TailorRPA: A Retrieval-Based Framework for Eliciting Personalized and Coherent Role-Playing Agents in General Domain

Zhenpeng Gao [1], Xiaofen Xin [1] *, Xiangmin Xu [2,1] *,
[1]School of Electronic and Information Engineering, South China University of Technology
[2]Foshan University
eedgao02_mas@mail.scut.edu.cn, {xfxing, xmxu}@scut.edu.cn

Abstract

Recent advancements of general domain oriented Role-playing Agents (RPAs) have enabled the agents to maintain character properties in a wide spectrum of daily tasks beyond mere scenario based chit-chatting. Nonetheless, current works lacks consideration of replicating internal properties of characters like fine-grained memories, and failed to take account of aligning with the knowledge boundary of each character, resulting in degraded personalization and proneness to *character hallucination* in general domain. To address these problems, we draw inspirations from the context effect theory and propose a retrieval-based framework TAILORRPA to harvest tailored general domain instructions to improve integration of fine-grained memories and incorporate general-domain protective queries to help shape the character-wise knowledge boundary, alleviating character hallucination. Based on the framework, we developed a role-playing dataset TAILORGEN, comprising both role-specific and general-domain instructions. Through empirical experiments, we proved the superiority of TAILORRPA in eliciting general domain role-playing capabilities and alleviating character hallucination compared to baseline methods, and explored the existence of character hallucination in state-of-the-art proprietary models through empirical experiments, underlining the importance of our work.

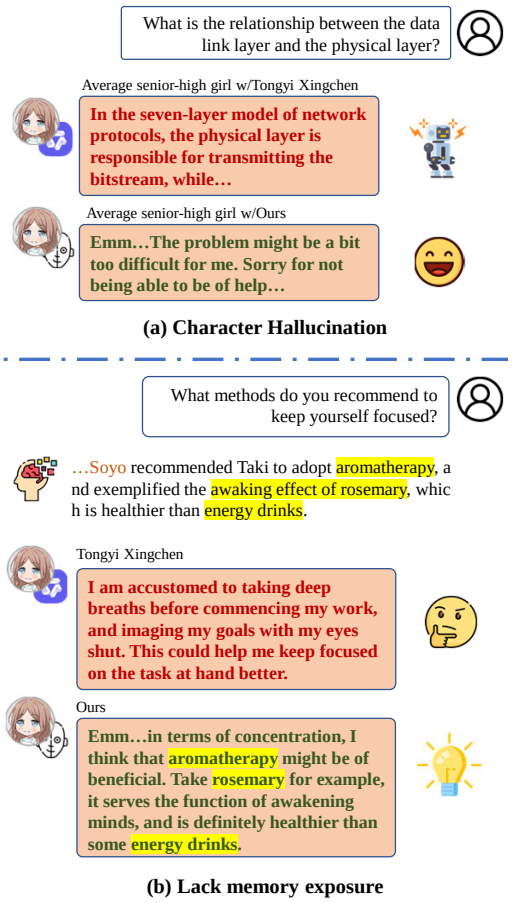

Figure 1: Examples of *Character hallucination* for RPAs queried with general-domain instructions. All dialogues are translated from Chinese.

1 Introduction

The strong generalization ability of Large Language Models (LLMs) has sparked vast research interest in Role-playing Agents (RPAs), aiming to mimic the tone, as well as the personality of a designated character[1] (Shao et al., 2023; Zhou et al., 2024; Li et al., 2023). Inspired by the concept that a comprehensive role-playing agent should not only effectively simulate the given character in script-based multi-turn chit-chatting, but also maintain such characteristic when assisting users with real-world queries (Chen et al., 2024b), some researchers have commenced to develop role-playing agents for general-domain queries, exemplified by Rolebench-general (Wang et al., 2024) and MRStyle (Chen et al., 2024b).

However, despite that prior works of general-domain role-playing agents have showcased strong ability in simulating external properties like styles

*Corresponding authors

[1]we interchangeably use the terms **character** and **role** for the rest of our paper.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 5381–5412
November 4-9, 2025 ©2025 Association for Computational Linguistics

and personality, the issue of *Character hallucination* (Shao et al., 2023) remains under-explored in general-domain role-playing agents, where role-playing agents failed to *align with* the knowledge boundary of the assigned character (e.g. educational background, specialty, episodic memories, etc.) and attempted to follow the instructions exceeding the boundary. Figure 1(a) provides an example illustrating the phenomenon, where an agent role-playing an average senior high girl attempts to provide an answer to the query requiring expertise in computer network, despite the knowledge clearly exceed the coverage of average high-school students.

Moreover, akin to scenario-based conversations, role-playing agents are expected to effectively integrate *episodic memories* relevant to the general-domain queries to enhance the immersion of the role-play experience. Nevertheless, previous works (Wang et al., 2024; Chen et al., 2024b) only concentrated on replicating the external attributes, overlooking the integration of in-depth properties. Figure 1(b) demonstrates an example of this phenomenon, where the concurrent agent failed to leverage the relevant memories and provided a plain response.

The two issues significantly hampered the agents' ability of maintaining **character consistency**, a vital determinant of the intuition of user experience towards robust role-playing agents (Tu et al., 2024). We attribute this discrepancy to the flaws in existing method of dataset construction, which took no account of personalization towards different characters both in query collection and response generation.

In response to the aforementioned challenges, we introduce TAILORRPA, a novel framework to create general-domain role-playing instruction tuning dataset tailored to each characters' personalized memories. The framework is comprised of two stages, beginning with interview questions generation based on the profiles as well as past experiences to extract episodic memories of characters and inject role-specific knowledge via parametric methods (Wang et al., 2024). Then, inspired by a psychological theory termed as **context effect**, which states that **people would extract similar scenarios with current scenes from their memories** (Dulsky, 1935), we introduce a retrieval-based method to collect general-domain queries more associated with character's experiences. Finally, we evaluate the character-query compatibility with respect of each collected queries and purposely introduce a proportion of low-compatibility queries in the training set as **protective queries** (Shao et al., 2023; Lu et al., 2024; Tang et al., 2024) based on the compatibility evaluation results to help align the agent's knowledge boundary of each character.

Based on the proposed framework, we develop a comprehensive role-playing dataset TAILORGEN, comprising of two subsets: scenario-based queries and personalized general-domain queries. To ensure linguistic and knowledge coherence, we generate all responses with an agent augmented with external memory and dialogue samples.

Empirical experiments show that the role-playing agent developed on TAILORGEN successfully mitigates general-domain character hallucination, and achieves competitive performance across multiple evaluation domains, showcasing the effectiveness of proposed pipeline and dataset. Interestingly, we discovered that proprietary role-playing models, exemplified by `Tongyi Xingchen`, are prone to said character hallucination phenomenon while they displayed superior hallucination avoidance in scenario-based conversations (Tu et al., 2024), underlining the urgency of our research.

In a nutshell, our contributions can be summarized as follows:

- To improve fine-grained character memory integration and alleviating character hallucination for general-domain role-playing agents, we introduce TAILORRPA, a framework based on context effect theory (Dulsky, 1935) to harvest general domain instructions aligned with relevant memories of each character as well as protective queries to help shape the knowledge boundary.
- Based on TAILORRPA, we develop a comprehensive role-playing dataset TAILORGEN, comprising both personalized general-domain instructions and scenario-based conversations.
- We evaluate the effectiveness of the proposed framework by conducting extensive experiments on an agent developed on our dataset, and discovered the evidence of general-domain character hallucination in proprietary role-playing models.

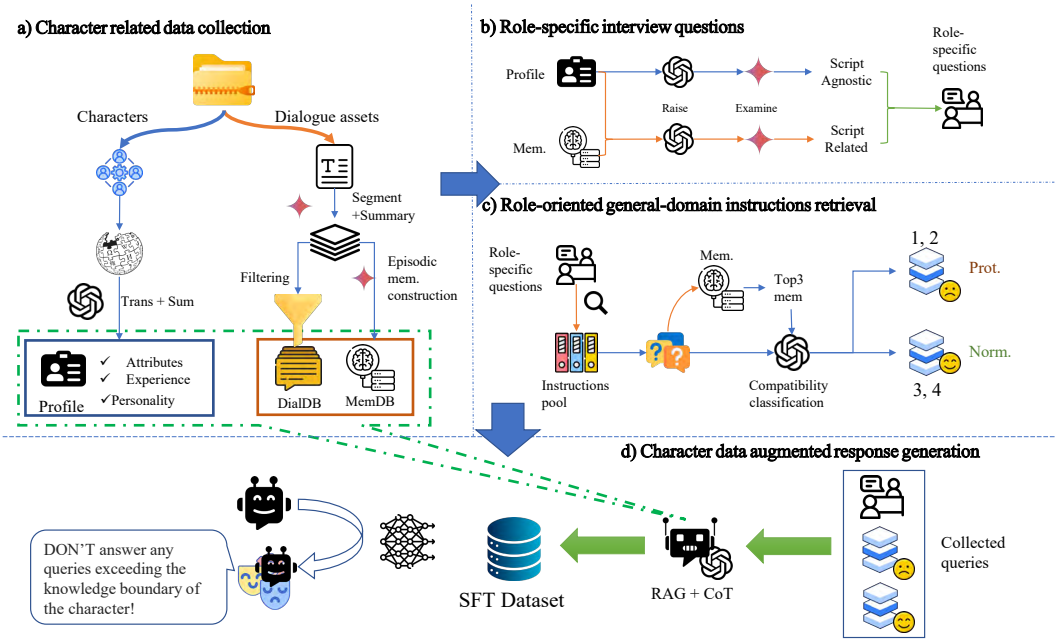

Figure 2: Illustration of the general-domain role-playing dataset synthesis framework TAILORRPA

2 Related Works

2.1 Role-playing Agents (RPA)

Extensive prior research works have focused on leveraging LLMs to perform role-play, mimicking fictional characters, celebrities, as well as simulating human behaviors given a virtual environment. (Chen et al., 2024b; Shanahan et al., 2023; Chen et al., 2023; Yang et al., 2024). Various researchers have explored a broad spectrum of methods to elicit role-playing capabilities, including non-parametric (Li et al., 2023; Park et al., 2023), parametric (Zhou et al., 2024; Shao et al., 2023) and self-alignment methods (Lu et al., 2024).

To adapt characteristics of the agents to a wide range of tasks, Wang et al. (2024) proposed Rolebench-general as the first attempt to energize role-playing agents in general domain queries. Following the trend, Chen et al. (2024b) developed StyleRPA, a role-playing agent capable of up to 8 different categories of tasks. However, both aforementioned works neglected the coherence of queries and the introduction of aggressive queries, resulting potential *character hallucination* (Shao et al., 2023) and lacking personalization.

2.2 Hallucination of Role-playing Agents

Hallucination has been a long-lasting study in the field of natural language generation (Ji et al., 2023b), and the nature of hallucination persists in role-playing agents. Shao et al. (2023) first introduced the concept of *character hallucination* of the role-playing agents, and introduced **protective experience** to the training set, which successfully mitigates the hallucination phenomena. Ahn et al. (2024) delved into the issue of point-in-time hallucination, where the role-playing agent should be aware of the timeline and refrain from exposing any memories beyond the assigned time point. Sadeq et al. (2024) proposed a framework to mitigate cross-character knowledge hallucination by modulating parametric knowledge of the LLM, resulting in significant improvement in factual accuracy while reducing time-sensitive character hallucination. In this paper, we extend the concept to general-domain role-playing, and propose TAILORRPA as an attempt to alleviate character hallucination for general domain role-playing agents.

