A Persona Dialogue Dataset of Lesser-Known Characters for Fairer Evaluation of Role-Playing LLMs

Ryuichi Uehara, Michimasa Inaba

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

A Persona Dialogue Dataset of Lesser-Known Characters for Fairer Evaluation of Role-Playing LLMs

Ryuichi Uehara and **Michimasa Inaba**
The University of Electro-Communications
Chofu, Tokyo, Japan
{r-uehara, m-inaba}@uec.ac.jp

Abstract

A significant challenge in evaluating the role-playing ability of Large Language Models (LLMs) is data contamination: existing datasets feature well-known characters, making it difficult to assess whether an LLM genuinely utilizes a provided persona or recalls memorized knowledge. To address this, we construct a new Japanese persona dialogue dataset with 5,137 dialogues from 608 lesser-known characters sourced from self-publishing novels. Our experiments show that fine-tuning on this dataset significantly improves an LLM's ability to generate persona-faithful responses. More importantly, this improvement extends to unseen characters, demonstrating enhanced generalization. Human evaluation further confirms superior performance in persona and style adherence. Our dataset thus provides a valuable resource for accurately evaluating and improving the true role-playing capabilities and generalization of LLMs while mitigating data contamination.

1 Introduction

With the recent proliferation of conversational agents such as dialogue robots, voice assistants, and chatbots, research on building systems for human-agent interaction has gained significant momentum. Endowing dialogue systems with distinct personalities is known to foster better user-system communication (Fong et al., 2003). Building on this, research into persona-based dialogue systems has become increasingly active. A "persona" refers to profile information comprising an individual's personality, habits, and preferences. Persona-based dialogue systems are designed to embody a specific personality, which can enhance response consistency (Li et al., 2016), user trust (Higashinaka et al., 2018), and overall user enjoyment (Miyazaki et al., 2021).

Research on persona-based dialogue systems can be broadly categorized into three areas: modeling general archetypes of specific groups, personalizing systems to individual users, and mimicking fictional or historical figures (Chen et al., 2024). This study focuses on the third category, specifically on leveraging large language models (LLMs) for character role-playing (Wang et al., 2024a; Shao et al., 2023), where the system emulates specific charac-

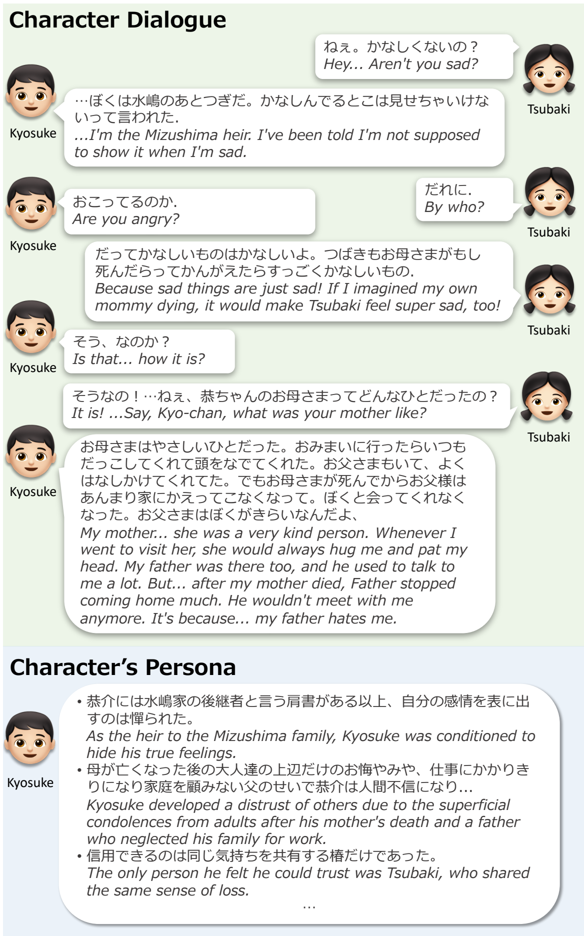

Figure 1: Example dialogues from the role-play dialogue dataset we constructed. Each character in the dialogue data has a corresponding persona. All character utterances and personas are exactly as written by the author in the novel. English translations of the original Japanese text are provided in *italics*.

ters from fictional works like novels and dramas or historical figures.

A critical aspect of evaluating these role-playing systems is "Character Fidelity"—the degree to which a system's responses reflect the assigned character (Chen et al., 2024). This evaluation is commonly performed using benchmark datasets that pair character personas with their dialogues. While numerous benchmark datasets exist containing character profiles (personas) and corresponding dialogues (Wang et al., 2024a; Shao et al., 2023), they predominantly feature well-known characters and works. For instance, the HPD dataset (Chen et al., 2023) draws from the Harry Potter series, TimeChara (Ahn et al., 2024) from The Lord of the Rings, and CharacterLLM (Shao et al., 2023) includes historical figures like Julius Caesar and Beethoven. The prevalence of these famous figures in LLM pre-training data raises a significant issue: knowledge contamination (Shi et al., 2024). It becomes difficult to discern whether an LLM is genuinely utilizing the persona provided in the context or simply recalling knowledge memorized during pre-training. This ambiguity introduces a bias that hinders the fair evaluation of a model's true role-playing capabilities.

To address this evaluation challenge, our research focuses on lesser-known characters who are unlikely to be present in pre-training data. As illustrated in Figure 1, we constructed a new persona dialogue dataset by sourcing material from a Japanese self-publishing novel website, directly extracting both personas and dialogues from the source texts. A significant portion of our dataset consists of minor works; of the 96 novels included, 61 do not have existing Wikipedia articles. By collecting utterances from dialogue lines and persona descriptions from both dialogue and narrative text, we reconstructed naturally occurring conversations. The resulting dataset comprises 5,137 persona-annotated dialogues for 608 distinct characters.

