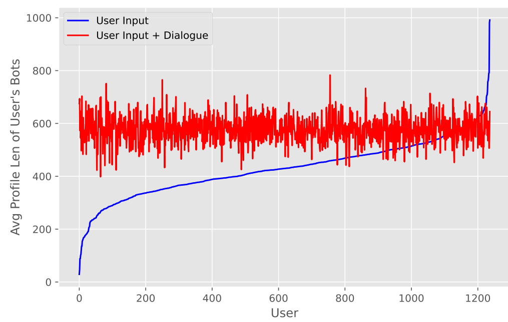

Figure 5: Impact of including dialogue on profile length.

dataset. The distribution of the number of bots exhibits a long-tailed pattern, while the distribution of dialogue rounds is more uniform. In fact, the average number of bots is 6.85 and the average dialogue rounds is 24.96, indicating that users tend to engage in longer conversations with fewer bots. This behavior suggests a deep level of engagement and commitment to bot-based interactions among users.

In Section 3.1, we generate the profiles for all bots in MBTI dataset and here we generate word cloud of bot type, characteristics and communication style in fig. 4. The results indicate that the dataset includes a wide variety of bot types and bots with diverse personalities. The bot type are various ranging from anime, dating, game characters, movie star VTuber, Companionship, etc.

Dialogue Can Enrich Character Personality For certain test bot with initial user input x and dialogue history d, we use the same profile extractor $\mathbf{E}(\cdot)$ to generate standard p and $p^{\emptyset}$ where $p = \mathbf{E}(x, d)$ and $p^{\emptyset} = \mathbf{E}(x, \emptyset)$. Here $\emptyset$ means there is no dialogue yet.

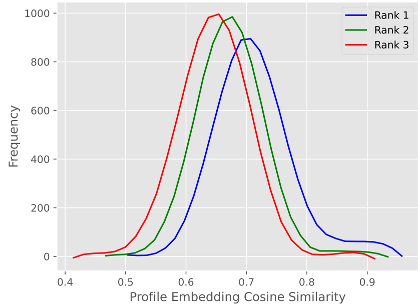

Figure 6: Histogram of similarity between profile and its neighbor

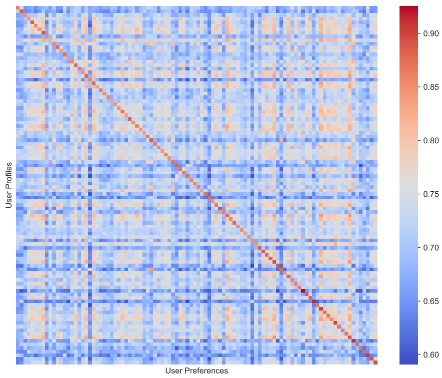

Figure 7: The intra-class and inter-class similarity of profile embeddings for 100 users.

We compare the length of $p^{\emptyset}$ and p to roughly estimate the information enhancement brought by dialogue d. In Fig. 5, p is longer than $p^{\emptyset}$ for most bots which means that p contains significantly more information, highlighting the need to enrich the initial profile and emphasizing the importance of this Next Profile Crafting task.

Cross-bot Persona Link In the MBTI dataset, we observe a strong cross-bot persona link among a user's different bots, providing a solid theoretical foundation for our Next Profile Crafting pipeline. As illustrated in Fig. 6, we calculated the similarity between each profile and its three nearest neighboring profiles. It was observed that the distribution shapes of similarity across different ranks are broadly similar. This suggests that, within the dataset, the majority of bots can identify one or two profiles that are most similar to them in the same user's historical bot interactions.

User Preference For each user, we summarize a preference from all the user's ground truth bot profiles. Specifically, for user u, as prompts in A.4, we feed LLM with all profiles $\{p_1, p_2, \ldots, p_{\mathrm{N}}\}$ and ask LLM to summarize an overall preference P_u, see Fig3. We ask LLM not to output basic info and whole background of the user preference profile cause even though different bots shares a persona link, their background story and basic info such as name and gender keep a slightly difference to maintain different role cores.

To examine if the bots in the same user's historical bot sequence can demonstrate consistency in certain fields, we obtain the embedding of all profiles, calculate the average bot profile embedding for each user to represent user preference and calculate the intra(inter)-user similarity. As the results shown in Fig. 7, the diagonal represents inter-user similarity, while other positions indicate intra-user similarity. It is evident that inter-user similarity is significantly higher than intra-user similarity.

4 Next Profile Crafting (NPC)

Task Definition The NPC task target to return an enhanced profile p^{*}_{N} for user u's query x_{N} based on historical profile sequence $Q_u = \{p_1, p_2, \ldots, p_{\mathrm{N}-1}\}$. Note that we will reformat the initial user input x_{N} to initial profile $p^{\emptyset}_{\mathrm{N}}$ where $p^{\emptyset}_{\mathrm{N}} = \mathbf{E}(x_{\mathrm{N}}, \emptyset)$.

