Identifying User-Inference-Dependent Out-of-Character Behavior in LLM-Based Role-Playing Agents

Shinji Iwata[1] **Shiki Sato**[1]
Jun Baba[1] **Koya Ihara**[1] **Michimasa Inaba**[2]
[1]CyberAgent
[2]The University of Electro-Communications
{iwata_shinji,sato_shiki,baba_jun,ihara_koya}@cyberagent.co.jp
m-inaba@uec.ac.jp

Abstract

For the practical deployment of the Role-Playing Agents with Large Language Models, prior work has mainly focused on suppressing out-of-character behavior relative to explicitly provided character information such as personas. However, as shown in psychological research, out-of-character behavior can arise not only from contradictions with explicit information but also from mismatches with character images implicitly inferred by users. Modeling user-inference-dependent out-of-character behavior is a challenging task, because it is unclear what kinds of inferences users make from existing information. To tackle this problem, we construct a dataset of 480 dialogue scenarios and collect 2,152 descriptions from users familiar with those titles. Using cases that annotators judged inappropriate despite the absence of contradiction with existing settings, we derive a 12-category typology, organized into general, title-specific, and individual-specific groups. We also conduct an exploratory evaluation of category-guided feedback, finding no consistent overall advantage over standard self-refinement but possible benefits for some characters. Our shareable code and dataset are at https://github.com/CyberAgentAILab/User-Inference-OOC.

1 Introduction

With the advancement of large language models (LLMs), research on role-playing agents (RPAs) that imitate specific individuals has also progressed rapidly. Because RPAs that emulate attractive personalities or fictional characters can enhance user engagement, they have been used in a wide range of real-world applications, including advertising (Mizukami et al., 2023), education (Noh and Hong, 2021; Pataranutaporn et al., 2023), and grief care (Xygkou et al., 2023). Accordingly, many studies have explored how to reflect character information in LLM-generated utterances by providing detailed

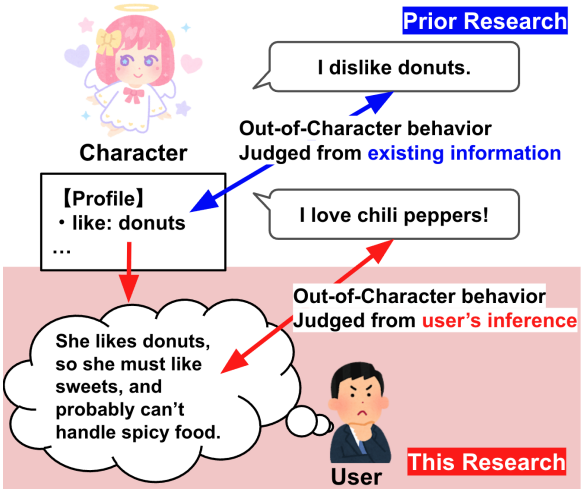


Figure 1: Two types of out-of-character behavior

character profiles and generating utterances that take them into account (see Sec. 2.1). Recently, researchers have pointed out that LLMs may generate out-of-character utterances that deviate from the input character profile, and have worked on suppressing such behavior (Shin et al., 2025).

However, it is important to account not only for out-of-character behavior based on explicitly provided character information, but also for deviations from the implicit character image that individual users derive by interpreting and extending that information. As has already been noted in psychology (Winter and Uleman, 1984; Uleman et al., 1996; Bott et al., 2024), humans interpret attributes about a person from the stimulus information they are given and form a character image of that person. An example is shown in Figure 1. Users form their own interpretations from the given settings and utterances, thereby constructing a character image. When the character behaves deviating from this user-constructed image, users perceive the behavior as inappropriate for the character. Modeling user-inference-dependent out-of-character behavior in LLM-generated utterances is challenging because

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4927–4943
November 6-10, 2026 ©2026 Association for Computational Linguistics

such inferences lack predefined categories and are diverse and sometimes illogical.

In this study, to better understand out-of-character behavior as inferred by users, we construct dataset from descriptions provided by users familiar with the games and develop a typology of out-of-character behavior based on it. Specifically, we prompt LLMs to generate advertising scenarios for multiple characters from several game titles, and ask users familiar with those games to annotate spans they perceive as out-of-character behavior. We conducted our study in the domain of advertising, one of the practical settings in which RPAs are already used. For each span, annotators indicate whether it explicitly contradicts the provided settings/past statements or is inappropriate only relative to their own interpretation and explain why. We then construct a typology of out-of-character behavior based on these descriptions. Finally, we identify 12 categories: those widely perceived as out-of-character behavior, those whose perception varies across titles, and those noticed by only a few users.

Furthermore, we conducted an A/B evaluation to explore whether these categories can be useful as feedback for generation. Although category-guided feedback did not show a consistent overall advantage over standard self-refinement, a post-hoc analysis suggested that it may help reduce perceived out-of-character behavior for many characters.

Our contributions are threefold: (i) We focus on out-of-character behavior from user inference, an aspect that has not been underexplored in prior RPA research. (ii) Through annotations by people familiar with actual games, we construct dataset and develop a typology of out-of-character behavior from user inference and examine how broadly they recur across our three games. (iii) We conduct an exploratory evaluation of category-guided feedback and find preliminary, character-dependent signals that it may help reduce perceived out-of-character behavior in some settings.

2 Related Work

2.1 Role-Playing Agents

Studies on Role-Playing Agents (RPAs) that imitate a specific person have a long history (Cavazza et al., 2002; Traum, 2006; Leuski et al., 2006). Later, with advances in neural networks and pre-trained models, many studies explored RPAs based on these methods. For neural dialogue models, prior work used speaker embeddings (Li et al., 2016) and self-awareness (Kim et al., 2020).

More recently, with the rise of large language models, researchers have tried to improve character fidelity by providing large amounts of profile information about a character or person as input. Starting from Character-LLM (Shao et al., 2023), prior studies have examined how to keep consistency with given utterances and settings (Liu et al., 2025; Shin et al., 2025; Jeong et al., 2025), as well as how to reduce deviations from existing character settings (Sadeq et al., 2024; Lu et al., 2024; Ahn et al., 2024; Kong et al., 2025; Shin et al., 2025).