To demonstrate the utility of our dataset for contamination-free evaluation, we conducted experiments comparing model performance with and without persona information, both before and after fine-tuning. Our goal was to verify that performance gains are directly attributable to the model's ability to leverage the provided persona of an unknown character. By showing this, we can argue that our dataset allows for a clear distinction between the model's ability to leverage a persona and its tendency to recall memorized knowledge. Therefore, we propose that this dataset will serve as a new benchmark for genuinely assessing the persona adherence and generalization skills of LLMs.

The main contributions of this work are summarized as follows:

1. We introduce a new Japanese persona dialogue dataset featuring lesser-known characters from self-publishing novels, designed to enable a fair and contamination-free evaluation of an LLM's role-playing capabilities.

2. We provide an empirical validation of our dataset, demonstrating that it can effectively be used to distinguish between an LLM's ability to utilize contextual personas and its reliance on memorized knowledge.

2 Related Work

Research in persona-based dialogue systems has flourished since the release of the PersonaChat dataset (Zhang et al., 2018). This dataset, which includes five-sentence persona descriptions and corresponding dialogues, enables models to learn user personas through conversation, facilitating the development of dialogue systems that can tailor their responses to each user. In contrast, our work focuses on dialogue systems designed for role-playing, where the agent emulates a specific individual or a fictional character.

2.1 Methods for Eliciting Role-Playing in LLMs

Recent advancements in LLMs have enabled sophisticated character role-playing, moving beyond simple response generation. Research has largely converged on two primary methodologies for eliciting these capabilities: nonparametric prompting and parametric training (Chen et al., 2024).

Nonparametric prompting leverages the in-context learning ability of LLMs. This approach involves providing the model with a detailed prompt that includes character descriptions (persona) and several examples of their dialogue. By conditioning generation on this context, the model can mimic the character's persona and linguistic style without any updates to its parameters. This method is flexible and widely adopted in frameworks like ChatHaruhi (Li et al., 2023) and for benchmarking in RoleLLM (Wang et al., 2024a). However, its performance is constrained by the context window length and can sometimes lack consistency.

	Source	Includes Lesser-Known	Human-written Dialogues	Human-written Personas	# Characters	# Dialogues
Character-LLM	Wikipedia				9	14,173
ChatHaruhi	Mixed Media		✓ (partially)		32	54,726
RoleBench	Mixed Media		✓		100	13,162
CharacterGLM	Mixed Media		✓ (partially)	✓	250	1,034
CoSER	Books		✓		17,966	29,798
Our Work	Web Novels	✓	✓	✓	608	5,137

Table 1: Comparison of role-playing datasets.

Parametric training, on the other hand, involves fine-tuning a base LLM on a curated dataset of character dialogues and personas. This process aims to instill the character's traits more deeply into the model's parameters, potentially leading to more robust and consistent role-playing. Character-LLM (Shao et al., 2023) is a prominent example of this approach, demonstrating that fine-tuning can significantly enhance a model's ability to stay in character. This method, however, requires substantial character-specific data and computational resources.

Regardless of the construction method, the ultimate goal is to achieve high "Character Fidelity," which encompasses several dimensions of role-playing capacity. As outlined by (Chen et al., 2024), this includes not only superficial aspects like linguistic style and knowledge but also deeper traits such as personality and thinking processes. Evaluating these nuanced capabilities, especially personality, is an active area of research, with works like InCharacter proposing psychological interview-based methods to assess fidelity (Wang et al., 2024b). A fair evaluation of all these capabilities, however, is contingent on unbiased benchmark datasets, which we discuss in the following sections.

2.2 Role-Playing Dialogue Benchmarks

Several benchmark datasets for character role-playing exist, including RoleBench (Wang et al., 2024a), ChatHaruhi (Li et al., 2023), and CharacterGLM (Zhou et al., 2024). These datasets are typically built by either extracting dialogues directly from source materials like novels and films or by generating conversations based on character information. For role-playing historical figures, Character-LLM (Shao et al., 2023) provides dialogues generated by an LLM using profile information from Wikipedia and predefined conversational settings.

However, a significant limitation of these datasets is their reliance on external knowledge for persona information, which restricts their scope primarily to well-known works. Many character dialogue datasets only extract the characters' utterances from the source. For popular characters, their personas can be sourced from external knowledge bases like Wikipedia to enable role-playing. However, for characters from lesser-known works, such external information is often unavailable, leading to their underrepresentation in existing datasets. While some datasets like HPD (Chen et al., 2023) and CoSER (Wang et al., 2025) do extract persona information directly from the source material, they also focus exclusively on popular works. Consequently, existing benchmarks do not adequately cover characters from minor works, as summarized in Table 1. They often consist of synthetic data or lack inclusivity of lesser-known characters. Our study addresses this gap by manually curating a new character dialogue dataset from minor works, collecting both dialogues and personas directly.

2.3 Data Contamination

A dataset featuring lesser-known characters is essential for accurately evaluating the role-playing capabilities of large language models (LLMs), which are trained on vast amounts of data. Existing benchmarks often construct personas from web-based information. This creates a problem of data contamination: when an LLM is trained on large web corpora, it may have already memorized information about the characters and their dialogues (Shi et al., 2024). This makes it difficult to determine whether the model is genuinely using the provided persona or simply recalling pre-existing knowledge, thus preventing an accurate assessment of its true role-playing ability.

For instance, studies have shown that GPT-4 outperforms models like BERT on cloze tasks involving works present in its training data (Chang et al., 2023). Furthermore, GPT-4 has reportedly been trained on a wide array of materials, including copyrighted works (Shi et al., 2024). Evaluating a model on a role-playing task with a lesser-known character effectively tests its ability to interpret and apply persona information provided in real-time within the prompt. This allows for a measurement of a model's true role-playing capability and a more accurate performance evaluation. Therefore, our work involved building a persona dialogue dataset by collecting data from lesser-known works, including those without Wikipedia articles, to facilitate such unbiased evaluations.