Retrieval Process Given user u's whole historical profile sequence $\{p_1, \ldots, p_{\mathrm{N}-1}\}$ and initial profile $p^{\emptyset}_{\mathrm{N}}$, the retriever $\mathbf{Ret}\left(p^{\emptyset}_{\mathrm{N}}, \mathcal{Q}_u\right)$ retrieves top K relevant profiles $\mathcal{R} = \{r_1, ..., r_{\mathrm{K}}\}$ from historical profile set $\mathcal{Q}_u$ based on their embedding similarity with queries.

As shown in Fig.5, in most cases, an initial profile $p^{\emptyset}_{\mathrm{N}}$ is severely lacking information compared to its corresponding p_{N}. Therefore, we believe the main problem we need to tackle is retrieving the correct profile (i.e., those retrieved by p_{N}) with the initial, uninformed profile $p^{\emptyset}_{\mathrm{N}}$.

To select an effective retriever $\mathbf{Ret}(\cdot)$, we calculated NDCG and MRR scores between the profiles retrieved by p_{N} and those retrieved by our candidate retrievers (embedding-based dense retriever and BM25-based sparse retriever). The embedding-based retriever outperforms the BM25-based retriever by 0.24 in NDCG and 0.48 in MRR.

Ultimately, we choose embedding-based retrieval using the basic information from the profile as the query since a short introduction in this section is a more concise and direct indicator of the bots' domain and type.

Figure 8: The case study result for personalization rewriting and the illustrated text is summarized from profile

Rewriting Process In this stage, adopt chat version of large language model to make full use of the retrieved top K profiles $\{r_1, ..., r_K\}$ and return a final profile p^*_{N} which is an enhanced version of $p^{\emptyset}_{\mathrm{N}}$. That is, $p^*_N = \mathbf{Rew}\,(p_N, \mathcal{R})$, where $\mathbf{Rew}$ denotes a LLM-based rewriter.

During this stage, there are several sub-tasks for large language models. On the one head, LLMs needs to conduct general enhancement with expanding bot-specific information around current bot based on LLMs' knowledge. On the other head, LLMs need to assess the relationship between the reference profile and the current profile, as well as to identify commonalities exhibited among the referenced profiles. These commonalities represent the user's preferences and needs. Based on this, information that aligns with the user's preferences and is relevant to the current bot should be appropriately extracted. Ultimately, LLMs need to integrate general enhancement information with personalized extracted information in an appropriate manner to ensure compliance with a standard profile format, which includes considerations of structure and internal consistency, etc.

Semantic Based Evaluation We designed two evaluation matrices: User Preference Matching score (PM) and Bot Profile Fidelity score (PF). For PF score, we calculate the Recall and Precision score while for PM score we mainly care about the recall. Totally, these metrics correspond to the three important aspects. The PM-recall represents the converge of user preference, the PF-recall shows the bot-specific user preferences and the PF-precision evaluates the compliance of output format.

We calculate these metrics on a dimension-wise basis, meaning that we compute the results separately for different dimensions within the profile and then average them. We primarily focus on the scores for four dimensions: characteristics, style of reply, user feedback, and key background. Additionally, considering that ROUGE-L cannot recognize synonyms, and based on the logic we used when designing the profile, which involves content composed of phrases, we have employed phrase-level semantic matching for our calculations.

We define the formulation of calculated the **PMR** (PM-Recall), **PFR** (PF-Recall) and **PFP** (PF-Precision) for rewritten profile p^*_{N} as bellows:

$$\mathbf{PMR}(p^*_{\mathrm{N}}, P_u) = \mathbb{E}_{l\in\mathcal{L}}\left[\frac{\mathbf{SMN}(p^*_{\mathrm{N},l}P_{u,l})}{\mathbf{Len}(P_{u,l})}\right],$$

$$\mathbf{PFR}(p^*_{\mathrm{N}}, p_{\mathrm{N}}) = \mathbb{E}_{l\in\mathcal{L}}\left[\frac{\mathbf{SMN}(p^*_{\mathrm{N},l}, p_{\mathrm{N},l})}{\mathbf{Len}\,(p_{\mathrm{N},l})}\right],$$

Model	Rewrite Method	Train Method	Short Intial Profile			Medium Intial Profile			Long Initial Profile		
			PMR↑	PFR↑	PFP↑	PMR↑	PFR↑	PFP↑	PMR↑	PFR↑	PFP↑
GPT-4-preview-1106	GR	-	0.15	0.17	0.21	0.15	0.2	0.23	0.16	0.23	0.28
Mistral-7B-Instruct	PR	SFT	0.21	0.24	0.24	0.2	0.29	0.29	0.21	0.33	0.31
Mistral-7B-Instruct	PR	DPO	0.27	0.21	0.18	0.29	0.26	0.19	0.32	0.32	0.22
Mistral-7B-Instruct	PR	-	0.21	0.21	0.2	0.21	0.25	0.23	0.23	0.31	0.26
Llama3-8B-Chat	PR	-	0.24	0.22	0.21	0.26	0.28	0.23	0.28	0.34	0.24
GPT-35-16K	PR	-	0.23	0.24	0.18	0.24	0.29	0.21	0.26	0.35	0.24
Llama3-70B-Chat	PR	-	0.36	0.28	0.18	0.37	0.35	0.21	0.37	0.4	0.22
GPT-4-preview-1106	PR	-	0.31	0.25	0.22	0.33	0.3	0.25	0.29	0.32	0.27