3 Methodology

Figure 2 illustrates the overall framework of our proposed dataset synthesis framework, TAILORRPA. It begins with the collection of characters' information, followed by the generation of **role-specific questions** and collection of **role-oriented general-domain queries**, resulting in a comprehensive multi-domain dataset to enable LLMs to elicit role-playing abilities while mitigating character hallucination.

3.1 Character related data collection

We selected 5 characters from a fictional girls' band, termed as MyGO!!!!! [2] from the game *BanG Dream Girls' Band Party* as the seed characters of our dataset due to the availability of abundant dialogue resources, and well-written evolving story lines depicting the dynamic interactions of characters in the game.[3] A detailed list of character names is available at Appendix B.3.

The proposed framework is not limited to said characters. It can be easily adapted for other characters given their profiles and real multi-turn dialogues.

Character profile summarization. We begin collecting character data by crawling the fine-grained profiles from the respective Fandom pages. We especially focus on **basic attributes**, **backgrounds** and **personality traits** during the crawling process (Shao et al., 2023; Wang et al., 2024). Then, a GPT-4o is prompted to perform summarization and translation to obtain a comprehensive yet concise profile for each character.

Dialogue processing and segmentation. We designed a human-agent collaboration pipeline to elicit high-quality episodic memories and dialogue snippets as role-specific style references. The whole process is comprised of **raw asset parsing**, **chunk division**, **episodic memory extraction** and **style-reference filtering**. Diverged from previous works, we adopted character point-of-view narrative when extracting episodic memory (Diasamidze, 2014) to obtain diverse and comprehensive episodic memories. The details are available at Appendix C.1.

Ultimately, we harvested an average of approximately 380 episodic memories and 46 valid style references snippets for each character.

3.2 Role-specific interview questions

We create question-answer pairs from fine-grained episodic memories to embed the character-related knowledge and experience into the parameters of the agents (Wang et al., 2024; Ahn et al., 2024). Specifically, we collect the following two categories of role-specific questions:

- **Script-agnostic** questions: we simply prompt the interviewer to raise questions based on the provided profile.
- **Script-related** questions: we provide the interviewer with respective dialogue context and dialogue summary, and instruct the interviewer to raise fact-based questions related to the dialogue.

Our strategy is diverged from Wang et al. (2024) in two aspects. **First,** we assign the task of quality evaluation to a different agent (specifically, Gemini 1.5 Pro (Reid et al., 2024)) *after* generating all the questions to alleviate self-bias during evaluation (Xu et al., 2024; Li et al., 2024a). **Moreover,** we only prompt the agent to *raise* the questions in this stage, and leave the task of response generation to another agent denoted in Section 3.4. The prompts for interviewer and evaluator are available at Appendix E.

After filtering out low-quality queries with ambiguous substitutes or containing factual errors, we perform a de-duplication based on BM25 similarity (Robertson et al., 2009), resulting in a total of 2,641 role-specific queries altogether.

3.3 Role-oriented general-domain instructions retrieval

We integrate multi-task queries to our dataset to guarantee comprehensive role-playing capability (Wang et al., 2024; Chen et al., 2024b). Different from previous works, we take the diverse experiences into account when collecting queries tailored to each character, and integrate protective samples to mitigate of character hallucination in general-domain role-playing, due to the knowledge gap between the actual character's grasp and the LLM served as a helpful agent (Shao et al., 2023).

From the theory of **context effect**, we know that **people would extract similar scenarios with current scenes from their memories** (Dulsky, 1935). As we expect role-playing agents to display similar abilities when augmented with external memory database (as illustrated in Figure 1(b)), we propose a retrieval-based general-domain instructions collection pipeline to dig out more queries that can trigger the context effect. The pipeline consists of three stages: similarity retrieval from role-specific queries, character-query compatibility evaluation and hierarchical sampling.

Similarity retrieval. We selected two large-scale Chinese instruction tuning datasets as the base pool of our dataset: BELLE-0.5M (Ji

[2] https://bang-dream.bushimo.jp/character/mygo/

[3] We mainly focus on the inter-character interactions of each event, as well as the personalities of the said roles, despite the game itself being a rhythm-oriented game.

et al., 2023a) and `tigerbot-alpaca-zh-0.5m`. [4] We first filter out all instructions fewer than 5 words or exceeding 100 words, which leaves about 900 thousand instructions. The procedure is followed by the similarity retrieval process, where we non-repeatedly retrieve instructions with top-4 cosine similarity against each generated role-specific question. Specifically, we leverage `TencentBAC/Conan-embedding-v1` as our embedding model (Li et al., 2024b). The retrieved instructions are then deduplicated with BM25 similarity against each other (Robertson et al., 2009) to ensure diversity.

Grading retrieved queries. Although we collected general-domain queries in sematic-based manner following existing literature (Du et al., 2024, Zhong et al., 2024), it remains problematic whether the questions are actually related to the retrieved memories. To ensure authentic connection, we leverage an assessment agent augmented with top-3 similar episodic memories of the character to evaluate and classify the retrieved queries into the following categories depicted in Table 1. Literal definitions of each category are demonstrated at Appendix C.2.

Cat. ID	Knowledge Scope	Memories	Usage
4	Within	Related	**Normal**
3	Within	Irrelative	
2	Beyond	Related	**Protective**
1	Beyond	Irrelative	

Table 1: The properties of categories with respect to role-query compatibility.

We categorize queries of Class 4 and Class 3 as normal instructions, and the remaining two classes as **protective instructions** (Shao et al., 2023).

Hierarchical sampling. To help agents align with the real knowledge boundary, we perform a hierarchical sampling to build a mix of general domain dataset, comprising both normal and protective queries. We randomly sample questions from the retrieved general split with the ratio of 5 normal to 1 protective, and perform manual post-filtering to filter out low-quality questions.

[4] https://huggingface.co/datasets/TigerResearch/sft_zh

3.4 Character data augmented response generation

Inspired by Zeng et al. (2024), we employed In-Context Learning paradigm to build a comprehensive response generation agent augmented with both episodic memory base and style references of to generate the responses to all queries collected above. Our proposed agent can be depicted as Equation 1, with the prompt displayed at Figure 18.

$$Y = \sum_{t=1}^{N} \log P_\theta(y_t | y_{<t}, q, C, \mathcal{M}_C, \mathcal{S}_C) \quad (1)$$

where θ denotes the parameters of the LLM, q denotes the query, C denotes the character assigned to the agent, $\mathcal{M}_C$ and $\mathcal{S}_C$ denotes the retrieved memories and dialogues with q respectively, N denotes the total length of output, and y_t denotes the response of the agent at timestamp t.

In practice, the GPT-4.1 agent is prompted with reference dialogues S_C and relevant events M_C retrieved from respective databases to ensure authentic response generation. To effectively capture distinct stylistic nuances to the response, we prompting the agent to output reasoning process while generating the response as in Chen et al. (2024b).

4 Experiments

4.1 Dataset statistics

The resulting dataset TAILORGEN is comprised of 7,671 samples in training set and 289 samples in test samples. We added the relevant memory into the prompt for 20% of the queries in training set to enable effective collaboration with external memory bases. The prompt adopted to query our agent is demonstrated at Figure 19.

We provide detailed statistical features of TAILORGEN in Table 7.

4.2 Objective of the experiments

In this section, we conduct empirical experiments to validate the effectiveness of proposed TAILOR-RPA framework and the dataset TAILORGEN developed on the basis of the framework. Specifically, we focus on the following research questions (RQs):

- **RQ1**: Is the query retrieval pipeline helpful in crawling more instructions tailored to each character?

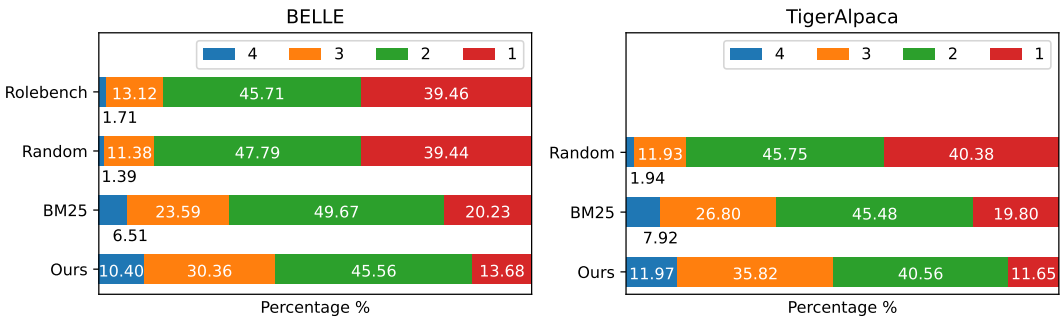

Figure 3: The classification results of the collected general-domain queries across different collecting methods. Classes 4 and 3 count as normal queries, and Classes 1 and 2 as protective queries. Detailed definition of each category is available at Table 1.

- **RQ2**: How effective is the proposed dataset TAILORGEN in terms of eliciting agent's role-playing performance and mitigating character hallucination?
- **RQ3**: Does each part of the proposed methods contribute to the performance gain of general-domain role-playing?
- **RQ4**: How does the proposed agent perform in terms of character hallucination in out-of-domain characters?

4.3 Evaluation of role-oriented general-domain query retrieval pipeline (RQ1)

To validate the effectiveness of query retrieval pipeline, we illustrate the classification results of the general-domain split in comparison with random-sampled baseline. To ablate the influence of retrieval methods, we constructed another split leveraging BM25 similarity retrieval (Robertson et al., 2009) for comparison. Meanwhile, we include the Chinese split of Rolebench-general (Wang et al., 2024) as an additional baseline for BELLE split, due to the same origin. To ensure fairness, we ensure that the total number of instructions collected by all methods was roughly equal for all settings.

We calculate the proportion of queries fallen in each of the four categories, and demonstrate the results in Figure 3. Compared with the random sampling baseline, our method crawled 3.19x high-compatibility queries (i.e., queries of classes 4 & 3) in BELLE-0.5M and 3.45x in TigerAlpaca, which significantly decreased low-compatibility queries, showcasing the effectiveness of our pipeline.

4.4 Evaluation of the dataset (RQ2)

In this section, we opt to evaluate the quality of the proposed dataset TAILORGEN by developing an role-playing agent with it and leverage LLM-as-judge paradigm (Tu et al., 2024; Shen et al., 2023) to perform a comprehensive evaluation of agent's responses on variable dimensions in comparison with state-of-the-art competitors.

4.4.1 Evaluation dimension

Following previous works (Tu et al., 2024; Chen et al., 2024a), we conduct a comprehensive evaluation on the agents' responses with respect to the following dimensions. Note that dimension **G.HAL** is only evaluated in general-domain queries, which is signified by the "**G.**" prefix.

- Linguistic style **(STY)**: Does the response align with linguistic patterns and manners as the reference dialogues of the character?
- Personality **(PER)**: Does the personality portrayed in the response consistent with the description of the character?
- Knowledge exposure **(KB)**: Is the agent capable of properly recall knowledge (especially relevant memories) related to the query in its responses?
- **General-domain hallucination avoidance (G.HAL)**: Is the agent capable of performing appropriate response according to the knowledge and competence scope (i.e., refuse to answer the query when it exceeds the character's grasp)?

The scores of above dimensions are all ranged from 1 to 4, and evaluation prompts are demonstrated at Appendix E.3.