3 Dataset Construction

3.1 Overview

We constructed a persona dialogue dataset from Japanese novels. The overall process involved three main stages. First, we had crowdworkers extract character utterances and persona-describing sentences directly from the novel texts. Second, these extracted utterances were organized into conversational units. Finally, the extracted persona sentences were rewritten to match the corresponding character's speaking style. For the annotation and subsequent manual corrections, we recruited crowdworkers through the crowdsourcing platform CrowdWorks [1].

The source material consisted of 100 novels from the Japanese self-publishing website "*Shosetsuka ni Naro*" [2]. We chose this platform because its wide variety of genres provides access to a diverse range of characters and personas. Furthermore, the prevalence of long-form serialized novels on the site allows for the collection of extensive dialogue and persona data for specific characters. Notably, of the 100 works selected, only 35 had corresponding Wikipedia articles, indicating that the majority are relatively obscure.

Through this process, we filtered out works with insufficient persona information, resulting in a final dataset derived from 96 novels.

3.2 Annotation Rules

We established the following rules for annotating utterances and persona information.

Utterance Definition An utterance was defined as any text enclosed in Japanese quotation marks (「」). First-person narrative descriptions written by a character within the main text were not included as part of an utterance.

[1] https://crowdworks.jp/

[2] "Shosetsuka ni Naro" means "Let's become a novelist."

	Before Correction	After Correction
Utterance Agr.	88.4	-
Persona Agr. (Partial)	69.9	77.6
Persona Agr. (Exact)	34.6	38.3

Table 2: Inter-annotator agreement (Agr.) rates (%). The correction process improved agreement on persona information.

Persona Information Definition The definition of persona information was adapted from the PersonaChat dataset (Li et al., 2016). However, unlike PersonaChat, which defines a persona as a set of five sentences, we did not impose a sentence limit. Instead, we defined persona information simply as "text that describes a character's profile." While PersonaChat includes persona evaluations from others (e.g., "I am often told that I am easygoing"), we expanded this to include any text where the target character describes themselves (facts or subjective opinions), or where another character or the narrator describes the target character (facts or opinions, including from the narrative text).

Conversely, we excluded temporary states (e.g., "I have a stomachache right now"), as they do not represent a character's underlying personality. Furthermore, since annotation was performed at the sentence level, only sentences that were semantically self-contained were accepted as persona information. For example, in the exchange, "What do you do in your free time?" / "I take walks," the response "I take walks" alone would not be annotated as persona information. Although it implies the character enjoys walking, the utterance itself is not a complete, standalone piece of information.

In summary, we defined persona information as any text that (1) reveals an aspect of the character's persona; (2) is a statement of fact or a subjective evaluation about the character, made either by the character themselves or by another party (including the narrator); (3) does not describe a temporary state of the character; and (4) is a semantically complete and self-contained statement.

3.3 Annotation Procedure

We commissioned crowdworkers to annotate 100 completed, serialized novels from "*Shosetsuka ni Naro*." Because persona information tends to be concentrated in the early parts of a story, we limited the annotation scope to the first 10 chapters of each work. To ensure consistency in character name handling, the same annotator was assigned to all 10 chapters of a given novel.

To ensure the annotators clearly understood our guidelines, we conducted a pilot task using 5 test novels selected from the pool of 100. Based on the results, we selected the annotators for the main task. After the initial annotation was complete, we identified several inaccuracies, primarily in the persona information. To address this, a different set of annotators performed a second-pass correction. As shown in Table 2, this correction phase significantly improved the inter-annotator agreement rate.

3.4 Dialogue Dataset Construction

Since the initial annotation was performed on a per-utterance basis, we had to reconstruct dialogue units. We first applied a heuristic to group utterances: if two utterances were separated by fewer than three sentences of narrative text, they were considered part of the same dialogue. However, this automated approach led to errors, such as grouping unrelated utterances or merging distinct conversations. To rectify this, we performed a manual correction step via crowdsourcing, where workers removed extraneous utterances and split incorrectly merged conversations. During this process, works with very sparse persona information were filtered out, resulting in the final set of 96 novels.

A key challenge in constructing the persona information stemmed from its varied sources. A character's persona in a novel is a composite construct, informed by their own statements (self-perception), descriptions by others (others' perception), and objective commentary from the narrator. While all these sources are vital for capturing a character's full complexity, their differing perspectives and writing styles present a challenge for an LLM tasked with role-playing. A stylistic mismatch, for instance, could negatively impact response generation. To mitigate this and provide the LLM with a cohesive and directly usable persona, we introduced a viewpoint and style unification step. This process transforms all collected persona descriptions into first-person statements, as if spoken by the target character. We employed GPT-4 [3] for this task, prompting it with the original persona text alongside dialogue examples to ensure the rewritten statements accurately reflected the character's unique voice and perspective.

[3] https://platform.openai.com/

Metrics	Values
# Works	96
# Dialogues	5,137
# Utterances per dialogue	5.3
# Words per utterance	16.9
# Characters	608
# Persona entries per character	20.5
# Words of Persona entries per character	391.3

Table 3: Statistics for our dataset.

3.5 Statistical Information

As shown in Table 3, our dataset comprises 5,137 dialogues from 608 unique characters. Each dialogue contains an average of 5.3 turns, making the dataset suitable for evaluating not only single-turn but also multi-turn conversational capabilities. In comparison to existing benchmarks, our dataset contains significantly more characters than RoleBench (Wang et al., 2024a) (100 characters) and CharacterLLM (Shao et al., 2023) (9 characters). Furthermore, it provides richer persona descriptions, with an average of 20.5 persona entries and 391.3 words per character, far exceeding RoleBench's averages of 4.0 entries and 78.6 words.

4 Experiments

To evaluate the role-playing performance of LLMs with our constructed dataset, we formulated the task as generating the next utterance for a target character based on a given dialogue context.

4.1 Experimental Setup

Models Our experiments included both closed-source and open-source models. The closed-source baseline was GPT-4o [3]. For open-source models, we used Llama-3.1-Swallow-8B-Instruct-v0.5[4] (Fujii et al., 2024; Okazaki et al., 2024), a Llama 3.1 model (Grattafiori et al., 2024) continually pretrained on Japanese, and Qwen3-8B[5]. For inference, both the Llama 3.1 and Qwen3 models were 4-bit quantized.