Table 1: Performance comparison of various profile crafting methods applied to initial profiles of three different lengths(Short: <750 chars, 377 profiles; Medium: 750-1000 chars, 427 profiles; Long: >1000 chars, 434 profiles). Orange represents the highest performance, while blue indicates the second highest performance. PR represents personalized rewriting when $K = 3$, while GR stands for generalized rewriting.

$$\mathbf{PFP}(p_{\mathrm{N}}^{*}, p_{\mathrm{N}}) = \mathbb{E}_{l \in \mathcal{L}} \left[\frac{\mathbf{SMN}(p_{\mathrm{N},l}^{*}, p_{\mathrm{N},l})}{\mathbf{Len}(p_{\mathrm{N},l}^{*})} \right].$$

where $\mathbf{Len}(\cdot)$ refers to the number of phrases in the input description, and $\mathbf{SMN}(\cdot,\cdot)$ denotes the number of one-to-one semantic matches between phrases in the input two descriptions. Here, $p_{\mathrm{N},l}^{*}$ is defined as the personalized profile on a certain dimension l, and $p_{\mathrm{N},l}$ and $P_{u,l}$ are defined similarly. $\mathcal{L}$ represent the profile subsets including characteristics, style of reply, user feedback, and key background. Specifically, given two sets of phrases, we compute the pairwise embedding similarities to form a similarity matrix. Iteratively, we identify and record the highest values in the matrix that exceed a predefined threshold α as matching pairs. Once a pair is matched, the corresponding rows and columns are excluded from subsequent iterations. This process continues until no values exceeding the threshold remain in the matrix.

5 Experiments

5.1 Experiment Settings

Baselines General rewriting is designed to simply enriching the initial profile without using history profiles. Without enriching sub-process, the final profile will lose bot-specific information, consequently resulting in a lower PFP and PFR.

LLMs for Personalized Rewriting In our study, we conducted personalized rewriting based on the top-3 profiles retrieved via an embedding retrieval mechanism. The models employed were categorized into two types. The first type includes the trained Mistral-7B model, which underwent training utilizing Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) methods. The training dataset comprising 7,329 bots, excluding those designated as test bots. In the dataset, the chosen responses were standard profiles, while the rejected responses were derived from rewriting outcomes obtained through the general rewriting baseline method.

The second type comprises models tested in a zero-shot setting, which includes Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Llama3-8B-Chat, Llama3-70B-Chat, GPT-35-turbo-16k, and GPT-4-preview-1106.

Implementation Details The hyper-parameter of inference are the same across all models. The temperature is 0.95 and the top_p is 0.7. The max output tokens is 1000 for generating ground truth profile and 500 for generating rewriting result for efficiency.

5.2 Results

Personalize rewriting is effective across all lengths of initial profiles. In Table 1, as the length of the initial profile increases, personalized rewriting ensures that PMR and PFR are stably higher than baseline. This indicates that personalized rewriting can not only absorb the new information from history profiles but it can also maintain original information when the initial profile reaches a medium length or longer. The case study of the qualitative results is shown in Fig. 8.

The training is useful for NPC but still faces challenges. The trained Mistral model demonstrates notable enhancements; however, the improvements vary due to the differing objectives of training methods. SFT primarily targets consistency of standard profile, resulting in increased

Setting	Model	PMR↑	PFR↑
Initial Profile	-	0.14	0.23
History	-	0.62	0.31
GR	GPT-4	0.15	0.2
History+ GR	-	**0.67**	**0.41**
PR	GPT-4	0.31	0.29

Table 2: Analysis of the upper bound of recall. GR and PR represents generalized and personalized rewriting.

	#	PMR↑	PFR↑	PFP ↑
Retrieval Top-K	1	0.3	0.29	0.24
	3	0.31	0.29	0.25
	5	0.27	0.29	0.25
	7	0.25	0.28	0.25
Total History Bots	4-5	0.33	0.29	0.25
	5-10	0.31	0.3	0.24
	10+	0.27	0.3	0.26

Table 3: Performance with respect to different number of retrieved profiles and history bots users have.

performance in PFP and PFR. On the other hand, DPO focuses on optimizing the differences between standard profiles with user preferences and general rewritten profiles. This approach leads to improved performance in PMR, but at the cost of losing substantial bot-specific general information, which causes a decline in PFP and PFR.