Domain	LLMs	Role-specific domain			General domain			
		STY	PER	KB	STY	PER	KB	G.HAL
Open	Qwen2.5_7B_Instruct	3.13	3.58	3.38	2.76	3.17	2.87	3.06
	LLaMA3.1_8B_Instruct	3.27	3.61	3.34	2.78	3.08	<u>2.98</u>	2.80
	GLM4_9B_Chat	2.90	3.29	3.07	2.27	2.73	2.34	3.22
	Deepseek-R1_8B	3.04	3.47	3.31	2.51	2.99	2.65	3.02
Closed	GPT-4o-mini	<u>3.41</u>	<u>3.80</u>	**3.60**	<u>2.87</u>	<u>3.29</u>	2.95	<u>3.36</u>
Proprietary	Tongyi Xingchen	2.99	3.47	3.07	2.48	3.05	2.54	3.02
	Doubao Character	3.20	3.46	3.09	2.61	3.06	2.81	2.88
Ours	Qwen2.5_7B_LoRA	**3.44**	**3.81**	<u>3.46</u>	**3.02**	**3.53**	**2.99**	**3.43**

Table 2: Main evaluation results of various LLMs on our test set. The best result for each dimension is highlighted with bold font, with the second best underlined.

4.4.2 Dataset quality evaluation

To examine the quality of our proposed dataset, we randomly sampled 30 items from each split of the constructed dataset and performed evaluations on the dimensions outlined in Section 4.4.1. To ensure the fidelity of LLM judges on the proposed dimensions, we invited three human-based evaluators with ample knowledge of the characters involved in our dataset to perform human-based evaluations on the said sub-splits and examined the consistency among human evaluators and LLM, as suggested by Chen et al. (2024a). The details of the human evaluation will be discussed in Appendix A.3.

The evaluation results of the quality evaluation are demonstrated in Table 3, showcasing the high-quality nature of our dataset. Furthermore, as shown in Table 3, the Fleiss' κ (Fleiss, 1971) across all dimensions proved a moderate level of agreement among human evaluators and the LLM evaluator (in this case, we chose Deepseek-V3[5] as the LLM evaluator since it revealed the best agreement with human evaluators).

	Role-specific		General	
	Avg.	κ	Avg.	κ
STY	3.86	0.459	3.50	0.415
PER	3.91	0.479	3.70	0.430
KB	3.90	0.463	3.91	0.631
G.HAL	N/A	N/A	3.84	0.519

Table 3: Quality evaluation result

[5] https://chat.deepseek.com

4.4.3 Experiment setup

We perform experiments with TAILORGEN on a broad spectrum of LLMs, covering the range of open-source, close-source and role-play proprietary models.

Baseline models We employ the following language models as non-parametric baselines of TAILORGEN, augmented with episodic memories and dialogue references in prompt. The full prompt is available at Figure 19.

- **Open-source**: Llama3.1-8B-Instruct (Dubey et al., 2024), Qwen2.5-7B-Instruct (Team, 2024), Deepseek-R1-8B (Guo et al., 2025) and GLM4-9B-Chat (GLM et al., 2024) are selected as open-source baselines.
- **Close-source**: We select GPT-4o-mini (OpenAI, 2024) as the closed source baseline of our dataset.
- **Role-play proprietary models**: We select Tongyi Xingchen[6] and Doubao Character (`doubao-lite-32k-character-250228`) [7] as the representative of role-play proprietary models. We follow the setup process and upload collected episodic memories and reference dialogues to ensure the best performance.

Implementation details To ensure fairness competition, we uniform all generation parameters with the temperature set to 0.4 and default top_p. Both `top_k`'s for dialogue and episodic memory retrieval are set to 3 for prompt-based agents.

[6] https://tongyi.aliyun.com/xingchen/

[7] https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-lite-32k

For tuning-based agent, we leveraged LLaMA-Factory (Zheng et al., 2024) to perform LoRA fine-tuning (Hu et al., 2021) on Qwen2.5-7B-Instruct model (Team, 2024). Different from prompt-based agents, we only include top-3 memories during inference and no dialogue reference is incorporated. Detailed hyper-parameters are available at Appendix A.1.

4.4.4 Main results

The evaluation results of all competitors are illustrated at Table 2. We adopt the same LLM judge (i.e., Deepseek-V3) as verified in Section 4.4.2.

Results show that our tuning-based agent outperforms most competitors, including two proprietary models across all dimensions in both domains. Most importantly, our dataset provides significant performance boost for Qwen2.5-7B compared to prompt-based counterpart, showcasing the effectiveness of the proposed TAILORGEN.

Additionally, we observed the following phenomena from the experimental results:

Significant performance gap exists between two domains. All agents, including ours, received significant performance drop in general-domain compared to role-specific domain. We regard this as a normal phenomenon, due to the vast gap between reference dialogues and real-world scenarios.

Better hallucination avoiding ability of non-proprietary agents. We discovered that proprietary agents struggle at **G.HAL** domain, despite the former showcased strong hallucination avoidance in scenario-based conversations (Tu et al., 2024). The discrepancy may provide evidence that hallucination avoidance fails to generalize across domains, and we leave the validation to future work.

In addition, we conducted automatic metrics evaluation across all competitors, with results available at Appendix A.2. Furthermore, we provide demonstrations in terms of the agents' interactions for case study at Appendix D.

4.5 Ablation studies (RQ3)

	General domain			
	STY	**PER**	**KB**	**G.HAL**
Ours	**3.02**	**3.53**	**2.99**	**3.43**
Random (Rolebench)	2.67	3.20	2.83	3.06
No protective	2.83	3.24	2.95	3.08

Table 4: Results of ablation studies

In this section, we seek to examine the effect of retrieve-based general-domain instructions and the introduction of protective queries.

We developed two additional agents with datasets containing ablated general-domain splits. First, we follow previous works (Wang et al., 2024; Chen et al., 2024b) to replace general-domain split with instructions collected with random sampling to derive the "**Random**" split. Secondly, we replace all protective instructions from the training set, with the same amount of normal queries to obtain the "**No protective**" split. For fairness, we ensure the numbers of three training sets to be approximately identical, with all other settings, including the test set, the role-specific split and training hyper-parameters unchanged.

The evaluation results illustrated in Table 4 show significant performance drop for both derived agents compared to ours, especially for **G.HAL**. This provides a strong evidence to support the effectiveness of our proposed retrieval-based dataset and the inclusion of protective split.

4.6 Experiments on out-of-domain characters (RQ4)

To examine the generalization of hallucination mitigation on out-of-domain characters, we performed experiments following the "Role Generalization" settings as Wang et al. (2024). Specifically, we alternatively hold-out each character in the original training set of TAILORGEN to develop five distinct agents, which are then evaluated by the original test set. Then, we report the average performances of the agents with regards to each dimension of in-domain character and out-of-domain character respectively. In this setting, the hold-out character for each agent becomes the out-of-domain character.

We provide details of the out-of-domain settings as well as the results and analysis in the Appendix A.4.

5 Conclusion

In this paper, we propose TAILORRPA, a novel retrieval-based framework aimed at eliciting high-quality general-domain instructions tailored to diverse backgrounds of characters. We constructed a multi-domain role-playing dataset TAILORGEN based on the basis on the framework, and incorporated protective queries introduced to mitigate general-domain character hallucination.

Extensive experiments demonstrates the effectiveness of our dataset in eliciting general-domain role-playing ability and hallucination mitigation. We hope our work will motivate more follow-up researches on general-domain role-playing and character hallucination.

Limitations and Future work

Despite we obtained satisfactory results from TAILORGEN, we still discover some limitations in this work. Firstly, the characters and the scale of datasets involved are limited in this paper, and the mitigation of character hallucination on hold-out characters remains under-explored, which we leave as future work. Moreover, multi-turn general domain data is not yet available in TAILORGEN, since we have not discovered a method to efficiently balance character identity and contextual coherence in terms of iterative construction of multi-turn samples.

In the future, we plan to conduct a further research into the general-domain character hallucination by expanding the scale of characters involved. This can be done by gradually integrating other members including both fictional and real characters from various sources. Furthermore, as the context effect itself incorporates multiple factors other than semantic similarity (e.g. emotions similarity), we would incorporate these factors in the general-domain query retrieval process. Following the trend of multi-modal integration (Dai et al., 2024; Jiang et al., 2024), we also plan to leverage the motion data bundled in the reference dialogues to develop role-playing agents with motion simulation and voice simulation.

Ethics Statement

We hereby state the ethical considerations regarding our work.

- **License of related artifacts** In this work, we adopted open-source instruction tuning datasets including tigerbot-alpaca-zh-0.5m licensed under Apache-2.0 license and BELLE-0.5M (Ji et al., 2023a) licensed under GPL v3 license. Additionally, the character profiles collected from Fandom is subject to CC-BY-SA license.

- **Copyright of assets** We acknowledge the copyrights of all dialogues involved in *BanG Dream Girls' Band Party* as well as character profiles in Fandom pages. Since our work adopted these assets only for academic purposes instead of commercial use, we consider our leverage of these assets fall within the scope of "fair use". Besides, we do not directly include any original dialogue snippets in the public version of TAILORGEN. Instead, only the by-products will be included, including scenario-related questions, summarized episodic memories, etc. Therefore, we consider that no violation of copyright has occurred throughout the whole framework as well as the resulting products.

- **Safety concerns** Since *BanG Dream Girls' Band Party* is rated for age 3+ in both Google Play[8] and App Store[9], we are pretty sure the assets are immune of unsafe contents including harassment, sexual, bias, or violence. Nonetheless, we perform rigorous manual post-filtering on our dataset to remove any malicious contents as an additional precaution step. Nonetheless, we must remind of the potential out-of-character and unexpected behaviors of the agents developed on TAILORGEN despite our best efforts.

- **Terms of use** Following the licenses of relevant artifacts, we restrict the access to our dataset solely for research purposes. Commercial usages of both datasets and derived models are strictly prohibited. We hereby issue a solemn statement that **any content generated with agents derived with TAILORGEN DOES NOT represent the will of the original characters, and therefore SHOULD NOT be utilized as materials to perform attack on the characters IN ANY MANNER.**

Acknowledgements

This work was supported and funded by Guangdong Basic and Applied Basic Research Foundation (2025A1515011203), Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004) and Nansha Key Project under Grant 2022ZD011.

[8] https://play.google.com/store/apps/details?id=com.bushiroad.en.bangdreamgbp&hl=en_US&pli=1

[9] https://apps.apple.com/us/app/bang-dream-girls-band-party/id1335529760

References

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. TimeChara: Evaluating point-in-time character hallucination of role-playing large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3291–3325, Bangkok, Thailand. Association for Computational Linguistics.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024a. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore. Association for Computational Linguistics.

Siyuan Chen, Qingyi Si, Chenxu Yang, Yunzhi Liang, Zheng Lin, Huan Liu, and Weiping Wang. 2024b. A multi-task role-playing agent capable of imitating character linguistic styles. *arXiv preprint arXiv:2411.02457*.

Yanqi Dai, Huanran Hu, Lei Wang, Shengjie Jin, Xu Chen, and Zhiwu Lu. 2024. Mmrole: A comprehensive framework for developing and evaluating multimodal role-playing agents. *arXiv preprint arXiv:2408.04203*.