However, prior work often assesses whether a generated utterance is character-like only with respect to the explicitly provided profile, rather than the character image that users infer from that information. Therefore, beyond these research directions, identifying the factors that make users feel that "the character is not behaving as they normally would" in response to LLM-generated utterances would help in developing RPAs that feel more natural and less jarring to a wide range of users.

2.2 Evaluation and Analysis of Character-Likeness

As RPAs that imitate existing characters have advanced, researchers have begun to develop benchmarks for evaluating character-likeness and to study how character-likeness can be understood. Previous work has evaluated character-likeness from multiple dimensions such as style, personality, knowledge, behavior, and dialogue ability (Wang et al., 2024b; Tu et al., 2024; Lu et al., 2024; Wang et al., 2024a; Wu et al., 2025; He et al., 2025). Furthermore, Iwata et al. (2025) decomposed character-likeness into 27 components. These studies have clarified important aspects to evaluate in RPA development.

However, it is still unclear what kinds of quantitative or qualitative outcomes make users feel that a character is not acting like itself. In this study, we provide qualitative insights and an empirical basis for future quantitative investigations of how users infer and perceive out-of-character behavior in utterances generated by LLMs.

3 Data Collection

We constructed scenarios in which specific promotional materials are advertised and collected annotations of out-of-character behavior. This section

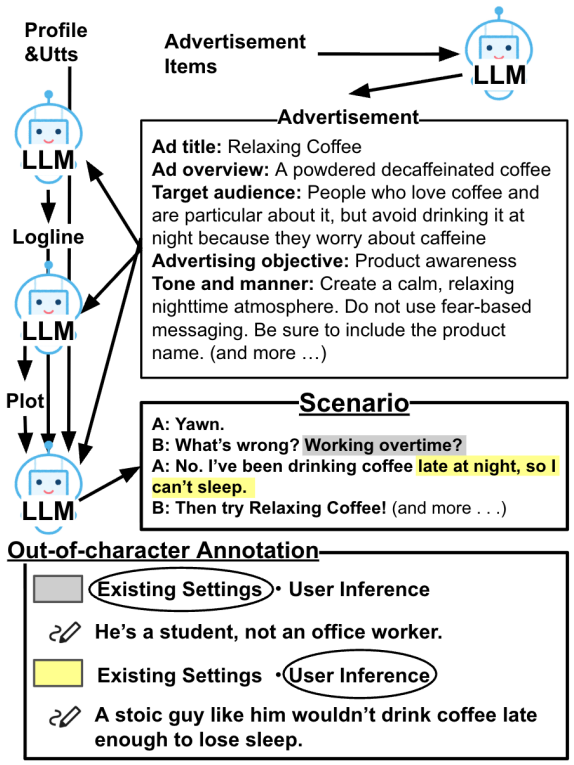


Figure 2: Overview of data collection.

describes the dataset construction procedure, illustrated in Figure 2.

We selected 18 characters from three game titles (Sec. 3.1) and used LLMs to generate dialogue scenarios in which they advertise specific promotional materials (Sec. 3.2)[1]. We then asked users who had read scenarios of the games to annotate out-of-character behavior spans and explain why they felt so (Sec. 3.3).

3.1 Titles and Characters

We used 18 characters from three game titles with different world settings and character compositions: IDOLY PRIDE[2], 100 Sleeping Princes and the Kingdom of Dreams[3], and JewelSaviorFree[4]. The titles cover real-world setting and fantasy setting. We selected six characters from each title.

3.2 Scenario Generation for Annotation

The dataset consists of 480 one-on-one dialogue scenarios generated with LLMs. For each title, we created six character pairs from the six characters described in Section 3.1 and generated scenarios for each pair. Following prior work (Yang et al., 2022, 2023), we used a staged process that first generated a three-line logline, then a detailed plot, and finally a dialogue script.

At each stage, we used character profiles and advertisement information as input to the LLM. In addition, we used generated logline as input during plot generation, and both the logline and the plot during scenario generation. When character profiles were unavailable, we generated them from dialogue utterances following Yuan et al. (2024). We used four LLMs: GPT-5.2[5], Gemini-3-Pro[6], Grok-4.1-fast-reasoning[7] and Claude-Opus-4.6[8]. The advertisement information also included advertising materials automatically generated by GPT-5.2 from items derived from marketing guidelines published by an expert association (Digital Marketing Institute, 2023).

Number of Scenarios and Descriptions	
# Scenario	480
# Annotators per Scenario	5
# Annotated Descriptions (N)	2152
Number of Descriptions by Title	
IDOLY PRIDE ($N = 1092$)	
# Contradictions with existing settings	455
# Contradictions with user inference	637
100 Sleeping Princes and the Kingdom of Dreams ($N = 664$)	
# Contradictions with existing settings	324
# Contradictions with user inference	340
JewelSaviorFree ($N = 396$)	
# Contradictions with existing settings	193
# Contradictions with user inference	203

Table 1: Dataset and annotation statistics.

3.3 Annotation Process

We recruited annotators who had read the game scenarios of each title and assigned five annotators to each target scenario. During annotation, we provided annotators with the advertisement information (its name and a short overview), character profiles, and the target scenario. We asked them to select any number of utterance spans they judged to exhibit out-of-character behavior. For

[1] Prompts are described in Appendix A.2
[2] https://idolypride.jp/
[3] https://www.yume-100.com/
[4] https://jewel-s.jp/
[5] https://developers.openai.com/api/docs/models/gpt-5.2
[6] https://ai.google.dev/gemini-api/docs/gemini-3
[7] https://x.ai/news/grok-4-1-fast
[8] https://www.anthropic.com/news/claude-opus-4-6

each selected span, we asked them to indicate in a select box whether the out-of-character behavior was due to a contradiction with existing settings or past statements, or whether it was not explicitly contradicted by them but was still considered inappropriate based on their own interpretation. We also asked them to provide a free-text explanation for each judgment.