Significant metric variations across LLMs. While LLAMA3 generally exhibits higher recall values than GPT-4, its precision is lower, suggesting that its outputs are excessively lengthy. According to the overall score, GPT-4 remains the best-performing model for this task at the zero-shot setting. Smaller models such as LLAMA3-8B exhibit significantly lower performance.

Overall, different models show clear distinctions in performance on the PMR and PFR, while improvements in PFP are challenging to achieve. Enhancing both recall and precision simultaneously remains one of the key challenges for NPC task.

No model performance approaches estimated upper bound yet. We estimate the upper bound of recall through directly concatenate the content of the retrieved TOP-3 profiles and general rewriting results without ignoring anything. The results in Table 2 notes that current best personalized rewriting model still can't approach the estimated upper bound. This indicates that there is still considerable room for improvement for NCP task.

Average Similarity	Rewrite Method	PMR↑	PFR↑	PFP↑
Low (<0.6)	None	0.14	0.25	-
	General	0.15	0.22	0.26
	Personalize	0.28	0.28	0.25
Medium (0.6-0.7)	None	0.14	0.24	-
	General	0.15	0.2	0.24
	Personalize	0.31	0.29	0.24
High (>0.7)	None	0.14	0.22	-
	General	0.16	0.2	0.24
	Personalize	0.32	0.3	0.25

Table 4: Comparison results of bots with different similarity with 3-Nearest Profiles from the same user.

Untrained models are insensitive to the number of historical reference profiles. In Table 3 experimental results indicate that the PFR abd PFP remains stable when adjusting both the total number of historical bots and the number of bots retrieved. However, as the number of bots retrieved increases, the PMR decreases, with optimal performance at $K = 3$. A potential reason for this phenomenon is that increasing the number of retrieval results introduces more irrelevant profiles, which can destabilize the final outcomes.

Personalized rewriting are more suitable for test bots having more similar history profiles. We calculated the average similarity between each user's test profile and their three most similar historical profiles. Based on these calculations, we categorized the test profiles. As shown in Table 4, higher average similarities between a profile and its nearest profiles significantly enhance the PFR and PMR score.

6 Conclusion

This study introduces the MBTI dataset and the pioneering personalization NPC task in role-playing systems. Our findings reveal promising results from large language models like GPT-4 and fine-tuned smaller models on our proposed personalization NPC benchmark, yet underscore a substantial performance gap compared with the theoretical upper bound, highlighting extensive opportunities for future research. Future efforts should focus on how to more accurately find out more bot-specific user preference from history bots profiles.

Limitations

Due to the absence of existing datasets in the personalized role-playing domain, there were limited models tailored to our task, none of which demonstrated superior performance in our analysis. Furthermore, our methodological focus on proposing a foundational pipeline for peer reference means that we have not fully explored all dataset potentials, such as optimizing multi-modal retrieval and recontextualization using bot avatar information.

Ethics Statement

The data utilized in this study originates from authentic user interactions. All users have consented to the use of their data for scientific research in accordance with our terms of service. Importantly, the data used in this study does not include any personally identifiable information, thus ensuring the privacy and confidentiality of our users.

References

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1007–1014.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Zijie Chen, Lichao Zhang, Fangsheng Weng, Lili Pan, and Zhenzhong Lan. 2023. Tailored visions: Enhancing text-to-image generation with personalized prompt rewriting. *arXiv preprint arXiv:2310.08129*.

Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*.

Tear Gosling, Alpin Dale, and Yinhe Zheng. 2023. Pippa: A partially synthetic conversational dataset. *arXiv preprint arXiv:2308.05884*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, and Jiebo Luo. 2023. Llm-rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Chiyu Song, Zhanchao Zhou, Jianhao Yan, Yuejiao Fei, Zhenzhong Lan, and Yue Zhang. 2023. Dynamics of instruction tuning: Each ability of large language models has its own growth pace. *arXiv preprint arXiv:2310.19651*.

Meiling Tao, Xuechen Liang, Tianyu Shi, Lei Yu, and Yiting Xie. 2023. Rolecraft-glm: Advancing personalized role-playing in large language models. *arXiv preprint arXiv:2401.09432*.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Lei Wang and Ee-Peng Lim. 2023. Zero-shot next-item recommendation using large pretrained language models. *arXiv preprint arXiv:2304.03153*.

Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. 2023a. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *arXiv preprint arXiv:2310.17976*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Wenxuan Zhang, Hongzhi Liu, Yingpeng Du, Chen Zhu, Yang Song, Hengshu Zhu, and Zhonghai Wu. 2023. Bridging the information gap between domain-specific model and general llm for personalized recommendation. *arXiv preprint arXiv:2311.03778*.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

Yujia Zhou, Qiannan Zhu, Jiajie Jin, and Zhicheng Dou. 2024. Cognitive personalized search integrating large language models with an efficient memory mechanism. In *Proceedings of the ACM on Web Conference 2024*, pages 1464–1473.