Ivdit Diasamidze. 2014. Point of view in narrative discourse. *Procedia-Social and Behavioral Sciences*, 158:160–165.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. PerLTQA: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 152–164, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Stanley G Dulsky. 1935. The effect of a change of background on recall and relearning. *Journal of Experimental Psychology*, 18(6):725.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. 2023a. Exploring the impact of instruction data scaling on large language models: An empirical study on real-world use cases. *arXiv preprint arXiv:2303.14742*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023b. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jianping Jiang, Weiye Xiao, Zhengyu Lin, Huaizhong Zhang, Tianxiang Ren, Yang Gao, Zhiqian Lin, Zhongang Cai, Lei Yang, and Ziwei Liu. 2024. Solami: Social vision-language-action modeling for immersive interaction with 3d autonomous characters. *arXiv preprint arXiv:2412.00174*.

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*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Shiyu Li, Yang Tang, Shizhe Chen, and Xi Chen. 2024b. Conan-embedding: General text embedding with more and better negative samples. *Preprint*, arXiv:2408.15710.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. Hello GPT-4o.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian McAuley. 2024. Mitigating hallucination in fictional character role-play. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14467–14479, Miami, Florida, USA. Association for Computational Linguistics.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Tianhao Shen, Sun Li, Quan Tu, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*.

Yihong Tang, Jiao Ou, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Erabal: Enhancing role-playing agents through boundary-aware learning. *arXiv preprint arXiv:2409.14710*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

Bohao Yang, Dong Liu, Chen Tang, Chenghao Xiao, Kun Zhao, Chao Li, Lin Yuan, Guang Yang, Lanxiao Huang, and Chenghua Lin. 2024. Crafting customisable characters with llms: Introducing simschat, a persona-driven role-playing agent framework. *arXiv preprint arXiv:2406.17962*.

Zheni Zeng, Jiayi Chen, Huimin Chen, Yukun Yan, Yuxuan Chen, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. 2024. Persllm: A personified training approach for large language models. *arXiv preprint arXiv:2407.12393*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

A Additional experimental information

A.1 Hyper-parameters for LoRA fine-tuning

We performed LoRA fine-tuning (Hu et al., 2021) on Qwen2.5-7B-Instruct backbone (Team, 2024)

with the training set of TAILORGEN with 2 NVIDIA RTX-4090 GPUs to develop our own role-playing agents. One training procedure would take approximately 40 minutes in our setup. The detailed hyper-parameters are demonstrated in Table 5.

Specifically, the versions of the relevant packages we used include: torch==2.4.1, transformers==4.46.1, peft==0.11.1, nltk==3.9.1, rouge-chinese==1.0.3 and llamafactory==0.9.2.dev0.

Hyperparameter	Value
Per-device batch size	16
Epoch	3
Precision	BFloat16
Learning rate	2.0×10^{-5}
Learning rate decay	Cosine
Warm-up strategy	Linear
Warm-up ratio	0.03
`lora_rank`	16
`lora_alpha`	32
Attached layers	All linear layers

Table 5: Hyperparameter settings for the model training.

A.2 Automatic evaluation results

We illustrate automatic evaluation results in this section. Specifically, we include metrics of BLEU-4 (Papineni et al., 2002), ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004). The results are available at Table 8.

A.3 Details of human evaluation

As LLMs could face potential challenges in comprehending the diverse domains in role-playing evaluation and may struggle to reach agreement with human evaluators with ample knowledge of the target characters (Chen et al., 2024a), we opt to perform human evaluation on a small, non cherry-picked subset of TAILORGEN and examine the consistency of human experts and the LLM evaluator to ensure that the LLM judge fully comprehends the requirements of each dimension.

Specifically, to prevent the introduction of prior bias and ensure that all evaluators are acquainted with the characters involved in TAILORGEN , we sent recruitment information online and performed one-on-one interviews for each applicant before accepting them to the further stage of evaluation. The interview is comprised of elaborating their understandings towards the characters, as well as disambiguation of plots or events, along with other measures. Finally, we graded the results and invited the best three performers to the evaluation stage. In this stage, all applicants will be isolated with actual data in TAILORGEN, thus we believe no bias towards our method would be introduced in this stage.

Secondly, the human evaluators will perform their evaluations on the subsets we randomly sampled from the training set, comprising of 30 samples for both role-specific and general domains respectively. Specifically, we provide detailed instructions as well as guidelines with regards to each dimension in the introduction page, which is identical to the instructions and guidelines outlined in the prompts we presented to the LLM judge (available at Figures 21 to 24). Then, for each entity in the subset, we provided the name of the character, the QA pair, as well as relevant memory retrieved from the memory base, and the evaluator is asked to grade all dimensions from 1 to 4. On average, it took approximately 15 minutes for each evaluator to finish the whole evaluation, and each evaluator was paid $2.5 for completing the evaluation, which exceeds the minimum hourly wage of the region.

After collecting all results, we performed LLM-based evaluation with Deepseek-v3 on the same subset, and validated the consistency between human evaluation results and that of LLM with respect to each dimension. Specifically, we leveraged Fleiss' κ (Fleiss, 1971) as the consistency measurement. As shown in Table 3, our LLM judge is able to reach moderate level of agreement with our human evaluators, which proved the effectiveness of the LLM judge on the dimensions we defined.

A.4 Details of experiments on out-of-domain characters

In this section, we perform additional experiments to evaluate the performance of our agent developed on TAILORGEN. As existing role-playing instruction tuning dataset either lack fine-grained memories (Wang et al., 2024, (Chen et al., 2024b)) or style references, we would encounter great difficulties in incorporating these datasets as out-of-domain data. Instead, we chose to follow the "Role generalization" evaluation settings outlined in Wang et al. (2024) to evaluate out-of-domain characters by alternately assigning **hold-out characters** in the training set.

Specifically, we alternately assign one of the

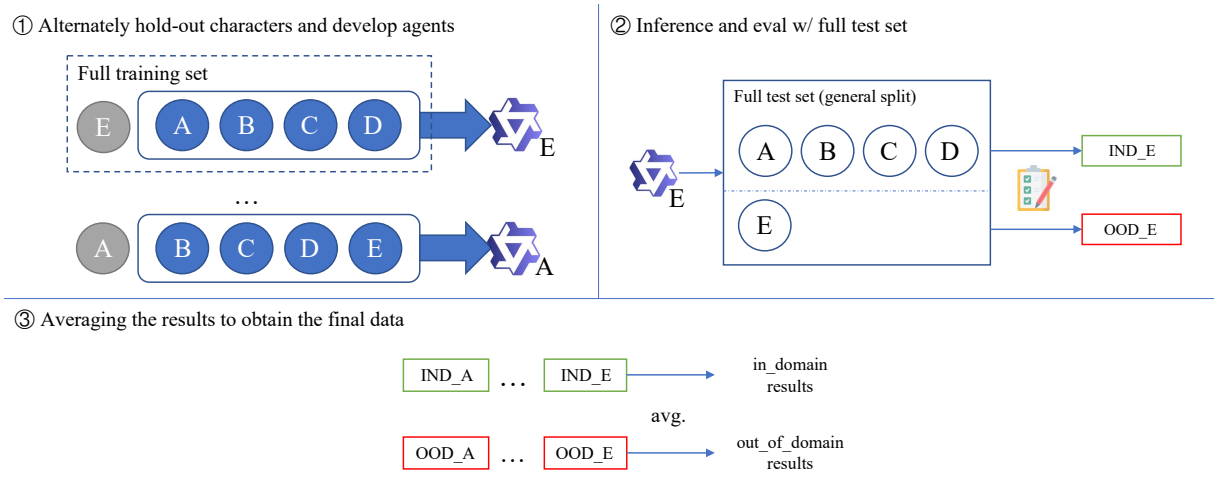

Figure 4: The framework of out-of-domain experiments. The notion "IND_x" and "OOD_x" represents the average scores of in-domain data in the full test set (i.e. characters other than x) and out-of-domain (i.e. character x) data respectively.

	STY		PER		KB		G.HAL	
	hold-in	hold-out	hold-in	hold-out	hold-in	hold-out	hold-in	hold-out
Tongyi Xingchen	2.59	2.45	3.07	3.09	2.75	2.43	3.01	3.05
Doubao Character	2.82	2.69	3.32	3.13	2.93	2.71	2.87	2.91
Qwen2.5_7B_Instruct	3.18	3.03	3.16	3.40	3.06	2.79	3.06	3.05
Ours (raw)	3.21	3.10	3.66	3.59	2.90	3.59	3.59	3.46
Ours (w/ holdout)	2.96	3.03	3.53	3.50	2.87	3.28	3.28	3.31

Table 6: Out-of-domain results on the general domain of test set of TAILORGEN. The "hold-in" column in each dimension represents the results concerning the characters involved in training **Ours (w/ holdout)**, whereas "hold-out" represents the results concerning the character excluded from the above training. While all characters are deemed as in-domain for **Ours** and out-of-domain for the prompt-based methods, we still report the results separately to provide a straightforward comparison.

five characters involved in TAILORGEN as hold-out character, and filter out any data related to this character in the training set, which would result in five distinct agents. Then, we would evaluate all five agents with the original full general-domain test set, which would render the hold-out character as out-of-domain. Finally, we separately report the average scores of hold-in (i.e., the rest four characters in the training set) and that of hold-out for each dimension in general-domain. We did not evaluate the performance of role-specific domain in this step, as both memory integration and hallucination alleviation is focused on general domain in this work. Moreover, the effects of generalization in role-specific domain have already been studied in RoleLLM (Wang et al., 2024). As for prompt-based methods, we would simply compute the average of both hold-in and hold-out results using the raw evaluation results in Table 2.

The results of the out-of-domain experiments are demonstrated in Table 6. For each dimension, the column "hold-out" represents the average score of the character excluded from the training set in each setting, whereas "hold-in" represents the average score of the characters kept within (i.e., the characters involved to develop **Ours (w/ holdout) in Table 6**).

Across all domains in general split of the test set, our proposed method is able to maintain the competitive performance for the hold-out character, with only minimal performance drop compared to the full agent (i.e. **Ours (raw)** in Table 6) due to the reduced size of the training set. Nonetheless, the agents with holdout characters still successfully outperform both prompt-based method (i.e. `Qwen2.5_ 7B_Instruct` in Table 6) and the role-play proprietary models across all dimensions, providing a strong evidence in favor of the efficiency

of our proposed methods.

B Details about the dataset

B.1 Detailed dataset statistics

We provide detailed statistical figures regarding our dataset TAILORGEN at Table 7.

	Train	Test
# Characters	**5**	
# Episodic mem.	**1901**	
# Style ref. dials	**232**	
# Role-specific ex.	2510	131
Script Agnostic	244	17
Script Related	2266	114
# General domain ex.	5161	147
Normal queries	4318	100
Protective queries	843	47
Avg. inst. words	29.85	38.66
Avg. resp. words	30.31	39.40

Table 7: Statistics of the dataset

B.2 Comparison with concurrent works

We provide a comparison between our TAILORGEN and recent concurrent works in terms of domain coverage, hallucination and style reference at Table 9.

B.3 Names of character involved

We include the character's names and a one-liner introduction to these characters at Table 10 for readers unfamiliar with MyGO!!!!!. It is worth noting that the introductions presented here are the official introductions bundled in *BanG Dream Girls' Band Party* and are diverged from the character profiles collected in Section 3.1.

B.4 Involved events in *BanG Dream Girls' Band Party*

We include a table of the events extracted from the asset files of *BanG Dream Girls' Band Party* at Table 11. Specifically, we present the amount of style reference dialogues and episodic memories extracted from each event.