The number of annotators differed across titles: 25 for IDOLY PRIDE and 12 each for 100 Sleeping Princes and the Kingdom of Dreams and JewelSaviorFree. Accordingly, the number of annotations also varied: IDOLY PRIDE yielded 1,092 descriptions from 240 scenarios, while 100 Sleeping Princes and the Kingdom of Dreams and JewelSaviorFree each yielded 664 and 396 descriptions from 120 scenarios. Statistics on the annotations are in Table 1.

4 Categories of User-Inferred Out-of-Character Behavior

We created categories of out-of-character behavior derived from annotations of LLM-generated utterances collected in Section 3. As noted above, this study aims to clarify mismatches between LLM-generated behavior and users' implicitly inferred character images. We therefore used only descriptions labeled as "not explicitly inconsistent with existing settings or past statements, but still inappropriate based on the annotator's own interpretation. " From these descriptions, we constructed categories of out-of-character behavior and retained only those observed in at least two of the three titles. Using cases from multiple titles, rather than only one, reduces the risk that the resulting categories depend too much on a particular game. Table 2 shows the resulting categories. Additionally, to provide a simple check of the objectivity of the categories, we asked another annotator to classify a subset of the descriptions into the categories. As a result, Cohen's κ was 0.65, suggesting a moderate inter-coder agreement.

We divided the categories into three groups based on their generality, ranging from those commonly mentioned across many cases, to those characteristic of particular titles, and to those specific to individual users. We defined General Categories as those mentioned by a majority of annotators in all three titles, Title-Specific Categories as those mentioned by a majority in one or two titles, and Individual-Specific Categories as those for which no game had a majority. Using descriptions in which annotators judged a scenario to exhibit out-of-character behavior, we counted how many annotators mentioned each category. We treated a category as having majority support when it was mentioned by more than half of all annotators, since this suggests that the judgment is likely to be shared by a broad range of users. Below, we describe each category in detail.

4.1 General Categories

The General Categories comprised three categories: (i) mismatch with perceived emotional expression or personality, (ii) mismatch with the way things, events, or objects are described, or with the word choice used and (iii) mismatch with daily habits and everyday behavior. All three were mentioned by a majority of annotators within each title, suggesting that they are shared dimensions of out-of-character behavior beyond differences in title-specific settings, individual characters, and users.

In utterances expressing a character's emotions or personality, we observed cases where the emotional expression was too strong or the utterance did not match the character's usual personality. Annotators also repeatedly pointed out that the type or intensity of the expressed personality or emotion differed from their expectations. We therefore created the category "Mismatch with perceived emotional expression or personality. " This suggests that out-of-character behavior depends not only on what personality or emotion is expressed, but also on its intensity. In other words, if the degree of expression is not properly controlled, such as making the character sound brighter or less negative, many users may perceive the character as out-of-character.

When a character described a particular matter, event, or object, we observed patterns such as overly casual word choices that would not normally be expected from the character and unnatural figurative expressions. Annotators also often pointed out that the vocabulary or way of expression differed from their expectations. We therefore created the category "Mismatch with the way things, events, or objects are described, or with the word choice used." For example, this category applies when a well-educated upper-class character uses impolite expressions such as "jerk."

When a character spoke about their daily habits or everyday behavior, we observed cases where the habits or actions differed from the expected charac-

Category name	Explanation	
General Categories		
Mismatch with perceived emotional expression or personality	Mismatch in the degree or content in which the character's emotions or personality are expressed. Example: The character is cheerful, but gives a response that is overly exaggerated.	✓
Mismatch with the way things, events, or objects are described, or with the word choice used	Expressions or word choices do not fit how the character usually speaks (e.g., tendency to use certain figures of speech or ways of saying things, whether their vocabulary is slang-like, and their general word choice). Example: A well-educated character uses impolite expressions such as "suck."	✓
Mismatch with daily habits and everyday behavior	The daily actions or habitual behaviors expressed in the utterance do not fit the character. Example: A disciplined person eats nutritionally unbalanced meals.	-
Title-Specific Categories		
Mismatch with values	The values the character uses to judge actions or think about things do not fit the character (e.g., motives for action, priorities, or interests). Example: A wealthy character worries about whether a discount is available.	-
Mismatch with the tone of speech and the form of address used toward a specific character	The expressions or behaviors suggested by the relationship between characters, such as their level of closeness or relative position, do not fit the character (e.g., whether polite language is used, or the choice of honorifics or forms of address). Example: Calling a close best friend only by their family name.	✓
Mismatch with the character's sentence endings	The sentence endings used in utterance do not match the character's expected sentence endings. Example: A strict person ends a sentence with "yo."	✓
Mismatch with utterance planning and dialogue ability	Character's utterance structure, flow of logic, topic connection, and response pattern do not fit the character (e.g., unnatural omission, failure to keep consistency within an utterance, or mismatched responses). Example: A character who can normally hold a conversation does not answer the previous question and instead starts talking about themselves.	✓
Mismatch with level of knowledge or proficiency, including the appropriateness of expressions or actions of knowledge or proficiency	The character speaks unnaturally about things they are unlikely to know well or be familiar with. Conversely, they may speak as if they do not know things they would normally be expected to know or be familiar with. Example: A character who is bad with machines talks about a product's performance at an expert level.	✓
Individual-Specific Categories		
Mismatch with the sense of relational distance toward a specific conversation partner	The sense of psychological distance between the two characters, as suggested by their relationship such as closeness or relative position, is not expressed in a way that fits the character. Example: Two rival characters act in an overly dependent or clingy way toward each other.	✓
Mismatch with the naturalness of how the character's own traits or background are expressed	When traits or background information that define the character are included in an utterance, the way they are brought up should be natural in context and consistent with the character. Example: A character with high self-esteem suddenly says out of nowhere that they have high self-esteem.	-
Mismatch with multiple stereotypical attributes	When a character has multiple stereotypical attributes, none of those attributes should feel unnatural or out of place. Example: A character who is a "wealthy engineer" spends an hour worrying about using an API that costs only a few dozen dollars (this may not feel strange from the viewpoint of an engineer, but it does feel strange from the viewpoint of a wealthy person).	-
Mismatch with the background, world setting, or time period of the world the character belongs	The characters refer to things that do not fit, or do not exist, in their world setting, background, or time period. Example: A character from a medieval fantasy world is familiar with mobile phones.	-

Table 2: Categories of out-of-character behavior derived from user inference. The rightmost column indicates whether each category has been discussed in prior work (Sec. 2.2). ✓: discussed in prior work; -: not discussed in prior work.

ter image. Annotators repeatedly pointed out that such behavior was hard to imagine for the character, or that they instead imagined the character acting differently. We therefore organized this as the category "Mismatch with daily habits and everyday behavior." When an LLM generates new daily habits or everyday behaviors not found in existing information, we should check each behavior against the character image users have in mind.