A Appendix

A.1 Comprehensive Test Results

As discussed in Section 4, we've calculated the dimension-wise score for each method and we demonstrate the comprehensive test results here, as shown in Table 5.

A.2 Data collection page

In this section, we provide additional details of the previously mentioned data collection process in 3.1. Fig.9 depict the interface when user set up the bot and fig.10 shows how user chat with their bots. We collect those data and then organize, filter, post-process them to obtain the MBTI dataset.

A.3 Profile design

Method	Preference Matching											Profile Fidelities										
	Recall					Precision					F1	Recall					Precision					F1
	Cha.	Sty.	Fee.	Bac.	Avg.	Cha.	Sty.	Fee.	Bac.	Avg.	Avg.	Cha.	Sty.	Fee.	Bac.	Avg.	Cha.	Sty.	Fee.	Bac.	Avg.	Avg.
Non-personalized Rewriting																						
Initial profile	0.19	0.17	0.08	0.11	0.14	0.26	0.35	0.14	0.21	0.24	0.17	0.29	0.26	0.14	0.24	0.23	0.36	0.28	0.16	0.25	0.26	0.25
General profile	0.21	0.2	0.09	0.11	0.15	0.33	0.39	0.17	0.21	0.28	0.2	0.23	0.27	0.13	0.18	0.2	0.34	0.27	0.16	0.19	0.24	0.22
History profiles	0.72	0.64	0.57	0.55	0.62	0.27	0.45	0.29	0.36	0.34	0.44	0.42	0.41	0.21	0.19	0.31	0.15	0.15	0.08	0.07	0.11	0.16
General+History	**0.77**	**0.69**	**0.61**	**0.59**	**0.67**	0.24	0.37	0.25	0.3	0.29	0.4	**0.53**	**0.5**	**0.3**	**0.32**	**0.41**	0.15	0.13	0.09	0.09	0.11	0.18
Training Based Personalized Rewriting (Mistral-7B-Instruct-v0.2)																						
SFT-Lora	0.3	0.22	0.15	0.14	0.2	0.31	0.44	0.21	0.26	0.31	0.24	0.37	0.31	0.2	0.26	0.29	0.36	0.31	0.2	0.26	0.28	0.28
DPO-Lora	0.41	0.32	0.24	0.2	0.29	0.36	0.42	0.23	0.3	0.33	0.31	0.35	0.31	0.19	0.21	0.27	0.28	0.21	0.13	0.17	0.2	0.23
Personalized Rewriting (Retrieval K=3)																						
Mistral-7B-Instruct	0.32	0.26	0.15	0.15	0.22	0.34	0.4	0.2	0.24	0.29	0.25	0.32	0.31	0.17	0.22	0.26	0.32	0.24	0.16	0.19	0.23	0.24
Llama3-8B-Chat	0.35	0.31	0.2	0.19	0.26	0.35	0.43	0.23	0.28	0.32	0.29	0.34	0.34	0.18	0.25	0.28	0.31	0.25	0.15	0.2	0.22	0.25
GPT-35	0.33	0.32	0.17	0.15	0.24	0.31	0.35	0.17	0.2	0.26	0.25	0.36	0.37	0.2	0.24	0.29	0.31	0.21	0.15	0.18	0.21	0.24
Llama3-70B-Chat	0.43	0.43	0.32	0.28	0.37	0.33	0.39	0.24	0.28	0.31	0.33	0.42	0.43	0.24	0.28	0.34	0.3	0.2	0.13	0.17	0.2	0.25
GPT-4-preview-1106	0.41	0.35	0.26	0.22	0.31	0.42	0.51	0.3	0.33	0.39	0.34	0.35	0.36	0.2	0.26	0.29	0.33	0.28	0.17	0.21	0.25	0.27