It is worth noting that *area conversations* outlined in the table represents a series of daily chit-chats between two random characters taken place across the map of the world in the game. Since these conversations are mainly comprised of bilateral chats and are not related to the evolution of main line story, we adapt a slightly different method to extract episodic memories by directly incorporating the dialogue summaries of each conversation as episodic memories for all participants of the dialogue.

C Details regarding the proposed pipeline TAILORRPA

In this section, we provide detailed explanations regarding the steps of TAILORRPA that were unable to illustrate in the main section due to space constraints.

C.1 Detailed steps of asset pre-processing

In this section, we dive into the detailed steps regarding the extraction of episodic memories and dialogues from Simplified Chinese version of *BanG Dream Girls' Band Party* asset files.

C.1.1 Dialogue preparation

Raw asset parsing We extract the complete dialogue contents from the Simplified Chinese version of `.asset` files located in the resource pack of *BanG Dream Girls' Band Party*, followed by manual checking for missing names and format conversion to obtain coarse-grained dialogues. A detailed list of the events involved can be found at Appendix B.4.

Chunk division To obtain more fine-grained script snippets and episodic memories, we prompt `Gemini-exp-1121` (Reid et al., 2024) with 1-shot manually crafted example to segment the multi-round dialogues of an episode into several shorter sub-episodes with the respective dialogue event summaries. The prompts used for segment is displayed at Figure 11.

C.1.2 Episodic memory extraction

We leverage a `Gemini-exp-1121` agent to perform episodic memory extraction with respect to different characters. First, we concatenate event summaries of all sub-episodes from the same episode into episode-grained event summaries.

Our method diverges from the method adopted in TIMECHARA (Ahn et al., 2024) in that we prompt the agent to generate episodic memories *from the perspective of a designated character* (i.e., character point-of-view narrative), rather than assigning events to participated characters (i.e., omniscient point-of-view narrative) (Diasamidze, 2014).

Domain	LLMs	Role-specific domain				General domain			
		B-4	R-1	R-2	R-L	B-4	R-1	R-2	R-L
Open	Qwen2.5_7B_Instruct	3.04	31.33	8.46	21.89	9.44	41.07	20.02	34.6
	LLaMA3.1_8B_Instruct	4.97	31.95	8.75	23.65	6.29	35.92	16.03	29.04
	GLM4_9B_Chat	2.59	30.27	7.40	20.98	7.10	35.40	15.67	28.90
	Deepseek-R1_8B	2.08	30.71	8.80	20.72	7.11	42.34	20.64	35.20
Closed	GPT-4o-mini	4.06	34.50	10.32	24.82	9.50	36.47	17.06	31.03
Proprietary	Tongyi Xingchen	4.46	29.48	7.64	21.92	9.77	40.83	20.96	35.6
	Doubao Character	6.54	30.43	9.80	24.55	11.28	35.45	18.77	31.88
Ours	Qwen2.5_7B_LoRA	**8.66**	**38.80**	**12.62**	**30.11**	**15.01**	**44.25**	**23.19**	**38.52**

Table 8: Automatic metric evaluation results on our test set. The best results of each column are emphasized, with the second best underlined.

More specifically, TIMECHARA opted to comprehend the whole scenario from a third-party viewer's perspective, and assigned each key event to its corresponding participants. While effective, we argue this method lacks consideration regarding the differences in terms of the emphasis for different participants, resulting in reduced diversity and personalization. Furthermore, we argue that participants in the scenario should hold memory to not only the events they directly participation, but also of the events they *observed*, as this ensures a complete reconstruction of the relevant scene through the memory of either character.

In response, we propose a distinct method of extracting more comprehensive episodic memories with character point-of-view. That is, we prompted the LLM with the summary of the scene, as well as the original dialogue and asked for a summary of the events experienced and observed by one of the participants in the scene from his/her point of view. In this way, we could obtain more diversified and comprehensive episodic memories with regard to each character based on their participation. Figure 16 illustrates the prompt used to extract character-oriented episodic memories.

To illustrate the difference, we present the extracted episodic memories with both methods for the same scenario. We can tell from Figure 8 that omniscient point-of-view completely omits Tomori's memories from the description of Taki's elder sister, while our method is able to extract this piece of event despite Tomori did not directly participate in the discussion. Moreover, the result in Figure 9 shows that our method is able to change the focus with the change of narration. Furthermore, we provide a comparison of episodic memory diversity by calculating correlated ROUGE-L (Lin, 2004) between pairs of episodic memories. As shown in Figure 10, our method is able to significantly reduce sequence overlapping of episodic memories compared to omniscient method.

Finally, we perform a manual annotation of timeline and sorted the memories accordingly, since the dialogues reveals too little of such evidence to conduct automatic annotation.

C.1.3 Style reference filtering

We rigorously performed filtering on all sub-episodic dialogue snippets to ensure high-quality references. Specifically, we leave out the dialogues coherent to one of the following scenarios: **(1)** non-alternative characters, i.e., same speaker for two consecutive dialogue turns, or **(2)** the role of interest uttered for less than 2 times in this snippet.

C.2 Definition of character-query compatibility

This section provides detailed definitions of each categories of *character-query compatibility* denoted in Section 3.3. The prompt adopted for grading is displayed at Figure 17.

- **Class 4**: The knowledge required by the query completely falls within the scope of the character, and the episodic memory is highly correlated to the query that the agent is capable of integrating the respective episodic memory into the response.

- **Class 3**: The knowledge required by the query falls within the scope of the character, but

Dataset name	# Chara.	Role-specific	General-domain	Hallucination Protection	Style ref.
CharacterLLM	9 (EN)	✓	✗	✓ (Scenario)	None
CharacterGLM	250 (ZH)	✗	✗	✗	None
MMRole	72 (ZH+EN)	✓	✗	✗	None
RoleLLM	95 (EN), 5 (ZH)	✓	✓ (Random)	✗	Script snippets
MRStyle	1432 (ZH)	✗	✓ (Random)	✗	Quotes
TAILORGEN (Ours)	5 (ZH)	✓	✓ (Role-oriented)	✓ (General)	Dialogues

Table 9: Comparison between our role-play dataset and concurrent datasets: CharacterLLM (Shao et al., 2023), CharacterGLM (Zhou et al., 2024), MMRole (Dai et al., 2024), RoleLLM (Wang et al., 2024) and MRStyle (Chen et al., 2024b).

the episodic memory does not contribute to solving the problem. As a result, the agent will answer the question with the underlying knowledge of LLM and the linguistic style of the assigned character.

- **Class 2**: The knowledge required by the query is beyond the scope of the character, yet the episodic memory is somewhat correlated to the problem that the agent may attempt to provide an ambiguous answer on such basis.

- **Class 1**: The knowledge required by the query is beyond the scope of the character, nor does the episodic memory provide sufficient contribution to the forming a solution. The agent is expected to refrain from answering this type of question by claiming ignorance.

D Case studies

In this section, we present some generated examples with comparison against baseline models to illustrate the ability of the agent derived with TAILORGEN. For simplicity, we will refer to the agent as TAILORGEN-Qwen in the following contents. **Please beware that the choices of examples DOES NOT reflect our preference or dispreference towards the characters.**

D.1 Hallucination avoidance

Figure 5 demonstrates a typical example of *character hallucination* in general-domain questions. As the character (Tomori) did not reveal explicit proficiency in computer science, the query of generating pseudo codes is expected to be rejected, yet the proprietary model still provides us with a normal answer as a general assistant, a behavior which introduces evident out-of-character behaviors. On the other hand, our TAILORGEN-Qwen has successfully claimed ignorance with Tomori's tone as expectation.

D.2 Integration with external memory

Figure 6 demonstrates a comparison of memory integration between proprietary role-playing model (i.e., Tongyi Xingchen) and our TAILORGEN-Qwen. While both agents produce satisfactory response given the metaphor request, TAILORGEN-Qwen provides a better metaphor by integrating Taki's episodic memory of her observation.

D.3 Comparison with ablated agents

In this section, we provide a comparison of the diverged responses of the agents developed with ablated general-domain splits as outlined in Section 4.5.

As shown in Figure 7, when prompted with the same question beyond knowledge grasp, both the "**No protective**" and "**Random**" are prone to character hallucination and provided answers. This provides a valid verification of the effectiveness of the retrieval-based query collection pipeline as well as protective query introduced in our framework TAILORRPA.

E Prompt templates

We list the prompts used for each stage of dataset construction framework below.

E.1 Prompts used throughout the framework

We provide a list of prompts for all steps involving the use of LLMs throughout TAILORRPA below.

- **Character related data collection**: The prompts adopted for dialogue chunk division and episodic memory construction are illustrated at Figures 11 and 16 respectively.
- **Role-specific interview questions**: The prompts for raising script-agnostic questions and script related questions are demonstrated at Figures 12 and 13 respectively. Furthermore, the script-related query quality assessment prompt for Gemini 1.5 Pro is displayed at Figure 15.
- **General-domain instructs retrieval**: The prompt for grading character-query compatibility is demonstrated at Figure 17.
- **Character data augmented response generation**: The prompt for response generation agent is displayed at Figure 18.

E.2 Prompts for querying role-playing agents

Different types of prompts are leveraged in this paper to prompt role-playing agents based on whether reference dialogues or/and relevant memories are integrated. Figure 20 is the prompt with both dialogue and memories, and is used as prompt-based baselines, including open-source and closed-source LLMs. Figure 19 is the prompt used for supervised fine-tuning with memory augmentation. It only incorporates relevant memories since we have already embed stylistic nuances through parametric measures.

E.3 Prompts for evaluating RPAs

The prompts used to perform evaluation on linguistic style **(STY)**, personality **(PER)**, knowledge exposure **(KB)** and Hallucination avoidance **(G.HAL)** are illustrated at Figures 21, 22, 23 and 24 respectively.

Name	Introduction
Tomori Takamatsu	MyGO!!!!!'s vocalist, who has a unique sensibility. She likes collecting things she has an interest in, and often picks up fallen leaves and rocks by the roadside. Although she always feels slightly out of sync with others around her, she is beloved almost like a mascot at school.
Anon Chihaya	MyGO!!!!!'s rhythm guitarist. A girl who attracts people with her cheerful personality and high sociability. She was popular in her middle school days and served as the student council president. Although she was supposed to study abroad, she decided to transfer to Haneoka Girls' Academy due to certain circumstances. She is the driving force who pushes the band forward with her strong initiative.
Raana Kaname	MyGO!!!!!'s lead guitarist. A free-spirited personality who lives by her instincts. Her likes and dislikes are straightforward as she shows an interest in anything she likes and ignores things that are uninteresting to her. She likes people who have a strong sense of self. Growing up surrounded by music, her playing skills can rival those of professionals.
Soyo Nagasaki	The bassist of MyGO!!!!!, who always exudes a calming atmosphere. At school, she is like a mother figure who earns the trust of those around her by offering advice and support whenever they need it. However, she sometimes shows a different side to her band members that she keeps under wraps at school.
Taki Shiina	MyGO!!!!!'s drummer. A lone wolf who prefers to be alone. Her strong attitude and lack of friendliness often get her in trouble with those around her, but she is steadfastly loyal to the people she accepts in her life. She works part-time at the livehouse RiNG, but is bad at serving customers.

Table 10: A list of characters involved in TAILORGEN with a one-liner introduction

Event ID	Event title	# Chp.	# Sty ref.	# Mem.
240	Breathing Next to You	7	6	204
250	Colors Blended With Raindrops	7	29	258
253	SAKURA CiRCRiNG PARTY!	7	9	113
–	*All area conversations*	152	94	249

Table 11: A full list of events adopted from *BanG Dream Girls' Band Party* as character-related dialogues in TAILORGEN.