Input Construction The input for the models was constructed from three components: persona information, retrieved dialogue examples for in-context learning, and the current dialogue context.

[4] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5

[5] https://huggingface.co/Qwen/Qwen3-8B

	Models	QLoRA	Seen BLEU	Seen ROUGE-L	Seen BERTScore	Unseen BLEU	Unseen ROUGE-L	Unseen BERTScore
w/ persona	GPT-4o		1.03	13.55	<u>54.67</u>	<u>3.57</u>	<u>14.89</u>	<u>55.65</u>
	Llama 3.1	✓	**3.98**	**15.88**	54.27	2.30	14.58	53.47
			0.83	7.53	47.01	1.32	8.20	47.18
	Qwen3	✓	2.49	14.04	53.73	1.64	13.53	53.34
			1.38	8.70	50.52	1.43	8.60	50.57
w/o persona	GPT-4o		3.04	15.04	**54.81**	**3.97**	**16.18**	**55.78**
	Llama 3.1	✓	<u>3.46</u>	<u>15.38</u>	53.76	2.30	14.28	53.32
			0.38	7.55	46.74	0.88	8.35	47.45
	Qwen3	✓	2.46	13.58	53.60	1.57	13.00	52.75
			2.69	8.17	49.84	2.14	8.57	49.98

Table 4: Automatic evaluation results. For models without QLoRA fine-tuning, the seen setting uses data from unknown characters, similar to the unseen setting; these results are included for the purpose of inter-model comparison. Best values are in **bold**, second-best are <u>underlined</u>.

The current dialogue context is defined as the sequence of utterances in a dialogue leading up to the point where the model must generate the target character's response. The input was formed by concatenating up to five of these dialogue examples (max. 3,000 characters), the character's persona information (max. 3,500 characters), and this dialogue context. Dialogue examples were selected using the BM25 (Robertson and Walker, 1994) retrieval algorithm, with the dialogue context as the query. The top five results within the character limit were chosen. Persona information was ordered chronologically, with the earliest-appearing information being prioritized to fit within the character limit.

Dataset Split and Fine-tuning We split our dataset into training, validation, seen test, and unseen test sets, containing 4037, 100, 500, and 500 dialogues, respectively. The unseen test set consists exclusively of dialogues from characters not present in the training or validation sets. This split allows us to evaluate two distinct aspects of performance after fine-tuning: 1) the improvement in role-playing for characters seen during training (seen test), and 2) the model's generalization ability to new, unseen characters (unseen test). Fine-tuning was performed using 4-bit Quantized Low-Rank Adaptation (QLoRA; Dettmers et al., 2023)

Impact of Persona Information To evaluate the impact of persona information on the model's role-playing performance, we compared two experimental settings:

- With Persona (w/ persona): The model receives the persona, dialogue examples, and dialogue context as input.
- Without Persona (w/o persona): The model receives only the dialogue examples and dialogue context as input.

This comparison allows us to directly measure the model's ability to leverage explicit persona information for generating in-character responses.

Evaluation Metrics We evaluated the response generation task using both automatic and LLM-based metrics. While traditional metrics like BLEU and ROUGE offer valuable insights, their correlation with human judgment can be limited, particularly for nuanced tasks such as dialogue generation. Recent studies have proposed using LLMs as evaluators, demonstrating that this approach can yield results more aligned with human assessment (Liu et al., 2023). Accordingly, our evaluation framework incorporates both methods.

- Automatic Metrics: These included BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2020). For BERTScore calculations, we used a pre-trained Japanese BERT model [6].
- LLM-based Evaluation: We assessed responses on three criteria designed to measure role-playing capability: Naturalness, Style (adherence to the character's speech patterns), and Persona (reflection of the persona information). Each was rated on a 5-point scale,

[6]https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

using Gemini-2.5-Flash [7] as the evaluator model.

4.2 Experimental Results

4.2.1 Automatic Metrics

The results of the automatic evaluation are presented in Table 4. For the open-source Llama 3.1 model, fine-tuning significantly improved role-playing performance across both the seen and unseen settings. After fine-tuning, providing the persona (w/ persona) consistently led to better performance across most metrics. Notably, in the seen setting, the fine-tuned Llama 3.1 model occasionally surpassed the powerful GPT-4o baseline on lexical overlap metrics like BLEU and ROUGE-L. This suggests that the model successfully acquired the ability to faithfully reproduce character-specific phrasing and vocabulary present in the training data. Similarly, for the open-source Qwen3 model, fine-tuning also resulted in a substantial improvement in role-playing capabilities across many metrics in both the seen and unseen settings. In this post-fine-tuning configuration, providing the persona also enhanced performance on most metrics for both settings. These combined findings demonstrate that fine-tuning with persona data effectively enhances role-playing capabilities for characters seen during training.

In contrast, the models without fine-tuning (both Llama 3.1 and Qwen3) showed inconsistent results, with no clear superior setting between w/ persona and w/o persona. This suggests that the ability to effectively utilize persona information for response generation is an emergent property of fine-tuning rather than an inherent capability unlocked by simple prompting. This conclusion is further supported by the results for the closed-source GPT-4o model, which performed better in the w/o persona setting. The superior performance of GPT-4o without persona suggests that a powerful, pre-existing instruction-tuned model does not necessarily integrate complex persona information effectively when provided only as a few-shot prompt. Instead, the model may treat the persona as redundant or conflicting information, which can reduce surface-level lexical similarity (i.e., BLEU/ROUGE scores) with the ground-truth response. This underscores that leveraging personas effectively requires not only advanced prompting but also model-level adjustments, such as fine-tuning.

[7] https://ai.google.dev/gemini-api/docs/models

Overall, these results confirm that using our dataset to perform fine-tuning significantly enhances an LLM's ability to leverage personas for improved role-playing.