4.2 Title-Specific Categories

The Title-Specific Categories included five categories: (i) Mismatch with values, (ii) mismatch with the tone of speech and the form of address used toward a specific character, (iii) mismatch with the character's sentence endings, (iv) Mismatch with utterance planning and dialogue abil-

ity, (v) Mismatch with level of knowledge or proficiency, including the appropriateness of expressions or actions of knowledge or proficiency. These categories were mentioned by a majority of annotators in one or two titles. This suggests that, although they have a certain level of generality, how clearly they appear may change depending on the traits of the characters in each title.

In utterances describing a character's actions, we observed cases where the action policy, priorities, or interests inferred from those actions differed from the expected character image. Annotators also repeatedly pointed out that such action policies or interests felt unnatural for the character or that the character would be expected to act differently. We therefore created the category "Mismatch with values." When the standards for action inferred from utterance, such as what the character cares about or is strongly interested in, differ from users' expectations, the character may be perceived as out-of-character.

In dialogues with a specific conversation partner, we observed cases where the character omitted a form of address, added an unnecessary one, or used an inappropriate tone. Annotators also often reported discomfort with such expressions. We therefore introduced the category "Mismatch with the tone of speech and the form of address used toward a specific conversation partner." This kind of out-of-character impression arises when honorifics or speech tone, which should reflect the relationship with the other person, the speaker's role in the conversation, and the speaker's personality, differ from users' expectations.

We also observed cases where the sentence endings in the character's speech diverged from the character's established persona. Annotators likewise often pointed out that these sentence endings felt off. We therefore established the category "Mismatch with the character's sentence endings." This category arises when sentence endings that should reflect the character's prior manner of speaking (for example, "yo") differ from what users expect.

In a character's replies or utterances, we observed cases where the response or content differed from the level of dialogue ability normally expected. Annotators also repeatedly pointed out that some utterances did not properly answer the previous statement, contained unnatural omissions, or should have been more appropriate and natural for the character. We therefore organized this as the category "Mismatch with utterance planning and dialogue ability." For example, this category applies when a character who can normally hold a conversation starts talking about themselves without answering the previous question, or when a character with a normal speaking style produces an unnatural omission, such as saying "This pen" instead of "This is a pen."

	IDOL	Sleep	Jewel
Mismatch with values	✓	✓	
Mismatch with the tone of speech and the form of address used toward a specific character	✓	✓	
Mismatch with the character's sentence endings		✓	
Mismatch with utterance planning and dialogue ability	✓	✓	
Mismatch with level of knowledge or proficiency	✓		

Table 3: Titles in which each title-contingent category was mentioned by a majority of annotators. IDOL: IDOLY PRIDE, Sleep: 100 Sleeping Princes and the Kingdom of Dreams, Jewel: JewelSaviorFree

When a character spoke about matters requiring knowledge or skill, we observed cases where the explanation or reaction differed from the level of knowledge or proficiency normally expected. Annotators also repeatedly pointed out that such responses were hard to imagine given the character's usual level of knowledge or skill. We therefore introduced the category "Mismatch with level of knowledge or proficiency." For example, this category applies when a character who is bad with machines talks about a product's performance at an expert level.

Because the salience of these issues varies across works, it is necessary to identify the kinds of works in which they require particular attention and to decide on appropriate countermeasures accordingly. The next section therefore analyzes in which works the major categories appeared universally.

4.2.1 Discussion of Title-Specific Categories

Table 3 shows the titles in which a majority of annotators mentioned each category. First, a majority mentioned four of the five categories in IDOLY PRIDE and 100 Sleeping Princes and the Kingdom of Dreams, whereas no category was mentioned by a majority in JewelSaviorFree. One possible explanation is that the role of the target characters in each title. In IDOLY PRIDE and 100 Sleeping Princes and the Kingdom of Dreams, we used characters who appear in the main story. As a result, these titles provide detailed descriptions of

the characters' feelings and personalities, as well as how they change over the course of the story. In contrast, JewelSaviorFree contains almost no direct interaction with the characters in the main story, and character-related scenarios are developed mainly in side stories separate from the main plot. In this format, descriptions of personality and feelings are more limited, giving users more room for their own interpretations of the characters. For this reason, whether this category requires special care may depend on how fully the target characters are described in the original work.

4.3 Individual-Specific Categories

Individual-Specific included four categories: (i) mismatch with the sense of relational distance toward a specific conversation partner, (ii) mismatch with the naturalness of how the character's own traits or background are expressed, (iii) mismatch with multiple stereotypical attributes and (iv) mismatch with the background, world setting, or time period of the world the character belongs.

Although none of these categories was mentioned by a majority in any title, they were still a clear source of discomfort for some users across multiple titles. Therefore, Individual-Specific Categories should not be regarded as less important, but as subtle forms of out-of-character behavior that are noticeable to particular users.

In dialogues with a character who has a specific relationship with the target character, we observed cases where the utterance suggested a sense of relational distance different from what would normally be expected. Annotators also noted that the distance implied by the utterance differed from their expectations. We therefore defined the category "Mismatch with the sense of relational distance toward a specific conversation partner." Even when polite language or forms of address are formally appropriate, problems can arise when a particular pair of characters has its own unique relationship. For example, a character may suddenly show a level of closeness that feels too strong to users or, conversely, respond in a way that feels less close than expected. In such cases, the problem lies in a mismatch in the psychological distance between the two characters.