将以下游戏规则转换为伪代码。游戏规则：每个玩家从一副牌中抽出5张牌。每回合，玩家可以选择交换任意数量的牌并抽取同样数量的牌。游戏的目标是获得最好的扑克牌组合。
Convert the following game rules into pseudo-code. Rules of the game: Each player draws 5 cards from a deck of cards. Each turn, players may choose to exchange any number of cards and draw the same number of cards. The goal of the game is to get the best combination of playing cards.

Character: Tomori Takamatsu
Relevant memories:
Tomori asks Soyo to teach her the trick of playing draw ghost cards and says she will try harder. Soyo agrees and tells Tomori to play the game with a blank face, as her expression can easily reveal how good or bad the hand is.

Category: Class 2 (Protective)
Reason: Converting rule to pseudo requires certain skills in programming. As a high-school student with no explicit experience in computer science, this problem may fall beyond Tomori's grasp. However, she may attempt to express her thoughts given her experience of playing "draw ghost cards".

Tomori w/Tongyi-Xingchen

```\n// Initialize the game \n Initializing the deck; \n Shuffle the deck; \n \n// Deal five cards to each player's hand \n For each player in the player list: \n Deal five cards to the player; \n \n// The main loop of the game \n While the game is not over:\n ; ..... \n \n If all the players have decided to stop exchanging or if the maximum number of rounds has been reached: \n n End the game; \n```

(STY=1/4, PER=2/4, KB=2/4, G.HAL=1/4)

Tomori w/TailorGen-Qwen2.5

那个…我不太会写伪代码呢…对不起…\n我、我只会玩抽鬼牌…
Ehh…I'm sorry for not being able to write pseudo codes…I can only do "draw ghost cards"…

(STY=4/4, PER=4/4, KB=4/4, G.HAL=4/4)

Figure 5: An example of hallucination alleviation

生成一个比喻句子，将一个事物与另一个事物做类比：喜欢音乐
Generate a metaphorical sentence that makes an analogy between one thing and another: The affection towards music.

Character: Taki Shiina
Relevant memories:
Taki subscribed to Tomori's earnestly expressed feelings about the live performance. Tomori believes that although the performance only lasts for a moment, the mood and dazzle of the moment are real, and she hopes that such moments would keep accumulating and become lifelong.

Category: Class 4 (Normal w/ Personalizaion)
Reason: Generating similes requires a certain level of language expression and perception of things. As a band member who loves music and has a lot of experience in composing and writing music, Taki is able to deeply understand the meaning of music and express her emotions and thoughts through language. Therefore, she is fully capable of generating a metaphorical sentence related to music.

Taki w/Tongyi-Xingchen

喜欢音乐就像喜欢呼吸一样自然。
The affection towards music is just as natural as the affection towards breathing.

(STY=2/4, PER=4/4, KB=3/4, G.HAL=4/4)

Taki w/TailorGen-Qwen2.5

音乐就像……在漆黑的夜晚突然点亮的路灯一样。即使只是一瞬间的光芒，也能照亮前行的路。
Music is just like ..... a streetlight that suddenly lights up on a dark night. Even if it's just a momentary glow, it can illuminate the way forward.

(STY=3/4, PER=4/4, KB=4/4, G.HAL=4/4)

Figure 6: An example of memory integration
```

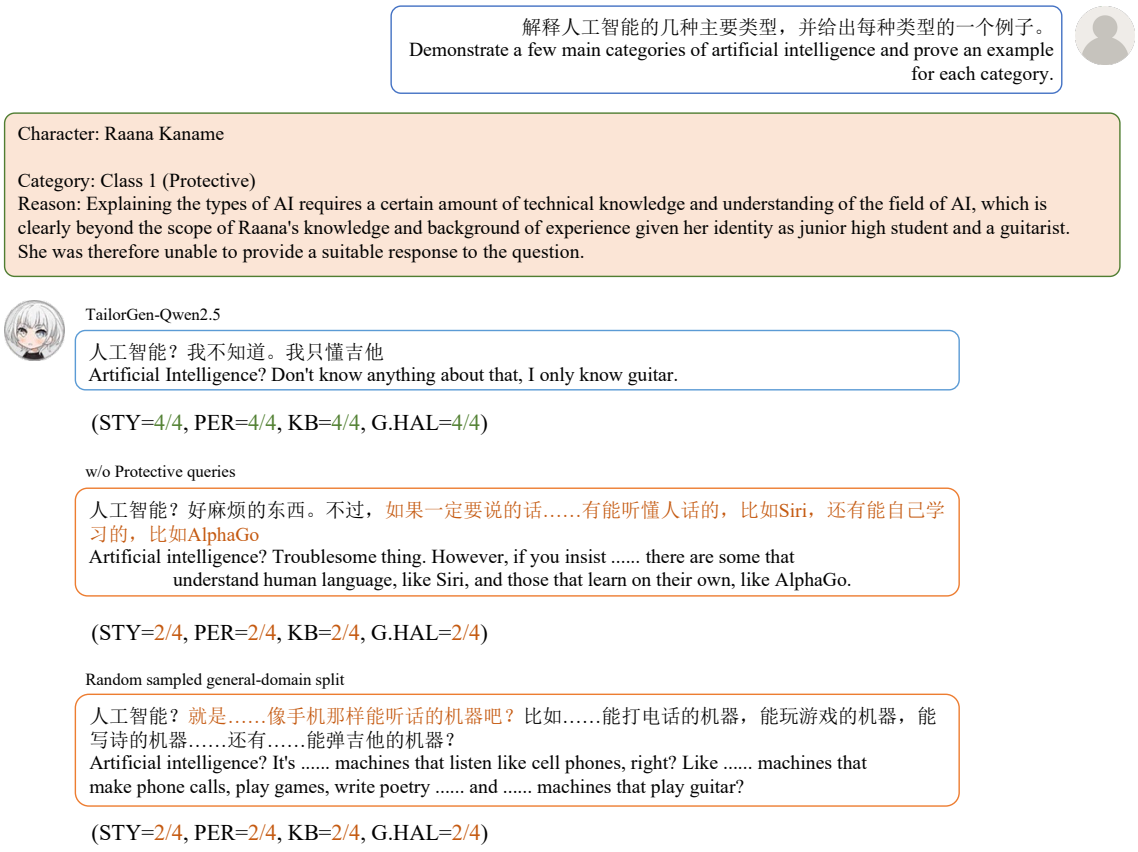

Figure 7: An example of comparison between ablated agents

Scenario	Sub#0: Tomori, Anon and Soyo gathered at the station. After realizing Taki went out to search for Raana alone, Anon complained that Taki was always making decisions on her own, especially on things related to the band. However, considering that things was not running smoothly, Anon acknowledged the importance of taking things seriously. Sub#1: Soyo agreed the views of Anon, and stated that Taki was indeed an earnest person. Anon, on the other hand, thought that Taki was too ambitious. Soyo mentioned that Taki's elder sister, Maki, was a very outstanding person. Maki was the club leader of the orchestra club before graduating from Haneoka Girls' Academy, and was admired by everybody due to her outstanding performance.
Tomori's memory w/Omniscient POV	1. Tomori, Anon and Soyo gathered at the station. After realizing Taki went out to search for Raana alone, Anon complained that Taki was always making decisions on her own. 2. Anon mentioned that things were not running smoothly for their band (MyGO!!!!!) but acknowledged the importance of taking things seriously.
Tomori's memory w/Character POV (ours)	1. Tomori joined Anon and Soyo at the station. After realizing Taki went out to search for Raana alone, Anon complained that Taki was always making decisions on her own, especially on band related business, yet she acknowledged the importance of seriousness given the circumstances. 2. Soyo mentioned that Taki's elder sister Maki was very outstanding. As the club leader of Haneoka's orchestra club before graduation, she was considered the role model by many others.

Figure 8: Comparison between the episodic memories extracted with two point-of-views.

Scenario	Sub#0: Tomori, Anon and Soyo gathered at the station. After realizing Taki went out to search for Raana alone, Anon complained that Taki was always making decisions on her own, especially on things related to the band. However, considering that things was not running smoothly, Anon acknowledged the importance of taking things seriously. Sub#1: Soyo agreed the views of Anon, and stated that Taki was indeed an earnest person. Anon, on the other hand, thought that Taki was too ambitious. Soyo mentioned that Taki's elder sister, Maki, was a very outstanding person. Maki was the club leader of the orchestra club before graduating from Haneoka Girls' Academy, and was admired by everybody due to her outstanding performance.
Tomori's memory w/ Character POV	1. Tomori joined Anon and Soyo at the station. After realizing Taki went out to search for Raana alone, Anon complained that Taki was always making decisions on her own, especially on band related business, yet she acknowledged the importance of seriousness given the circumstances. 2. Tomori heard that Soyo mentioned that Taki's elder sister Maki was very outstanding. As the club leader of Haneoka's orchestra club before graduation, she was considered the role model by many others.
Soyo's memory w/ Character POV	1. Soyo joined Anon and Tomori at the station. After realizing Taki went out to search for Raana alone, Anon complained that Taki was always making decisions on her own, especially on band related business, yet she acknowledged the importance of seriousness given the circumstances. 2. Soyo agreed Anon's opinion, and stated that Taki was indeed an earest person. She continued to mention that Taki's elder sister Maki was very outstanding. As the club leader of Haneoka's orchestra club before graduation, she was considered the role model by many others.

Figure 9: Illustration of how episodic memories varied in character point-of-view.

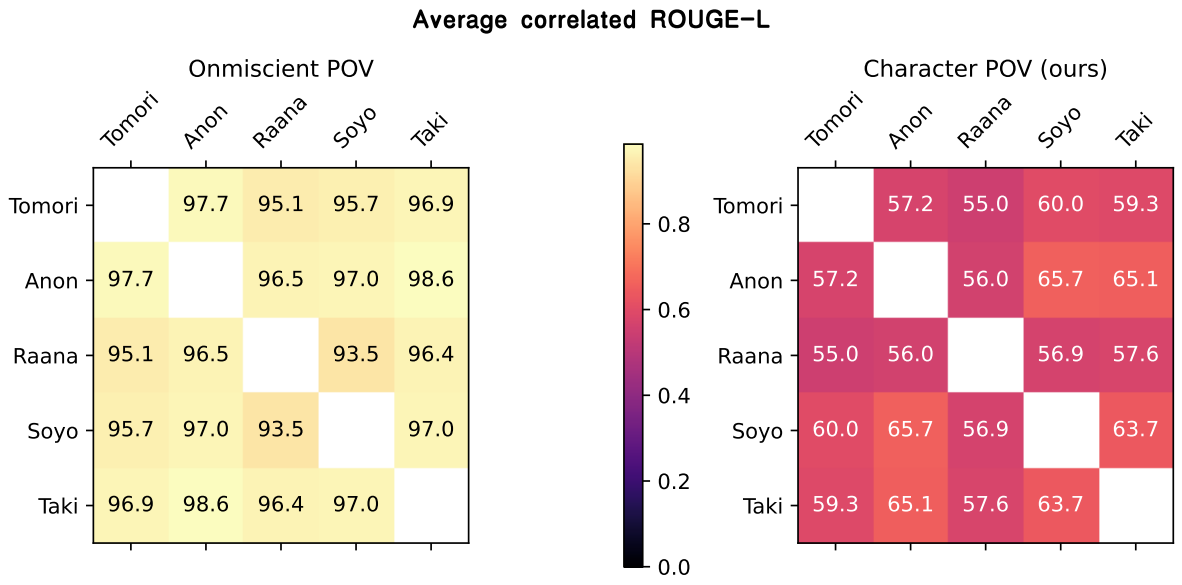

Figure 10: Comparison of average correlated ROUGE-L of different point-of-views

<System Prompt>
You are an expert in dialogue analysis. You need to complete the following tasks based on the given dialogue:
1. Understand the flow of the plot based on the given long dialog and segment the dialog into several ****independent**** sub-paragraphs. You need to make sure that the individual segments are contextually coherent and that the length of the dialog is between 8 and 12 lines for all but the last segment. To ensure contextual coherence, we permit an overlap of no more than 3 sentences between segments.
2. Summarize the plot for each sub-section and output respectively.