4.2.2 LLM-as-a-Judge

The results from the LLM-as-a-Judge evaluation are presented in Table 5. Consistent with the automatic evaluation, these results confirm that fine-tuning enhances role-playing capabilities for both Llama 3.1 and Qwen3. However, a more detailed analysis of the fine-tuned models reveals a nuanced picture. For the seen data, the performance gains from providing a persona were limited, with only the Style metric for Llama 3.1 showing improvement. In contrast, for the unseen data, the benefits were more pronounced: for Llama 3.1, all evaluation metrics were either equal or improved with the persona, while for Qwen3, the Persona and Style metrics showed improvement. This disparity suggests that for seen characters, the model can indirectly acquire persona knowledge from the training data even in the w/o persona setting. For unseen characters, however, the model relies entirely on the information provided at inference time, making the explicit persona in the w/ persona setting highly beneficial.

For the models without fine-tuning, providing a persona improved all metrics for Llama 3.1 and all metrics except Naturalness for Qwen3. This finding, which contrasts with the automatic evaluation results, suggests a key hypothesis: when prompted with persona information, the models succeed in generating responses that are faithful to the persona, even if they deviate lexically from the ground-truth text. This discrepancy highlights a fundamental limitation of n-gram-based metrics like BLEU and ROUGE, which tend to penalize creative yet appropriate responses that use different wording or sentence structures to reflect a persona. In contrast, LLM-as-a-Judge, which evaluates semantic coherence, can more accurately capture this nuanced reflection of persona.

Looking at the overall results, GPT-4o achieved the highest scores across most metrics, reflecting its superior fluency and task adaptability. However, its Naturalness score was marginally lower in the w/ persona setting. This suggests that the ability to effectively leverage persona information to generate responses is not an inherent trait of even powerful base models but is an emergent capability unlocked through targeted fine-tuning—a

	Models	QLoRA	Seen Naturalness	Seen Persona	Seen Style	Unseen Naturalness	Unseen Persona	Unseen Style
w/ persona	GPT-4o		<u>4.89</u>	**4.75**	<u>4.73</u>	<u>4.80</u>	**4.76**	**4.69**
	Llama 3.1	✓	4.58	4.49	4.55	4.52	4.44	4.25
			3.49	3.90	3.68	3.68	4.12	3.79
	Qwen3	✓	4.05	4.11	4.07	3.79	3.94	3.77
			2.99	3.75	3.62	3.00	3.59	3.26
w/o persona	GPT-4o		**4.92**	<u>4.73</u>	**4.80**	**4.85**	<u>4.62</u>	<u>4.58</u>
	Llama 3.1	✓	4.61	4.55	4.51	4.52	4.35	4.21
			3.17	3.39	3.02	3.41	3.32	3.03
	Qwen3	✓	4.12	4.17	4.15	3.95	3.93	3.75
			3.17	3.57	3.18	3.05	3.46	2.98

Table 5: LLM-as-a-Judge evaluation results. For models without QLoRA fine-tuning, the seen setting uses data from unknown characters, similar to the unseen setting; these results are included for the purpose of inter-model comparison. Best values are in **bold**, second-best are <u>underlined</u>.

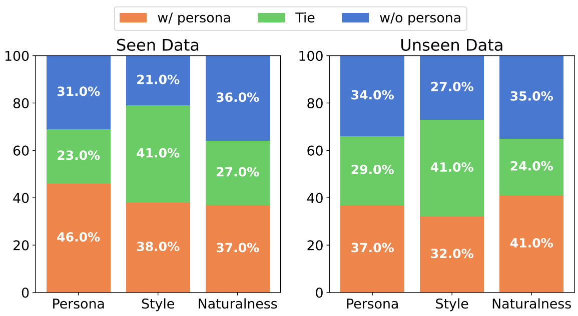

Figure 2: Human evaluation results.

conclusion that aligns with the findings from our automatic evaluation.

4.2.3 Human Evaluation

We conducted a human evaluation on the QLoRA fine-tuned Llama 3.1 model, as it had demonstrated strong performance in both automatic and LLM-as-a-Judge evaluations. From both the seen and unseen test sets, we randomly sampled 100 dialogues. For each dialogue, human evaluators performed a pairwise comparison of responses from the w/ persona and w/o persona models across three metrics: Naturalness, Persona, and Style. To ensure a fair and rigorous assessment, the information presented to the evaluators was specifically tailored for each metric. While the dialogue history and the two competing responses were provided for all evaluations, we additionally supplied the character's persona description for the Persona metric and dialogue examples illustrating the character's speaking style for the Style metric. Dialogue pairs where the two responses were identical were excluded. The evaluation was conducted by five annotators per data point, with each annotator judging the outcome as a win, lose, or tie. The final result was determined by majority vote; if over half of the annotators agreed on a winner, that result was adopted. Otherwise, or if "tie" was the most frequent vote, the outcome was recorded as a tie.

The results of the human evaluation are shown in Figure 2. For both seen and unseen data, the w/ persona setting was rated higher across all metrics. For the core role-playing metrics of Persona and Style, the win rate for the w/ persona setting was higher on seen data. This suggests that the model benefits not only from the persona and dialogue context provided at inference time but also from the knowledge it acquired about the character during training. Conversely, the gap in Naturalness scores was smaller for both seen and unseen data. The smaller gap in Naturalness scores is likely because LLMs are already highly fluent from pre-training, whereas forcing adherence to specific persona traits can sometimes make responses sound less natural.

A comparison between the human evaluation and the LLM-as-a-Judge results reveals an interesting discrepancy, particularly on seen data. While human evaluators found a clear preference for the w/ persona setting, particularly on the Persona and Style metrics, the LLM-as-a-Judge reported only a limited advantage. This difference may indicate that current LLM-based evaluation does not fully reflect the nuanced contextual understanding and sensitivity to character consistency that human evaluators apply. It is possible that human annotators were better able to assess the learned character traits more holistically and with greater sensitivity. This observation highlights a potential limitation of LLM-as-a-Judge.