There were cases where a character disclosed information about themselves in a context that felt unnatural. Annotators also pointed out that the disclosed content felt unnatural. We therefore defined the category "Mismatch with the naturalness of how the character's own traits or background are expressed." Even information that should help convey the character's identity can make the character feel out-of-character when it is introduced in an unnatural context.

When a character had multiple stereotypical attributes, we observed cases where an utterance felt unnatural from one attribute even if it did not from another. Annotators likewise pointed out that an utterance could seem natural from the viewpoint of one attribute but still feel out-of-character from another. We therefore defined the category "Mismatch with multiple stereotypical attributes." For example, if a character is a wealthy engineer, spending an hour worrying about an API that costs only a few dozen dollars may not seem strange from the viewpoint of an engineer, but it does from the viewpoint of a wealthy person.

When a character introduced a particular thing, we observed cases where the content or way of description did not fit the background, world setting, or time period to which the character belongs. Annotators also repeatedly pointed out that the thing described, or the way it was expressed, did not match the character's world setting or time period. We therefore defined the category "Mismatch with the background, world setting, or time period of the world the character belongs to." This was especially likely in settings like ours, where characters from existing works introduce promoted products/items, because modern product concepts or advertising expressions may appear in the speech of characters from fantasy worlds, creating friction with the world setting. In fact, this type of description was observed in 100 Sleeping Princes and the Kingdom of Dreams and JewelSaviorFree, which have stronger fantasy elements, but not in IDOLY PRIDE, which is set in a world closer to the real world. However, even in the two titles where such descriptions were observed, the number of participants who identified them as out-of-character did not exceed a majority, suggesting that these mismatches are salient to a subset of users rather than broadly shared.

4.4 Relationship with prior work

We examined the extent to which the categories identified in this study relate to points discussed in prior work on character-likeness (Sec. 2.2). However, this comparison requires caution. This is because our study focuses on categories of out-of-character behavior inferred by users. Therefore,

even when there is overlap with points discussed in prior work on character-likeness, our study defines what kinds of user inferences are assumed based on those points, and thus does not deal with exactly the same concept. Among our 12 categories, 7 correspond to points also discussed in prior work and the remaining 5 were not discussed in prior work (see Appendix A.4 for details). These results suggest that the descriptions collected in this study not only allowed us to extract the necessary points already discussed in prior work on character-likeness, but also captured new points that can be identified only by focusing on out-of-character behavior.

5 Experiment

We conducted an exploratory examination of whether a category-based model better suppresses out-of-character expressions than a category-agnostic model. Improved suppression would suggest that our categories may be useful for analyzing out-of-character behavior.

5.1 Experiment Settings

Characters. We used four characters from IDOLY PRIDE. To avoid direct overlap with the data used for category construction, we used characters different from those in Section 3.

Conditions. We compared three generation settings[9]. No Feedback uses the staged generation pipeline described in Sec 3, where the logline, plot, and scenario are generated sequentially. Self-Refine Feedback adds an iterative self-feedback loop based on Self-Refine (Madaan et al., 2023), where an LLM reviews the generated scenario and revises it accordingly. We repeated this refinement loop four times using the character profile and example utterances as input. Category-Guided Feedback extends the Self-Refine setting by additionally providing the categories of user-inferred out-of-character behavior introduced in Section 4, so that the model can generate feedback with explicit reference to these categories. We used four LLMs for generation: GPT-5.2, Gemini-3-Pro, Grok-4.1-fast-reasoning, and Claude-Opus-4.6. As in Section 3, automatically generated advertising information was also provided during scenario generation.

Participants and Evaluation Method. We recruited 22 participants for pairwise A/B comparisons under two of the three conditions described above. They were asked which scenario better matched each of six questions[10], including questions about out-of-character behavior. In total, we collected 188 comparisons.

5.2 Results

For the most basic question asking which scenario was more out-of-character, the response ratios for No Feedback vs. Self-Refine Feedback, No Feedback vs. Category-Guided Feedback, and Self-Refine Feedback vs. Category-Guided Feedback were 65% vs. 35%, 59% vs. 41%, and 44% vs. 56%, respectively. Overall, these results indicate that adding feedback was beneficial relative to no feedback, but do not show a consistent overall advantage of category-guided feedback over standard self-refinement.

We then conducted a post-hoc character-level analysis and found that the trend differed depending on whether one particular character, Chisa, was included. For comparisons between Self-Refine Feedback and Category-Guided Feedback, the response ratio was 56% vs. 44% when Chisa was excluded, but 32% vs. 68% when she was included. This suggests that the usefulness of category-guided feedback may be character-dependent rather than uniform across all settings.

Overall, category-guided feedback did not consistently outperform standard self-refinement, but character-level analysis suggested benefits for three of four characters by reducing perceived out-of-character behavior.

5.3 Discussion

One possible explanation for Chisa's exceptional case is that, for some characters, stronger trait consistency may conflict with the ad scenario's task demands. As described on the official character introduction page, Chisa has a shy personality and isn't very assertive[11]. In Chisa's case, a more character-faithful style may have made the product explanation less detailed or less assertive, which could have increased the perceived mismatch with the advertising role [12].

[9] Prompts are described in Appendix A.3.

[10] All questionnaires are described in Appendix A.5

[11] https://idolypride.jp/character/chisa-shiraishi/

[12] According to the authors' annotations, 33% of all cases in which Category-Guided Feedback was judged to be more out-of-character than Self-Refine Feedback were those where Chisa was responsible for introducing the product and the product explanation was simplified.

6 Conclusion

In this study, we constructed a dataset focused on out-of-character behavior as inferred by users. Our dataset covers three game titles and contains a total of 480 scenarios and 2,152 subjective annotation descriptions. Using this dataset, we identified 12 categories of user-inferred out-of-character behavior, including general, title-specific, and individual-specific categories. We hope this study offers an empirical foundation for future research on understanding and mitigating out-of-character behavior in RPAs.

7 Limitations

Despite the encouraging results, this study has several limitations.

Character Variety. Although this study covers a broad range of characters by using 18 characters from three titles, it is unclear whether these 18 characters are sufficient to cover the full diversity of characters. Expanding the dataset to include more characters and domains beyond games, such as novels, remains future work.