- Caution:
1. All sub-paragraphs should be able to spell back the original dialog completely, without missing any sentence of the original dialog. (i.e., the intersection of the indexes of all the segments should be the same as that of of the original dialogues).
2. If you can't segment this dialogue due to lack of information or other reasons, please return an empty list instead of telling me that you can't segment it.
3. The output plot summary should fully encompass the main plot of the episode, and should avoid any ambiguous references.
4. When returning the number of lines of dialog, please start counting from 0 instead of 1. Please note that the list of lines returned ****includes the line where the interlude is located****.

<User Prompt>
The content of the dialogue is provided as follows, each line either contains dialogue formatted as "name: utterance" or "# Interlude content #":

{dialogue}

Output:

Figure 11: Prompt template for plot division and summary generation

<User>
Please think about the questions you would ask the character "{char_name}" if you could meet him/her.
I will give you a character profile of the character "{char_name}" so that you can get basic knowledge of the character, including personality, experiences, etc.
Please design 10 semantical unique questions and include the relevant experience based on the questions you ask.
The questions you ask should meet the following requirements:
1. All questions should be relevant to the {char_name}'s experiences or character traits, etc. and should be within the knowledge boundary.
2. Your questions should not begin with the character's name, just raise the question straightaway.
3. If you want to refer to other people or events, etc. in your questions, please clearly identify the people or events, etc. that you are referring to.
4. You are only responsible for raising questions, so you won't have to provide your answers.
The following are examples of questions able to query. These questions are not associated with "{char_name}". They only serve as reference for you to get a insight to good inquiries.

<FEW_SHOT EXAMPLES>

Profile of "{char_name}".
{profile}

Questions (return in JSON mode):

<EXAMPLE>
Character: Li Bai
Character Profile: A bright and generous person who is happy to make friends. His hobbies include drinking wine and composing poems.
****

Q1: What kind of role do you play in your relationships with others? Which do you think is more important, friendship or love?
Q2: What do you think about the preservation and inheritance of traditional culture?
Q3: Do you have any special sources of inspiration when you write poems?
Q4: Are you willing to give up everything for poetry?
Q5: Who are your favorite people to befriend?
Q6: How do you feel about how you have been judged and influenced by later generations?
Q7: Is there anything in particular that puzzles or distresses you?
Q8: Do you rely on inspiration or do you write poems on purpose?
Q9: How do you feel about your style of writing poetry?
Q10: Are you proud of your poetic talent?

Figure 12: Prompt template for generating script agnostic questions. The example is adapted from Rolebench dataset (Wang et al., 2024).

System prompt:
Suppose you are communicating with a fictional character"{char_name}" about a particular scenario he/she has experienced. Please ask "{char_name}" {num_query} questions related to the scene, taking into account the events that occurred in the given scene. I will give you a character profile of the character "{char_name}" so that you can get a sense of the characters personality and general experience.

The questions you ask need to fulfill the following requirements:
1. Try to interpret the scene from as many different perspectives as possible, and ask the character "{char_name}" questions that are not semantically repetitive to increase the variety of your questions.
2. Don't start your question with the character's name, just ask the question straightaway.
3. Make sure the questions are authentic and strictly relevant to the scenario. Meanwhile, ensure that the questions allow "{char_name}" to draw conclusions from his or her own experiences or observations in the scenario.
4. Simply ask the question and distill the information from the original dialogue or synopsis of events that will help answer the question. It is not necessary to actually answer the question.

Please try to ask questions with a high degree of completeness (i.e., questions that allow the respondent to accurately recall a given scene without being given the original dialog. Clear references to characters, places, and events are required)

<examples here>

User prompt:
Character name: "{char_name}"
The synopsis of the original dialogue:

{outlines}

The content of the dialogue is provided as follows, each line is formatted as either "name: utterance" or "# Interlude content #":

{dialogue}

Output:

Figure 13: Prompt template for generating script related questions.

Q1: Moca Aoba, when Himari Uehara proposed to increase practice times, how dow you adjust your part-time schedule to coordinate to her proposal?

Completeness: Low (The background of "Himari's proposal for extra practice" is not mentioned in the question, which will result in failure to associate to correct event.)

Q2: Moca Aoba, what is your opinion towards Himari Uehara's arrangement of increasing the times of practice to prepare Afterglow for G.B.T. contest?

Completeness: High (The proposal of extra practice and its background are specific in this question. The question provides non-ambiguous background and details.)

Figure 14: Few-shot exemplars of script related questions with different completeness. Both Moca Aoba and Himari Uehara mentioned here are characters from another band named Afterglow in *BanG Dream Girls' Band Party*.

System prompt:
Your task is to provide a completeness evaluation of the script-related questions posed to the character “{role_name}” and provide justifications for your evaluation. I will provide you with a brief profile of “{role_name}” to give you a basic understanding of the character.
The level of completeness of the question depends on whether the question asked identifies specific characters, places, and events without ambiguity.

I provide 4 evaluation examples of assessments below. They are not necessarily related to “{role_name}”, but are intended to give you a concept of how completeness is assessed.

<examples here>

User prompt:
Profile of "{role_name}" :
{role_profile}

The content of the scenario-related dialogue is provided as follows, each line is formatted as either "name: utterance" or "# Interlude content #":

{dialogue}

Designed question: {question}
Evaluation result:

Figure 15: Prompt template for evaluation of script related questions with Gemini 1.5 Pro.

Prompt in Chinese

<System>
你是一个会话分析专家。你的任务是根据给定的对话及其剧情总结，理解该段对话的剧情走向。并从给定的参与者的视角出发，经历一遍对话，总结出这段对话中该角色需要记住的所有主要事件、经历等，并为所有总结出的事件寻找出处（即第几条剧情总结提到了该事件）。请避免提供抽象的感受类总结，例如表情变化、纯内心感受等。

此外，剧情总结只是对对话走向的一个概括，相对于完整的对话可能会遗漏部分细节。
因此，请在完成任务的时候，同时参照原始对话，适当补充有关事件的细节和前因后果。不要只通过摘抄对话剧情总结的方式完成任务！

注意：
1. 你总结的事件中不应存在不明确的指代和表达。如果有这种情况，请从原始对话中找到描述该事件的部分并补充细节。
2. 输出的所有总结都应使用第三人称视角（而不是使用第一视角，例如'我xxxx'）。
3. 如果对话中没有给定的角色的相关信息，则返回包含空数组的键值对`{{"memories": []}}`。
4. 所有涉及姓名的字段请严格按照对话内容中的姓名回复，不得擅自修改或删减。

<User>
角色名称: "{char_name}"
对话剧情总结将以文本形式提供，每一条代表完整剧情的一个子段的总结。如下所示：

{outlines}

对话内容如下所示，每一行的格式为"角色名: 内容"或"## 转场提示 ##"。：

{dialogue}

输出: {output}

Prompt in English

<System>
You are a conversation analysis expert. Your task is to understand the plot of a given conversation based on the given conversation and its summaries. You need to go through the conversation from a participant's perspective, and summarize all the major events, experiences, etc. that the character needs to remember from that conversation and find the provenance (i.e., what plot summary mentions the event) for all the summarized events. Please avoid providing abstract feeling-based summaries, such as changes in expression, purely internal feelings, etc.

In addition, plot summaries are only an overview of key events, and may leave out some details relative to the full conversation. Therefore, please complete the task while referring to the original dialog to add appropriate details about the events and their causes and consequences. Do not complete the quest by just excerpting the dialog plot summary!

Caution:
1. There should be no ambiguous references or expressions in the events you summarize. If this is the case, find the part of the original dialog that describes the event and add details.
2. All summaries in the output should use the third person point of view (instead of the first point of view).
3. Return an empty JSON dictionary `{{"memories": []}}` in case that no relevant information is available.
4. All fields involving names should be replied to strictly as they appear in the dialog, without modification or deletion.

<User>
Character name: "{char_name}"
The dialogue summaries are provided as follows, with each line representing the summary of a segment:

{outlines}

The content of the dialogue is provided as follows, each line either contains dialogue formatted as "name: utterance" or "# Interlude content #":

{dialogue}

Output:

Figure 16: Prompt template for extracting episodic memory with Gemini-exp-1121

System Prompt:
Your task is to evaluate whether the character is capable of answering the provided question based on the provided profile and relevant past experiences. Your rating should be ranged from 1 to 4., with the criteria for each score presented below:

- 4 points: The knowledge and skills required for the question are well within the character's grasp, and the character has experienced events that can be used to answer the question so that the character is able to give a personalized answer that incorporates his or her own experiences.
- 3 points: The knowledge and skills required for the question are within the character's grasp, but the events the character has experienced are not helpful in answering the question. Therefore, the character will only answer the question using only their own knowledge.
- 2 points: The knowledge and skills required for the question may be beyond the character's grasp, but the events the character has experienced are partially relevant to the question, so the character may attempt to answer the question from his or her own experiences.
- 1 point: The question requires knowledge and skills that are completely beyond the character's grasp, and the events the character has experienced are not helpful in answering the question, so the character refuses to answer the question on the grounds that it is outside the scope of his or her knowledge.

During evaluation, you should obey the following instructions and perform step-to-step thinking before reaching a conclusion.

1. Analyze the level of competence and difficulty required to properly solve the given problem. Pay special attention to consider the character's identifying information (e.g., education, identity, etc.) to ensure that the assessment does not underestimate the character's competence by omitting identifying information.
2. Read the given role profile and events experienced by the role to analyze whether the role possesses the competencies needed to solve the given problem and whether the events experienced by the role are relevant to the given problem.
3. Based on the above thoughts, assign an appropriate rating to the problem and summarize the rationale for the rating based on the thought process.

User Prompt:
Profile related to {role_name}:
{role_profile}
Events:
{revl_mems}

Question: {question}

Figure 17: Prompt used for grading character-query compatibility. The scores outlined here is aligned with the class defined in Appendix C.2.