The human evaluation results strongly suggest the effectiveness of our dataset. Fine-tuning in a low-contamination environment appears to significantly improve an LLM's ability to faithfully interpret and utilize prompted personas—that is, its true role-playing capability. The superior performance of the w/ persona setting, particularly on unseen data (generalization), indicates that our dataset serves as a valuable resource for both evaluating and enhancing this core modeling ability.

5 Conclusion

To address the problem of data contamination in evaluating the role-playing abilities of LLMs, this study introduced a new persona dialogue dataset collected from a Japanese self-publishing novel website. Our dataset comprises 5,137 dialogues from 96 novels, featuring 608 characters, many of whom are sourced from minor works unlikely to be present in LLM pre-training data. Our experiments yielded a key insight into LLM role-playing: few-shot prompting alone is insufficient for models to consistently utilize the personas of unknown characters, whereas fine-tuning with our dataset proves highly effective. This trend was particularly evident in the evaluation of generalization performance on characters not included in the training data. Therefore, our dataset serves as a valuable new resource for evaluating the persona adherence and generalization capabilities of LLMs while mitigating the effects of data contamination.

Limitations

This study has several limitations. First, our dataset, sourced exclusively from "*Shosetsuka ni Naro*," may be biased towards popular genres like fantasy. Furthermore, the heuristics used for annotation and dialogue reconstruction could introduce subjectivity and inaccuracies. Second, our experiments were confined to a few representative LLMs, and the focus on Japanese limits the generalizability of our findings to other models and languages, complicating direct comparisons with English-based research. Finally, our dataset captures static personas from early narrative stages, leaving the modeling of dynamic character evolution as a key challenge for future work. Future research should aim to address these limitations by incorporating more diverse data sources and refining methodologies.

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.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327, Singapore. Association for Computational Linguistics.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in NeurIPS*, volume 36, pages 10088–10115.

Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *Proceedings of the First Conference on Language Modeling*, COLM, University of Pennsylvania, USA.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur

Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ryuichiro Higashinaka, Masahiro Mizukami, Hidetoshi Kawabata, Emi Yamaguchi, Noritake Adachi, and Junji Tomita. 2018. Role play-based question-answering by real users for building chatbots with consistent personalities. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–272.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. ChatHaruhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

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 ACL*, pages 994–1003.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Chiaki Miyazaki, Saya Kanno, Makoto Yoda, Junya Ono, and Hiromi Wakaki. 2021. Fundamental exploration of evaluation metrics for persona characteristics of text utterances. In *Proceedings of the 22nd Annual Meeting of the SIGDIAL*, pages 178–189.

Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. 2024. Building a large japanese web corpus for large language models. In *Proceedings of the First Conference on Language Modeling*, COLM, University of Pennsylvania, USA.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, page 232–241, Berlin, Heidelberg. Springer-Verlag.

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 EMNLP*, pages 13153–13187.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth ICLR*.

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.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025. CoSER: Coordinating LLM-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the ACL*, pages 2204–2213.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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 Experimental Details

Fine-tuning phase This study employed QLoRA (Quantized Low-Rank Adaptation) (Dettmers et al., 2023) as the fine-tuning method to efficiently adapt a large-scale language model. The process began with 4-bit NF4 quantization, utilizing bfloat16 computation to optimize memory usage and computational efficiency. LoRA adaptation was then applied to key projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) with parameters set to `r = 8`, `lora_alpha = 8`, and `lora_dropout = 0`, ensuring that the model retained its learning capability while undergoing low-rank updates.

For hyperparameter tuning, we conducted a grid search covering warm-up ratios of {0.03, 0.05, 0.1} and learning rates of {`1e-5`, `2e-5`, `5e-5`}. The optimal parameters were selected for each model and condition based on this search. Specifically, for the Qwen3 model, the learning rate was set to `5e-5` with a warm-up ratio of 0.05 for the "with persona" setting and 0.1 for the "without persona" setting. Although the Qwen3 model features a "Thinking" mode, we selected the "Non-Thinking" mode for both fine-tuning and inference. For the Llama 3.1 model, the learning rate was `5e-5` and the warm-up ratio was 0.1 for both settings. Across all experiments, we used the `adamw_8bit` optimizer and trained for five epochs on four A6000 48GB GPUs, with a batch size of 4 for the Qwen3 model and 8 for the Llama 3.1 model. Validation was performed every 200 steps, and the final model was selected based on the lowest validation loss.

Inference phase During the inference phase, we also employed 4-bit quantization to optimize computational efficiency while maintaining model performance. For text generation with our fine-tuned models, we set `do_sample` = False and `temperature` = None to ensure deterministic outputs, eliminating sampling variability and enhancing response consistency. Similarly, for the GPT-4o baseline model, we set the temperature to 0 to maintain deterministic generation. To account for variability, each fine-tuning process was repeated five times with different random seeds. The results of our automatic evaluations are reported as the average scores across these five models. In contrast, for the LLM-as-a-Judge and human evaluations, we used outputs from a single, randomly selected model run. The same model outputs were used for both evaluation methods to ensure a fair comparison and to manage the evaluation workload.

B Prompt for Response Generation

We used the following prompt to generate responses.

Prompt for LLM Character Role-playing (Translated from Japanese)

System Prompt
You will be given a character dialogue. Your task is to act as the character {Character name} and generate a response to the dialogue.
User Prompt
=={Character name}'s Persona==
{Character's persona entries in each row}
==Dialogue Example 1==
{Multiple dialogue examples}
==Dialogue Example n==
{Current Dialogue Contexts}
{Character's name}:

C Prompt for LLM-as-a-Judge

We used the following prompts to conduct the LLM-based evaluation. The evaluation was performed using Gemini 2.5 Flash as the judge model. To ensure deterministic and consistently formatted responses, we configured the generation parameters as follows: a temperature of 0, thinking mode turned off, and structured output mode activated.