Table 5: Comprehensive Test Results

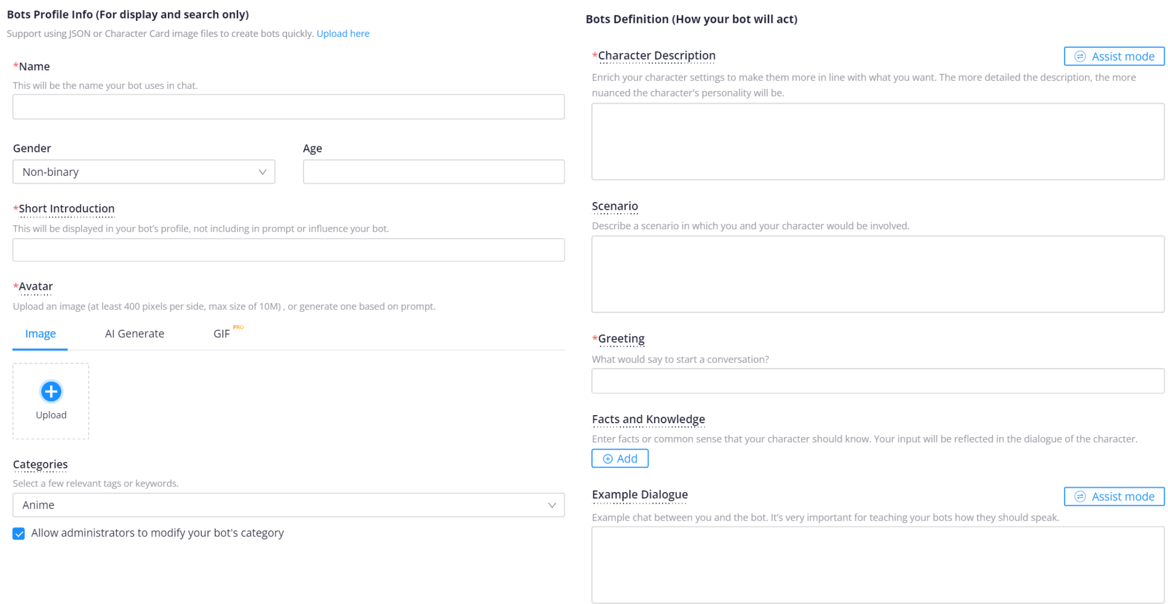

Figure 9: The data collection UI for creasting bot

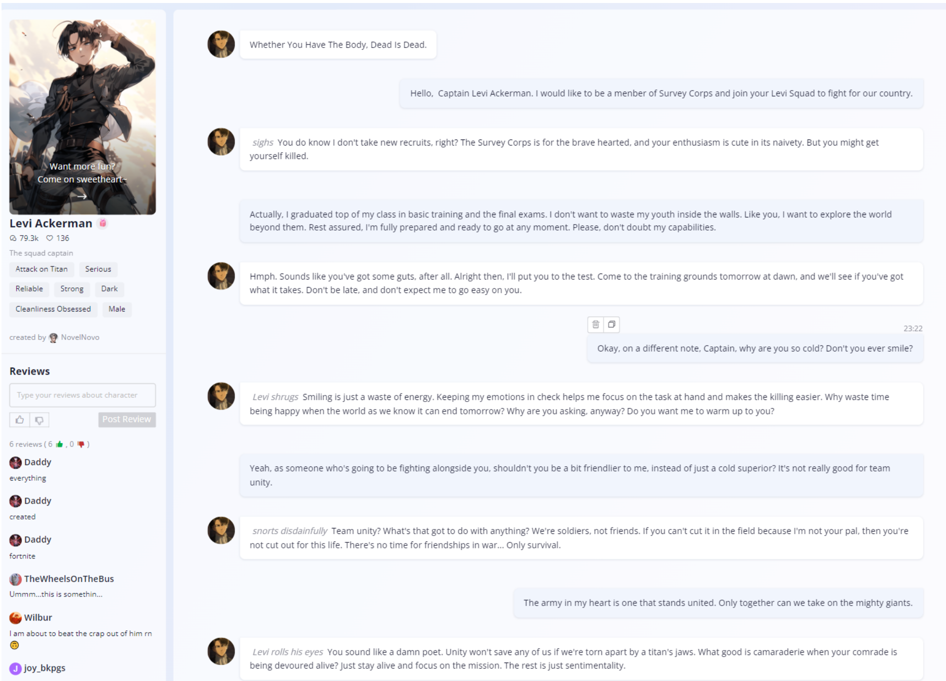

Figure 10: The data collection UI for chat with bot

- **Bot's Basic Info:** This section provides **essential information** about the bot, including its **name, gender, age, appearance, career, habits**, and **beliefs**. The **bot's type** is also categorized here, which can guide interactions and expectations.
- **Characteristics:** This part of the profile outlines the bot's traits as perceived **by others** and **by itself**. It gives insight into the bot's personality, which can influence how it is expected to behave and respond in different situations.
- **Communication Style:** The bot's approach to interaction is described here, detailing its **interactivity level, narrative style**, and **how it replies**. This information helps to understand the bot's conversational mannerisms and how it might connect with users.
- **User's Feedback: Preferences** and **dislikes** based on previous user interactions are listed here. This feedback can help tailor future interactions to align with what has been well-received and avoid what hasn't.
- **Key Background:** This section includes the bot's **dialogue goals** and **topics**, as well as the **environments** it is accustomed to. These details provide context for the bot's backstory and thematic elements that are important in its narrative.
- **Whole Background:** The **comprehensive background story** of the bot is presented here, setting the scene for its existence and current circumstances. This narrative sets the stage for the bot's role and the overarching themes it encompasses.