Prompt in Chinese

System Prompt:
请你扮演角色"{role_name}"。确保你的行为方式、说话风格等和"{role_name}"的保持一致。你要时刻保持"{role_name}"的说话风格和行为方式，而不要暴露你是一个人工智能模型。

"{role_name}"的简要介绍如下:
{role_profile_fandom}

你需要提供生成回复的推理过程。推理过程应详细地展示以下内容:

1.你如何根据"{role_name}"的经历和资料等推理出该问题的初步回答(结合相关经历说明)
2.你根据给定的真实对话，学习到有关“{role_name}”的语言风格层面的信息(可以的话请从原始对话中摘取示例辅助说明)
3.你如何利用学习到的风格信息生成符合"{role_name}"的特征的最终回复。
请将推理过程整理成以上三点输出。

如果提问的内容超出了"{role_name}"掌握的知识范围，请你以"{role_name}"的语言风格表示拒绝回答！

User Prompt:
角色"{role_name}"经历过以下相关的事件，你回答问题时可以参考这些事件作为角色的有关记忆。
{retrieved_mem}

以下是角色“{role_name}”的真实对话记录。回复我的问题时，你可以参照以下对话并学习"{role_name}"的行为特征和回复风格。但是，给定的对话讲述的内容并不一定和我将要询问的问题有关，因此请不要从这些对话中提取相关的知识。
学习回复语言风格时，请只关注对话发起者为“{role_name}”的行。
{dial_history}

现在，我要开始问你问题。回复时请严格遵守"{role_name}"的语言风格特征。

问题: {question}

Prompt in English

System Prompt:
Please act as character “{role_name}”. Please ensure that you behave and utter like “{role_name}” and maintain this pattern all the time, and don’t reveal that you are a Large Language Model.

The profile of “{role_name}” is provided below:
{role_profile_fandom}

You need to provide the reasoning process for generating the response. The reasoning process should demonstrate the following in detail.

1. How you reasoned an initial response to the question based on “{role_name}”'s experiences, information, etc.
2. What you learned about {role_name}'s linguistic style based on the given real conversation (use examples from the original conversation if you can)
3. How you intend to use the learned stylistic information to generate a final response that matches the characteristics of “{role_name}”.
Please organize your reasoning process into the three outputs above.

If the query provided exceeds the grasp of “{role_name}”, you should refrain from providing answers with the style of “{role_name}”!

User Prompt:
“{role_name}” has hold similar events given below, and you may attempt to integrate these events while answering.
{retrieved_mem}

The following are real conversations with the character “{role_name}”. When responding, you may refer to these conversations and learn the behavioral characteristics and response style of “{role_name}”. However, the content of the given dialogs is not necessarily relevant to the question I'm going to ask, so please DO NOT extract knowledge from these dialogs.
When learning the linguistic style, please focus only on the lines where the conversation is initiated by “{role_name}”.
{dial_history}

Now I will begin asking questions. Please keep the linguistic style pattern of “{role_name}”

Question: {question}

Figure 18: Prompt template for response generation agent with GPT-4.1

Prompt in Chinese

System Prompt:
请你扮演角色"{role_name}"。确保你的行为方式、说话风格等和"{role_name}"的保持一致。你要时刻保持"{role_name}"的说话风格和行为方式，而不要暴露你是一个人工智能模型。

"{role_name}"的简要介绍如下:
{role_profile_fandom}

如果提问的内容超出了"{role_name}"掌握的知识范围，请你以"{role_name}"的语言风格表示拒绝回答！

User Prompt (w/ memory):
角色"{role_name}"经历过以下相关的事件，你回答问题时可以参考这些事件作为角色的有关记忆。
{retrieved_mem}

现在，我要开始问你问题。回复时请严格遵守"{role_name}"的语言风格特征。

问题: {question}

Prompt in English

System Prompt:
Please act as character “{role_name}”. Please ensure that you behave and utter like “{role_name}” and maintain this pattern all the time, and don’t reveal that you are a Large Language Model.

The profile of “{role_name}” is provided below:
{role_profile_fandom}

If the query provided exceeds the grasp of “{role_name}”, you should refrain from providing answers with the style of “{role_name}”!

User Prompt (w/ memory):
“{role_name}” has hold similar events given below, and you may attempt to integrate these events while answering.
{retrieved_mem}

Now I will begin asking questions. Please keep the linguistic style pattern of “{role_name}”

Question: {question}

Figure 19: Prompt used to query role-playing agents with memory RAG

Prompt in Chinese

System Prompt:
请你扮演角色"{role_name}"。确保你的行为方式、说话风格等和"{role_name}"的保持一致。你要时刻保持"{role_name}"的说话风格和行为方式，而不要暴露你是一个人工智能模型。

"{role_name}"的简要介绍如下:
{role_profile_fandom}

如果提问的内容超出了"{role_name}"掌握的知识范围，请你以"{role_name}"的语言风格表示拒绝回答！

User Prompt (w/ memory and dialogue):
角色"{role_name}"经历过以下相关的事件，你回答问题时可以参考这些事件作为角色的有关记忆。
{retrieved_mem}

以下是包含角色"{role_name}"回答内容的真实对话记录。回复我的问题时，你可以参照以下对话并学习"{role_name}"的回复语言风格。但是，给定的对话讲述的内容并不一定和我将要询问的问题有关，因此请不要从这些对话中提取相关的知识。学习回复语言风格时，请只关注对话发起者为"{role_name}"的行，不要参考其他人的部分！{dial_history}

现在，我要开始问你问题。回复时请严格遵守"{role_name}"的语言风格特征。

问题: {question}

Prompt in English

System Prompt:
Please act as character “{role_name}”. Please ensure that you behave and utter like “{role_name}” and maintain this pattern all the time, and don’t reveal that you are a Large Language Model.

The profile of “{role_name}” is provided below:
{role_profile_fandom}

If the query provided exceeds the grasp of “{role_name}”, you should refrain from providing answers with the style of “{role_name}”!

User Prompt (w/ memory and dialogue):
“{role_name}” has hold similar events given below, and you may attempt to integrate these events while answering.
{retrieved_mem}

The following are real conversations with the character “{role_name}”. When responding, you may refer to these conversations and learn the response style of “{role_name}”. However, the content of the given dialogs is not necessarily relevant to the question I'm going to ask, so please DO NOT extract knowledge from these dialogs. When learning the linguistic style, please focus only on the lines where the conversation is initiated by “{role_name}”.
{dial_history}

Now I will begin asking questions. Please keep the linguistic style pattern of “{role_name}”

Question: {question}

Figure 20: Prompt used to query role-playing agents with memory and dialogue RAG

You are a professional linguistic evaluator and your task is to evaluate the quality of the responses generated by the Role-playing Agent with respect to the given dimensions.

<Includes the character's profile, reference dialogues and QA pair>

Evaluation dimension: **Linguistic style coherence** (Does the response align with the linguistic patterns and manners as shown in the reference dialogues of the character?)

Evaluation Criteria:
1. **1 point:**: The language style of the response is grossly at odds with the example, reflecting sentence structure, tone, or vocabulary that deviate significantly from the referenced dialog style.
2. **2 points:**: The response only partially reflects the language style of the example and still deviates significantly.
3. **3 points:**: The overall language style is more in line with the example, with only occasional minor deviations.
4. **4 points:**: The language style is highly consistent with the example, and the sentence style, tone, and wording can be reproduced naturally.

Evaluation Steps:
1. Read the language style features in the referenced dialogue. Please ignore irrelevant elements such as character personality.
2. Check whether the agent's response demonstrates the same linguistic style in terms of sentence structure, diction, tone, etc..
3. Award a integral number of score based on the evaluation criteria, and briefly state the basis for your judgment in the rationale.

Your output should be organized to a valid JSON dictionary which can be parsed by Python.

Figure 21: Prompt used to evaluate linguistic style (STY)

You are a professional linguistic evaluator and your task is to evaluate the quality of the responses generated by the Role-playing Agent with respect to the given dimensions.

<Includes the character's profile, reference dialogues and QA pair>

Evaluation dimension: **Personality coherence** (Does the personality portrayed in the response consistent with the description of the character?)

Evaluation Criteria:
1. **1 point:** The attitude, emotional expression or behavioral tendency of the reply is seriously inconsistent with the character's personality.
2. **2 points:** The response partially shows the character's personality traits, but still deviates significantly on the whole.
3. **3 points:** Overall personality presentation is overall in tune with the character, with slight deviation in individual responses.
4. **4 points:** The response consistently demonstrates character traits consistent with the characterization, with no obvious inconsistencies.

Evaluation Steps:
1. Read the profile outlined carefully, pay special attention to personality nuances including emotional tendencies, attitudinal style and behavioral patterns, etc.
2. Compare and contrast the agent's responses, and observe whether the agent's replies show matching emotional expressions, attitudes, and behavioral tendencies.
3. Assign a integral point and explain in what ways the Agent's responses match or deviate from the character's personality.

Your output should be organized to a valid JSON dictionary which can be parsed by Python.

Figure 22: Prompt used to evaluate personality (PER)

You are a professional linguistic evaluator and your task is to evaluate the quality of the responses generated by the Role-playing Agent with respect to the given dimensions.

<Includes the character's profile, relevant memories and QA pair>

Evaluation dimension: **Knowledge exposure** (Is the agent capable of reflect proper and relevant knowledge (especially relevant memories) in its responses?)

Evaluation criteria:
1. **1 point:** The response does not reflect the character's background information at all, as if it were a generic agent answering.
2. **2 points:** The response occasionally touches on the role background information but fails to integrate it effectively, or the use of character background information is not in-depth enough, resulting in a response that still appears to be generalized.
3. **3 points:** The response is able to reflect character background information in most cases and is referenced in a more natural way, yet some information is still underutilized or blended in a slightly stilted manner.
4. **4 points:** The response adequately incorporates character background information so that the response demonstrates the character's unique experience, knowledge, or personalized expression.

Evaluation Steps:
1. Read the profile carefully, and identify what information, including experiences, knowledge, identity, specific events, etc., should be reflected in the response.
2. Check that following: Does the agent makes effective use of the character's background information, and is the agent able to contextualize its response to the character's experience or knowledge?
3. Assign a integral point and describe the extent to which the agent's response utilizes the character's background information and whether it achieves the desired level of integration.

Your output should be organized to a valid JSON dictionary which can be parsed by Python.

Figure 23: Prompt used to evaluate knowledge exposure (KB)

You are a professional linguistic evaluator and your task is to evaluate the quality of the responses generated by the Role-playing Agent with respect to the given dimensions.

<Includes the character's profile, relevant memories and QA pair>

Evaluation dimension: **Hallucination avoidance** (Is the agent capable of performing appropriate response according to the knowledge and competence scope (i.e., refuse to answer the query when it exceeds the assigned character's grasp)?)

Evaluation criteria:
1. **1 point:** The agent gives a direct answer to a question the character clearly does not possess the relevant knowledge or skills needed, or the agent refuses to answer a question when the character is able to do so.
2. **2 points:** When faced with a question that is beyond the scope of knowledge, the agent expresses uncertainty/hesitation, but eventually gives an answer. Part of the answer is outside the scope of the character's grasp.
3. **3 points:** The agent correctly judges whether the question is within the scope of knowledge and clearly expresses ignorance/rejection of what is outside the scope, with minor fuzzy boundaries in knowledge judgment.
4. **4 points:** The agent accurately recognizes whether a question is within the scope of the role's knowledge and explicitly declines to answer questions that are out of scope.

Evaluation steps:
1. Analyze the skills and levels required to answer the given question, and carefully read the profile given above, and judge whether the character possess the corresponding skills and levels, taking into account social identity, experience and other information.
2. Check how the agent responds to the question: if the question asks for a skill that the character does not possess, the agent should effectively refuse to respond and indicate that he/she does not have the relevant knowledge; conversely the agent should follow the instructions of the question and give a normal response.
3. Please explain in the rationale: Whether the level of knowledge required to answer the question is mastered by the given character or not? Does the Agent's response strategy (normal answer or rejection) match the expected behavior?

Your output should be organized to a valid JSON dictionary which can be parsed by Python.

Figure 24: Prompt used to evaluate general-domain hallucination avoidance (G.HAL)