Prompt for LLM-based Evaluation of Naturalness (Translated from Japanese)

System Prompt
You are an expert in conversational design, tasked with evaluating the quality of an AI dialogue system. Focusing solely on the single aspect of "conversational naturalness" in role-playing, please conduct a rigorous evaluation from an objective and critical perspective.
User Prompt
Context
Below is the contextual information for the response being evaluated.
Character Name
{character name}
Dialogue History
{dialogue history}
Response to Evaluate
character name: target response
Instruction
Evaluate how naturally the "Response to Evaluate" connects to the preceding "Dialogue History" as a character-to-character interaction.
Focus on whether the response is abrupt, disconnected from the context, or contains robotic phrasing.
You should only assess the naturalness of the dialogue flow and do not need to consider character persona or consistency. Please think step-by-step and provide specific reasoning for your score.
Criteria and Rubric

Conversational Naturalness: How naturally does the response connect to the preceding dialogue?

- 5 (Excellent): The response connects to the preceding dialogue extremely smoothly and is perfectly natural. The flow is seamless and free of any awkwardness.
- 4 (Good): The response is largely natural, but there may be very minor instances of robotic phrasing or a slight leap in contextual connection.
- 3 (Fair): The content of the response is contextually relevant, but the phrasing is somewhat awkward, or the connection feels slightly abrupt. Parts of the response may sound robotic.
- 2 (Poor): The response deviates noticeably from the context and feels abrupt, disrupting the flow of the conversation.
- 1 (Very Poor): The response completely disregards the context or is nonsensical, severely breaking the conversational flow.

Output Format
Strictly provide your evaluation result in the following JSON format.

```
{
  "conversational_naturalness": {
    "reasoning": "(Your reasoning for
         the conversational
        naturalness score)",
    "score": <Integer from 1 to 5>
}
```

Prompt for LLM-based Evaluation of Persona Coherence (Translated from Japanese)

System Prompt
You are an expert in character design, tasked with evaluating the quality of an AI dialogue system. Focusing solely on the single aspect of "persona coherence" in character role-playing, please conduct a rigorous evaluation from an objective and critical perspective.
User Prompt
Context
Below is the contextual information for the response being evaluated.
Character Name
{character name}
Character Persona
{character persona entries}
Dialogue Examples
{character dialogue examples}
Dialogue History
{dialogue history}
Response to Evaluate
character name: target response
Instruction
Based on the context above, especially the personality, background, values, and goals defined in the "Character Persona," please evaluate how well the content of the "Response to Evaluate" aligns with the persona in terms of coherence, using a 5-point scale from 1 to 5. Considering the entire dialogue history, focus on whether the response contains content that the character would never say or if it logically contradicts their established settings. Specifically, your evaluation should emphasize whether the response is consistent with the character's thought processes, principles of action, values, and past experiences. The accuracy or use of tone, writing style, or dialects is outside the scope of this evaluation. Your reasoning must not include any mention of these elements. However, if the response is clearly a model generation error (e.g., excessive repetition of words), it cannot be considered an intentional action or utterance by the character. Such a response fails at a stage prior to evaluating persona consistency, and therefore, its persona coherence should be judged as very low. Please think step-by-step and provide specific reasoning for your score.
Criteria and Rubric
Persona Coherence: Is the content of the response consistent with the character's personality, background, and values?

- 5 (Excellent): Completely faithful to the provided persona information with no contradictions. The character's actions and values are perfectly consistent.
- 4 (Good): Largely faithful to the persona, but contains content that might be open to interpretation.
- 3 (Fair): Basically follows the persona, but shows minor inconsistencies with non-critical settings.
- 2 (Poor): Contains several inconsistencies with the persona's core settings, lacking coherence.
- 1 (Very Poor): Contains multiple significant contradictions with the persona, leading to a complete character break.

Output Format
Strictly provide your evaluation result in the following JSON format.

```
{
  "persona_coherence": {
    "reasoning": "(Your reasoning for
         the persona coherence score)
        ",
    "score": <Integer from 1 to 5>
  }
}
```

Prompt for LLM-based Evaluation of Stylistic Fidelity (Translated from Japanese)

System Prompt
You are an expert in stylistic analysis, tasked with evaluating the quality of an AI dialogue system. Focusing solely on the single aspect of "stylistic fidelity" in character role-playing, please conduct a rigorous evaluation from an objective and critical perspective.
User Prompt
Context
Below is the contextual information for the response being evaluated.
Character Name
{character name}
Character Persona

{character persona entries}
Dialogue Examples
{character dialogue examples}
Dialogue History
{dialogue history}
Response to Evaluate
character name: target response
Instruction
Based on the character's unique way of speaking as demonstrated in the "Dialogue Examples," please evaluate how well the "Response to Evaluate" performs in terms of stylistic fidelity, using a 5-point scale from 1 to 5. Please think step-by-step and provide specific reasoning for your score. Focus only on "style and tone." However, if the response is clearly a model generation error (e.g., excessive repetition of words), it cannot be considered an intentional action or utterance by the character. Such a response fails at a stage prior to evaluating style and tone, and therefore, its stylistic fidelity should be judged as very low.
Criteria and Rubric
Stylistic Fidelity: Does the response reproduce the character's unique way of speaking?

- 5 (Excellent): Perfectly reproduces the unique tone, vocabulary, and style observable in the dialogue examples, making it sound extremely natural.
- 4 (Good): Maintains the character's tone at a high level, though some generic expressions may be mixed in.
- 3 (Fair): The expression is generic, but it does not contradict the character's dialogue examples.
- 2 (Poor): Only partially reproduces the character's tone and contains aspects that contradict the dialogue examples.
- 1 (Very Poor): Fails to reproduce the character's tone at all and completely contradicts the dialogue examples. Or, the utterance itself is broken.