Content Domain Variety. This study focuses only on existing game titles. Characters in other forms of content, such as novels, may show different tendencies. In addition, our experiments were conducted in the domain of advertising. Examining how the findings may vary across different target domains remains an important direction for future work.

Benchmark Creation. In this study, our primary objective was to construct a set of categories; therefore, we did not develop a quantitatively measurable benchmark. Developing such a benchmark remains future work.

Taxonomy-Guided Profile Generation. In this study, we did not consider enriching the input profile descriptions, because covering all possible information that might be needed would be unrealistic. Assessing the extent to which current technologies can cover such information remains an important direction.

8 Ethical Considerations

The IDOLY PRIDE and 100 Sleeping Princes and the Kingdom of Dreams data used in this study were obtained and used with official permission. JewelSaviorFree is publicly available online, and its data were used in accordance with the applicable terms of use. In our annotations and experiments, we did not collect personality inventories (e.g., Big Five) or other sensitive individual difference measures. For human participants, we obtained consent after explaining the risks of taking part and how their information would be handled.

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.

Antonia Bott, Larissa Brockmann, Ivo Denneberg, Espen Henken, Niclas Kuper, Felix Kruse, and Juliane Degner. 2024. Spontaneous trait inferences from behavior: A systematic meta-analysis. *Personality and Social Psychology Bulletin*, 50(1):78–102.

Marc Cavazza, Fred Charles, and Steven J. Mead. 2002. Interacting with virtual characters in interactive storytelling. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1*, AAMAS '02, page 318 – 325, New York, NY, USA. Association for Computing Machinery.

Digital Marketing Institute. 2023. Launch of an orientation format for marketing operations: Created based on input from 30 practitioners from advertisers and partner companies (in japanese). (in Japanese). Accessed: 2026-03-17.

Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaxing Xu, Wenqiang Liu, and Mengling Feng. 2025. Crab: A novel configurable role-playing LLM with assessing benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15030–15052, Vienna, Austria. Association for Computational Linguistics.

Shinji Iwata, Koya Ihara, Shiki Sato, Jun Baba, Asahi Hentona, Masahiro Yamazaki, Yuki Shiotsuka, Takahiro Ishizue, and Akifumi Yoshimoto. 2025. Identification and analysis of identity-centric elements of character-likeness in game scenario. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–284, Avignon, France. Association for Computational Linguistics.

Yeonseok Jeong, Minsoo Kim, Seung-won Hwang, and Byung-Hak Kim. 2025. Agent-as-judge for factual summarization of long narratives. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23591–23608, Suzhou, China. Association for Computational Linguistics.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Chuyi Kong, Ziyang Luo, Hongzhan Lin, Zhiyuan Fan, Yaxin Fan, Yuxi Sun, and Jing Ma. 2025. SHARP: Unlocking interactive hallucination via stance transfer in role-playing LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 839–866, Vienna, Austria. Association for Computational Linguistics.

Anton Leuski, Ronakkumar Patel, David Traum, and Brandon Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27, Sydney, Australia. Association for Computational Linguistics.

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

Wenhao Liu, Siyu An, Junru Lu, Muling Wu, Tianlong Li, Xiaohua Wang, Changze Lv, Xiaoqing Zheng, Di Yin, Xing Sun, and Xuanjing Huang. 2025. Tell me what you don't know: Enhancing refusal capabilities of role-playing agents via representation space analysis and editing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5983–6005, Vienna, Austria. Association for Computational Linguistics.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Masahiro Mizukami, Hiroaki Sugiyama, Tsunehiro Arimoto, Ryuichiro Higashinaka, Koh Mitsuda, and Tessei Kobayashi. 2023. Construction of an Information Service Dialogue System in Collaboration with Local Governments using the Chatbot Platform of Role Play-based Question-Answering (in Japanese). *The Japanese Society for Artificial Intelligence*, 38(3):B–MA2_1.

Yeo-Gyeong Noh and Jin-Hyuk Hong. 2021. Designing reenacted chatbots to enhance museum experience. *Applied Sciences*, 11(16):7420.

Pat Pataranutaporn, Valdemar Danry, Lancelot Blanchard, Lavanay Thakral, Naoki Ohsugi, Pattie Maes, and Misha Sra. 2023. Living memories: Ai-generated characters as digital mementos. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 889 – 901, New York, NY, USA. Association for Computing Machinery.

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.

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.

Jisu Shin, Juhyun Oh, Eunsu Kim, Hoyun Song, and Alice Oh. 2025. Spotting out-of-character behavior: Atomic-level evaluation of persona fidelity in open-ended generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26312–26332, Vienna, Austria. Association for Computational Linguistics.

David Traum. 2006. Talking to virtual humans: dialogue models and methodologies for embodied conversational agents. In *Proceedings of the Embodied Communication in Humans and Machines, 2nd ZiF Research Group International Conference on Modeling Communication with Robots and Virtual Humans*, ZiF'06, page 296 – 309, Berlin, Heidelberg. Springer-Verlag.

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.

James S Uleman, Alex Hon, Robert J Roman, and Gordon B Moskowitz. 1996. On-line evidence for spontaneous trait inferences at encoding. *Personality and Social Psychology Bulletin*, 22(4):377–394.

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. 2024a. 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.

Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. 2024b. Characteristic AI agents via large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3016–3027, Torino, Italia. ELRA and ICCL.

Laraine Winter and James S Uleman. 1984. When are social judgments made? evidence for the spontaneousness of trait inferences. *Journal of personality and social psychology*, 47(2):237.

Bowen Wu, Kaili Sun, Ziwei Bai, Ying Li, and Baoxun Wang. 2025. RAIDEN benchmark: Evaluating role-playing conversational agents with measurement-driven custom dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11086–11106, Abu Dhabi, UAE. Association for Computational Linguistics.