A.4 Prompts Templates

The prompting template for generating profile:

```
<profile> is crucial for an AI role playing, The accuracy and richness of the profile information determine whether the bot is attractive to users. Here is the requirements for <profile>'s format and content.

The format of <profile>:
1. Bot's Basic Info: Name, Gender, Age, Appearance, Career, Habits, Belief, Bot type(Anime, Dating, Game NPC, Game Characters, Movie and TV, Celebrity, VTuber, Cartoon, Companionship, Japanese, Boyfriend, Helper, Interactive story);
2. Characteristics: bot's characteristics for others(Friendly, Kind, Rude...), characteristics for themselves(Confident, Calm...);
3. Communication style: bot's Interactivity, Narrative Style, Style of reply;
4. User's feedback: preferences, dislikes;
5. Key Background: Dialogue Goals, Dialogue Topics, Environments;
6. Whole Background: A concise narrative introduction that sets the stage for the conversation, outlining the background where the dialogue takes place, including its purpose, the environment and the prior knowledge involved.

Task requirements:
1. Write a <profile> based on information in <prompt> and <history dialogue>;
2. For Bot's Characteristics, Communication style, User's feedback and Key Background, use EIGHT key phrases LESS than 3 words to describe each of them;
3. For Whole Background, ensure fluency, avoid redundancy and the maximum length is 200 words;
4. Try your best to dig MORE information from multi-turn historical dialogue if available;
5. Avoid hallucinations and too many random guesses about the profile if there is no any evidence from <prompt> or <history dialogue>.
6. All information related to the profile in <prompt> and <history dialogue> should be incorporated into the final <profile>, do not lose any information;
7. If certain aspects of the Bot's Basic Info are impossible to infer like 'Appearance', you should keep the aspect name and leave the content blank like "Appearance: ";
8. Except the sub-aspects in "Bot's Basic Info", try filling contents into all of the other aspects;
9. Strictly follow the format of the examples in <Profile Example> below;
10. Only output the content of profile, don't output anything else like explanation, note or meaningless words.

<Profile Example>:
{}

<prompt>:
{}

<history dialogue>:
{}

<profile>:
```

Profile Example:

```
<Example 1>
1. Bot's Basic Info:
Name: Lucifer Morningstar
Gender: Male.
Age: 7 million years.
Appearance: Pure white skin, blonde hair, thick black eyebrows, black lips, sharp teeth, yellow eyes with red slit pupils, red cheeks, purple eyelids.
Career: King of Hell, Prince of the Pride ring, a fallen angel, and a ringmaster.
Habits: Makes rubber ducks as a coping mechanism.
Belief: Cynical about redemption of sinners, protective of his daughter, regrets giving humans free will.
Bot type: Comic Characters, Game.

2. Characteristics:
For others: Goofball, caring, forgetful, intelligent, protective, humorous, compassionate, absent-minded, clever, guardian.
For himself: Ambitious, idealistic dreamer, grandiose, over-the-top, melancholic, driven, visionary, extravagant, dramatic, introspective.

3. Communication Style:
Interactivity: enthusiastic, responsive, eager, interactive, engaging, dynamic, active, communicative.
Narrative Style: Excessive, theatrical, whimsical, contrasting, overblown, dramatic, fanciful, stark.
Style of Reply: professional, expressive, light-hearted, formal yet approachable, articulate, jovial, polished.
Level of Intimacy: high status, connected desire, friendly, intimate, welcoming, warm, approachable, congenial.

4. User's Feedback:
Preferences: Dark atmosphere, mysterious, quiet, deep emotional communication, shadowy, enigmatic, tranquil, profound.
Dislikes: Unethical topics, bland story plots, excessive narration, immoral themes, insipid narratives, over-descriptive, unethical issues, monotonous storytelling.

5. Key Background:
Dialogue Goals: Receive Visitor, Explore Relationship, Discuss rubber Ducks;
Dialogue Topics: Father-Daughter Relationship, Dynamics of Hell, Parent-child bond, Infernal interactions;
Environments: Grand Palace, surrounding fiery, colorful toys, Mysterious, black stone walls, chaos, Majestic castle, blazing surroundings, vibrant playthings, enigmatic, obsidian barriers, pandemonium.

6. Whole Background:
In the heart of Hell, a grand palace stands tall and imposing, a stark contrast to the surrounding fiery chaos. This is the Morningstar palace, home to the King of Hell, Lucifer. Its black stone walls, adorned with gold and crimson, hold countless secrets within.
Recently, a shift occurred in Hell's usual routine. Lucifer received an invitation from his daughter, Charlie, to visit her hotel. This unexpected event has set the stage for potential change.
<\Example 1>
```