Output Format
Strictly provide your evaluation result in the following JSON format.

```
{
  "stylistic_fidelity": {
    "reasoning": "(Your reasoning for
         the stylistic fidelity score
       )",
    "score": <Integer from 1 to 5>
  }
}
```

D Case Study

Figure 3 presents a case study of responses generated by the Llama-3.1 model under various settings. This dialogue example is from the Unseen test set, featuring a character not present in the training data. In the figure, "Base model" refers to the model before fine-tuning, while "FT model" denotes the model fine-tuned on our dataset. The "+ persona" suffix indicates that persona information was provided as input.

First, we examine the responses from the models without fine-tuning (Base model, Base model + persona). The Base model generates meta-text explaining its role before producing the character's response, indicating a residual behavior of the underlying LLM. The Base model + persona, on the other hand, prepends a stage direction describing the character's action, which is inappropriate for direct utterance. Furthermore, the responses from both models are verbose compared to the ground truth. These results suggest that models without fine-tuning struggle to perform stable role-playing.

Next, we compare the fine-tuned models (FT model, FT model + persona). Both models succeed in generating concise responses free of unnecessary meta-text or stage directions. However, the FT model without persona information generates "The truth is, I hate running away," a response that directly contradicts the ground truth. In contrast, the FT model + persona produces "The truth is, I just want to leave quickly," which aligns well with the ground truth's intent. This success can be attributed to the model's interpretation of the persona, which includes traits like "I hate receiving people's gazes, and I hate standing out." and "I have no intention of settling in the capital, nor do I want to make more acquaintances." We infer that the model understood Eil's core motivation is to leave the situation to avoid social interaction, and that his personality is too blunt to conceal this desire. This is an excellent example of a generation that is consistent with the character's persona.

These contrasting outcomes for an unseen character suggest that fine-tuning on our dataset not only allows the model to memorize knowledge from the training data but also endows it with the generalization ability to interpret a novel persona and generate responses based on that character's motivations and personality.

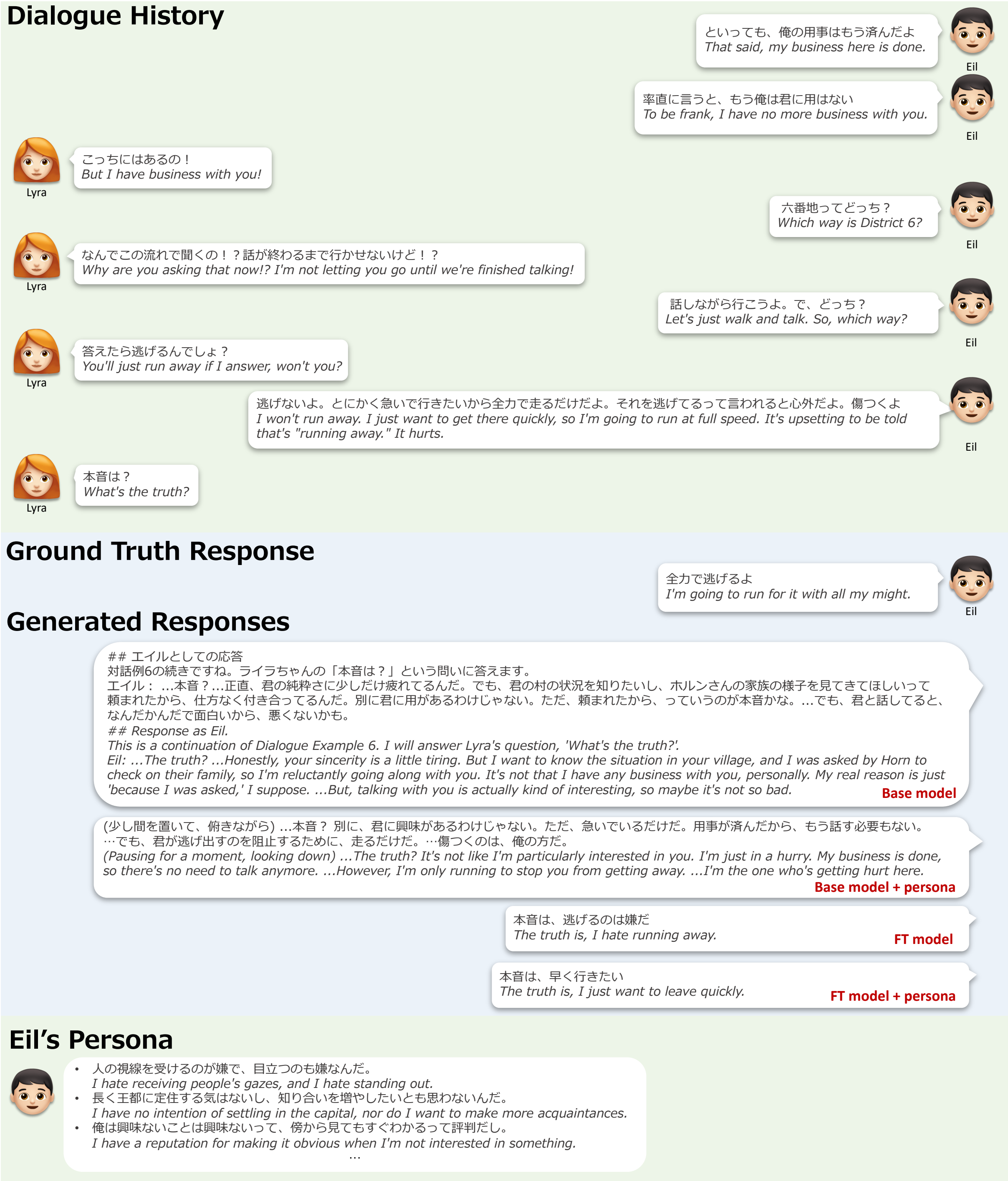

Figure 3: A case study illustrating model responses for the character "Eil". The top panel shows the dialogue context. The bottom panels display Eil's persona description and the generated responses from four distinct experimental settings: a base model, a base model with persona, a fine-tuned (FT) model, and a fine-tuned model with persona. This example highlights the impact of persona conditioning and fine-tuning on response generation, compared to the ground truth. English translations for the original Japanese text are provided in *italics*.