Anna Xygkou, Panote Siriaraya, Alexandra Covaci, Holly Gwen Prigerson, Robert Neimeyer, Chee Siang Ang, and Wan-Jou She. 2023. The "conversation" about loss: Understanding how chatbot technology was used in supporting people in grief. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. DOC: Improving long story coherence with detailed outline control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 Domain-General vs. Ad-Specific

We conducted a small-scale exploratory experiment to examine whether the categories developed in this study are domain-general or ad-specific.

A.1.1 Experimental Settings

Participants. The participants were three individuals who were considered to have knowledge relevant to the topic of this study. The first participant was a researcher specializing in conversational agents, while the second and third participants were game AI engineers with practical experience related to character utterances and scenarios.

Questions. We presented the participants with the category names and their descriptions, and asked them three questions for each category. The first question asked whether they thought the category represented an issue that generally arises in scenario generation beyond advertising contexts (Domain-General). The second question asked whether they thought the category was more likely to arise in advertising scenario generation than in general scenario generation (Ad-Salient). The third question asked whether they thought the category represented an issue specific to advertising contexts (Ad-Specific). The second question was designed to assess whether a category, while also appearing in general scenario generation, becomes particularly salient in advertising contexts. The third question was designed to assess whether a category was ad-specific. For each question, participants were asked to respond on a five-point Likert scale: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, and 5 = strongly agree.

A.1.2 Results

The results are shown in Table 4. First, for the first question, which concerned whether each category was domain-general, at least two of the three participants rated every category as 4 (agree) or higher. In addition, for the third question, which concerned whether each category was ad-specific, all participants rated every category as 3 (neither agree nor disagree) or lower. These results suggest that the categories developed in this study primarily consist of domain-general categories. For the second question, which concerned ad-salience, the response patterns differed across categories. These results preliminary suggest that the categories developed in this study may not be ad-specific but domain-general; however, their saliency may vary across domains. A broader validation using a wider range of domains therefore remains an important direction for future work.

A.2 Prompt used in Sec. 3

A.2.1 Advertisement Information Generation

Prompt

Please create a single example of promotional material and an advertisement you wish to commission, and fill in the required details. It has been decided that this will be an advertisement featuring a conversation between two specific characters, but all other details are yet to be determined. As the characters to be featured have not yet been decided, please do not assign them names.

- Name of promotional subject
- Overview of promotional subject
- Overview of promotion
- Target profile
- Needs & Insights
- Attitude Change (Customer Journey)
- Mandatory Items & Evaluation Criteria
- Industry-Specific Rules
- Brand Guidelines
- Assets
- Tone & Manner
- Advertising Objective
- Advertising Medium
- Advertisement Length

A.2.2 Logline Generation

Prompt

Please create the following items regarding the scenario in which the two characters promote the designated product. The specifications for this are also set out below.

- Scenario overview (3 lines)

* Please do not include information such as duration or on-screen displays; describe only the actions performed by the characters.

{Character Profile}
{Character Utterances}

Category name	Q1: Domain-General	Q2: Ad-Salient	Q3: Ad-Specific
Mismatch with perceived emotional expression or personality	4/4/5	4/4/4	1/1/1
Mismatch with the way things, events, or objects are described, or with the word choice used	4/2/5	5/4/2	1/1/1
Mismatch with daily habits and everyday behavior	4/4/5	4/4/4	1/1/1
Mismatch with values	4/5/5	4/3/4	1/1/1
Mismatch with the tone of speech and the form of address used toward a specific character	5/5/5	2/2/1	1/1/1
Mismatch with the character's sentence endings	5/1/5	2/3/1	1/1/1
Mismatch with utterance planning and dialogue ability	5/4/4	5/3/4	1/1/1
Mismatch with level of knowledge or proficiency, including the appropriateness of expressions or actions of knowledge or proficiency	4/4/5	5/4/4	2/1/1
Mismatch with the sense of relational distance toward a specific conversation partner	5/4/5	2/1/2	1/1/1
Mismatch with the naturalness of how the character's own traits or background are expressed	5/4/4	2/3/4	1/1/1
Mismatch with multiple stereotypical attributes	4/5/3	4/4/3	1/1/3
Mismatch with the background, world setting, or time period of the world the character belongs	4/4/5	5/4/4	2/1/1

Table 4: Responses from the three participants for each category. Ratings range from 1: strongly disagree to 5: strongly agree.

A.2.3 Plot Generation

Prompt

Please create a single plot outline for a scenario in which two characters promote a given product, based on the logline provided. Please also note the following requirements:
・A plot outline detailing the characters' actions within the scenario (the plot should describe the characters' actions that drive the narrative flow, rather than their dialogue)

【Sample Plot Description Format】
Character A does X. Character B does Y. ...
* Please do not include information such as timings or on-screen displays; describe only the actions performed by the characters.
{Character Profile}
{Character Utterances}
{Generated Logline}

A.2.4 Scenario Generation

Prompt

Please create a scenario for two characters promoting the given promotional subject, based on the logline and plot provided.
Do not output anything other than the characters' lines and full names. For the characters' full names, do not simply output 'A' or 'B'; please output their full names.

【Sample scenario format】
Character A's full name > Line A
Character B's full name > Line B
Character A's full name > Line C
...

{Character Profile}
{Character Utterances}
{Generated Logline}
{Generated Plot}

A.2.5 Character Profile Generation(Yuan et al., 2024)

Prompt

The content of the scenario is as follows:
—

{}

—

You are an assistant tasked with extracting character personas. Your task is to write a summary of the character as they appear in this scenario. If a character, location or other key element is mentioned for the first time in the summary, please provide a brief introduction. The story may include a non-linear narrative structure, flashbacks, and shifts between different worlds or perspectives. Therefore, please structure the summary as a coherent narrative following a consistent chronological order. The summary should be no more than words and may consist of multiple paragraphs.

Please output the summary in four specific sections, using the following headings as paragraph headings:

Attributes: // In 50 words or fewer, briefly describe the character's gender, skills, talents, goals and background.
Relationships: // In 50 words or fewer, briefly describe the character's relationships with other characters.
Events: // Organise the major events the character experiences or is involved in chronologically and describe them in 1,000 words or fewer.
Personality: // Briefly describe the character's personality in 50 words or fewer.