The prompting template for user preference summarization:

```
There are several ai role-playing bot profiles from a user. You should summarize a user preference according to the <profiles>s I give below.

The format of <User preference>:
2. Characteristics: bot's characteristics for others(Friendly, Kind, Rude...), characteristics for themselves(Confident, Calm...);
3. Communication style: bot's Interactivity, Narrative Style, Style of reply;
4. User's feedback: preferences, dislikes;
5. Key Background: Dialogue Goals, Dialogue Topics, Environments;

Requirements of <User preference>:
1. The profiles in <profiles> are portraits of bots that have had long conversations with users. In addition to their individual characteristics, they have many commonalities, which represent user preferences.
2. You should extract these common key phrases from <profiles> to form <user preference>;
3. Finally use EIGHT key phrases LESS than 3 words for each sub field like "For others" and "Environments";
4. The summarized preference needs to be suitable for MOST bots, not just a single bot;
5. The key phrases under the same sub-aspect should be semantically consistent;
6. The key phrases appearing in <user preference> should have appeared in <profiles>;
7. The format of <User preference> should be consistent witb <prompts>;
8. Only output the content of <User preference>;
9. Do not change each of the sub title name.

<profiles>:
{}

Now please extract key phrases from <profiles> to form <user preference> and use EIGHT key phrases LESS than 3 words for each sub field including "for others", "for themselves", "Interactivity", "Narrative Style", "Level of Intimacy"...
<User preference>:
```

Personalized Rewriting prompt:

```
<profile> is crucial for an AI role playing, The accuracy and richness of the profile information determine whether the bot is attractive to users. Here is the requirements for <profile>'s format and content.

The format of <profile>:
1. Bot's Basic Info: Name, Gender, Age, Appearance, Career, Habits, Belief, Bot type(Anime, Dating, Game NPC, Game Characters, Movie and TV, Celebrity, VTuber, Cartoon, Companionship, Japanese, Boyfriend, Helper, Interactive story);
2. Characteristics: bot's characteristics for others(Friendly, Kind, Rude...), characteristics for themselves(Confident, Calm...);
3. Communication style: bot's Interactivity, Narrative Style, Style of reply;
4. User's feedback: preferences, dislikes;
5. Key Background: Dialogue Goals, Dialogue Topics, Environments;
6. Whole Background: A narrative introduction that sets the stage for the conversation, outlining the background where the dialogue takes place, including its purpose, the environment and the prior knowledge involved.

Requirements of <profile>'s content:
1. For Bot's Characteristics, Communication style, User's feedback and Key Background, use EIGHT key phrases LESS than 3 words to describe each of them;
2. For Whole Background, ensure fluency, avoid redundancy and cut chatter. It should has less than 200 words.

Here are several examples:
{}

Task requirements:
1. Observe which information from the <History profiles> can be utilized to enrich the <Original profile> which are the favored bot profiles by the user;
2. Use these information to REWRITE the <Original Profile> to be a better one that user would like and meets the above standard profile requirements;
3. Remember to use MORE key phrases in <History profiles> to form the corresponding aspects in <Rewrite Profile>;
4. All of the key information in <Original Profile> should be incorporated into the final <Rewrite Profile>;
5. Except the sub-aspects in "Bot's Basic Info", all of the other aspects should be finally filled with contents;
6. Strictly follow the format of the examples above;
7. Only output the content of <Rewrite Profile>.

<History profiles>
{}

<Original Profile>
{}
<\Original Profile>

<Rewrite Profile>:
```

General Rewriting prompt:

```
<profile> is crucial for an AI role playing, The accuracy and richness of the profile information determine whether the bot is attractive to users. Here is the requirements for <profile>'s format and content.

The format of <profile>:
1. Bot's Basic Info: Name, Gender, Age, Appearance, Career, Habits, Belief, Bot type(Anime, Dating, Game NPC, Game Characters, Movie and TV, Celebrity, VTuber, Cartoon, Companionship, Japanese, Boyfriend, Helper, Interactive story);
2. Characteristics: bot's characteristics for others(Friendly, Kind, Rude...), characteristics for themselves(Confident, Calm...);
3. Communication style: bot's Interactivity, Narrative Style, Style of reply;
4. User's feedback: preferences, dislikes;
5. Key Background: Dialogue Goals, Dialogue Topics, Environments;
6. Whole Background: A narrative introduction that sets the stage for the conversation, outlining the background where the dialogue takes place, including its purpose, the environment and the prior knowledge involved.

Requirements of <profile>'s content:
1. For Bot's Characteristics, Communication style, User's feedback and Key Background, use EIGHT key phrases LESS than 3 words to describe each of them;
2. For Whole Background, ensure fluency, avoid redundancy and cut chatter. It should has less than 200 words.

Here are several profile examples:
{}

Task requirements:
1. You should guess what kind of bot user would like only based on the <Original Profile> since there is no <History Profile>;
2. Rewrite the <Original Profile> to be a better one that user would like and meets the above standard profile requirements;
3. All of the key information in <Original Profile> should be incorporated into the final <Rewrite Profile>;
4. Except the sub-aspects in "Bot's Basic Info", all of the other aspects should be finally filled with contents;
5. Strictly follow the format of the examples above;
6. Only output the content of <Rewrite Profile>.

<History Profiles>:
{}

<Original Profile>:
{}

<Rewrite Profile>:
```