Ensure that each section begins explicitly with the specified title, followed by the main text, and include a clear break (line break) between each section.

Summary:

A.3 Prompt Used in Sec. 5

Due to time constraints, we have imposed a constraint on the prompts generated during the experiment to ensure they are concise, so as not to present the participants with overly long scenarios.

A.3.1 Logline Generation

Prompt

Please briefly outline a scenario in which two characters promote the designated product. The requirements for this are set out below.
・Scenario overview (3 lines)
* Please do not include information such as timing or on-screen displays; describe only the actions taken by the characters.
{Character Profile}
{Character Utterances}

A.3.2 Plot Generation

Prompt

Based on the logline provided, please create a single, concise plot for a scenario in which two characters promote a given product, taking the logline into account. The requirements for this are outlined below.
・A plot describing the characters' actions within the scenario (the plot should describe the characters' actions that drive the flow of the scenario, not their dialogue)

【Sample Plot Description Format】Character A does X. Character B does Y. ... * Please do not include information such as timings or on-screen displays; describe only the actions performed by the characters.

【Sample Plot Description Format】
Character A does X. Character B does Y. ...
* Please do not include information such as timings or on-screen displays; describe only the actions performed by the characters.
{Character Profile}
{Character Utterances}
{Generated Logline}

A.3.3 Scenario Generation

Prompt

Please create a scenario of no more than 10 lines of dialogue for two characters promoting a given product, based on the logline and plot provided.
Do not include anything in the output other than the characters' lines and full names. When outputting the characters' full names, do not simply use 'A' or 'B'; please output their full names.

【Sample scenario format】Character A's full name > Line A
Character B's full name > Line B
Character A's full name > Line C
...

{Character Profile}
{Character Utterances}
{Generated Logline}
{Generated Plot}

A.3.4 Feedback Generation

Prompt

Please briefly describe any elements in the submitted advertising script that are inconsistent with the character's profile and dialogue examples, along with the reasons for this. 'Inconsistency' here does not refer to aspects that could make the character more true to form, but rather to whether there are any descriptions of speech or actions that appear to undermine the character's image. If there are no descriptions that are uncharacteristic, please output "None".
Advertising script:
{sentence}
Profiles of each character:
{Character Profile}
Sample dialogue for each character:
{Character Utterances}
Output format:
Comments regarding lack of character consistency:
- Descriptions that are inconsistent with the character:
- Reasons for judging them to be inconsistent:

A.3.5 Feedback Generation (w/ category information)

Prompt

Please briefly describe any uncharacteristic elements in the submitted advertising script, taking into account the character's profile, sample dialogue and evaluation criteria, and explain the reasons for them. 'Uncharacteristic' here does not refer to aspects that could make the character more true to form, but rather to whether there are any descriptions of speech or actions that appear likely to damage the character's image. Please consider whether any uncharacteristic elements exist for each evaluation criterion. If there are no descriptions that are uncharacteristic, please output "None"."
Advertising script:
{sentence}
Profiles of each character:
{Character Profile}
Sample dialogue for each character:
{Character Utterances}
Evaluation Criteria: {Categories}
Output format:
"Comments regarding lack of character consistency:
- Evaluation criterion name:
- Descriptions that lack character authenticity:
- Reasons for judging that the character lacks authenticity:

A.3.6 Scenario Refinement

Prompt

The following advertising script has been provided with comments highlighting aspects that are inconsistent with the character's personality. Please revise the script in accordance with these comments and submit the revised version. Ensure that you address all the points raised in the comments. Furthermore, please ensure that the revised script does not contain more lines of dialogue or longer sentences than the original.
Advertising scenario: {Scenario}
Comments pointing out aspects that are out-of-character: {Feedback}
Profiles of each character:
{Character Profile}
Sample dialogue for each character:
{Character Utterances}

A.4 Relationship between the categories constructed in this study and prior work (Sec. 4.4)

Table 5 shows the relationship between the categories constructed in this study and prior work.

A.5 Questionaries used in Sec. 5

Section 5 reports the analysis results for Q1, the most fundamental question.

Q1. Overall, which scenario is more out-of-character? (Please compare the two scenarios as a whole and choose the one that intuitively feels more out-of-character.)

Q2. Which scenario contains more out-of-character descriptions?

Q3. Which scenario contains the most out-of-character description?

Q4. Which scenario feels more natural as an interaction between the characters?

Q5. Which scenario is more appealing as a character scenario?

Q6. As an advertisement, which scenario makes you more willing to buy the product?

Category name	Explanation
General Categories	
Mismatch with perceived emotional expression or personality	Tu et al. (2024) (Personality Back-Testing), He et al. (2025)(Emotional Intelligence), Iwata et al. (2025) (Personality)
Mismatch with the way things, events, or objects are described, or with the word choice used	Tu et al. (2024) (Persona-Utterance), Wu et al. (2025) (Persona Language Style), Iwata et al. (2025) (Wording)
Mismatch with daily habits and everyday behavior	-
Title-Specific Categories	
Mismatch with values	-
Mismatch with the tone of speech and the form of address used toward a specific character	Iwata et al. (2025)(Habitual Attitudes toward Others)
Mismatch with the character's sentence endings	Iwata et al. (2025) (Wording)
Mismatch with utterance planning and dialogue ability	Tu et al. (2024)(Conversational Ability), He et al. (2025) (Language Fluency)
Mismatch with level of knowledge or proficiency, including the appropriateness of expressions or actions of knowledge or proficiency	Tu et al. (2024) (Know-Exposure, Know-Accuracy and Know-Hallucination), Lu et al. (2024)(Accurate Role-related Knowledge), Wu et al. (2025) (Script-Based Knowledge, Script-Agnostic Knowledge and Script-Contradictory Knowledge)
Individual-Specific Categories	
Mismatch with the sense of relational distance toward a specific conversation partner	Iwata et al. (2025) (Objective Relationships)
Mismatch with the naturalness of how the character's own traits or background are expressed	-
Mismatch with multiple stereotypical attributes	-
Mismatch with the background, world setting, or time period of the world the character belongs	-

Table 5: Relationship between the categories constructed in this study and prior work