Development of an Evaluation System for a Fan-Engagement Chat Application Using LLM-as-a-Judge

Yuki Fujita[1], **Yasunobu Sasaki**[2], **Ryota Arashi**[2], **Hokuto Ototake**[1] **and Shinya Takahashi**[1]

[1]Fukuoka University, Japan [2]OSHIAI Co., Ltd., Japan

Correspondence: td252012@cis.fukuoka-u.ac.jp

Abstract

To address challenges in objectivity and efficiency in evaluating the quality of generative AI chatbots, we developed an automatic evaluation framework using the "LLM-as-a-judge" approach. A User Simulator, built with In-Context Learning and LoRA tuning, was employed to generate pseudo-conversation logs of the fan-engagement application OSHIAI. These logs were then automatically evaluated by a Judge LLM across six dimensions, and the contribution of this method to quality management in real-world services was verified.

1 Introduction

In recent years, with the advancement of generative AI technology, character-based chatbots have become widespread. These bots imitate the style, tone of voice, and values of fictional characters or real people, and interact with users in natural language. However, there are few real-world examples of objectively and reproducibly quantifying aspects such as "character-likeness," usefulness, and safety of such dialogues, and practical deployments in applications remain limited.

In character-based dialogues, multifaceted evaluation is crucial, including (1) persona consistency, (2) appropriate information provision, and (3) suppression of responses to inappropriate input. Nevertheless, comprehensive manual evaluation is costly and poses challenges in scalability and reproducibility.

As a promising approach to this challenge, LLM-as-a-judge, which uses a large language model (LLM) as an evaluator, has been proposed(Zheng et al., 2023). While it is expected to reduce evaluation costs and improve scalability, correction methods to bridge the gap with human evaluation(Teshima et al., 2025) and frameworks for automatically generating evaluation aspects(Nishikawa et al., 2025) are also being researched.

Recently, LLM-enabled evaluation has been actively studied across dialogue settings, either as an automatic scoring mechanism or as a source of evaluation signals for iterative improvement. For example, Wang et al. introduce critic guidance for open-domain dialogues, where an LLM-based critic scores responses from multiple perspectives and the feedback is used to steer response regeneration and data construction toward user-oriented proactivity(Wang et al., 2025). In task-oriented dialogue, AutoEval-ToD presents an end-to-end evaluation pipeline that leverages a scenario-driven user simulator and generates multi-aspect evaluation reports using not only utterances but also internal states (metadata)(Jain et al., 2025). Taken together, these studies suggest an emerging practice of scaling evaluation (and related development cycles) by generating dialogue logs through user-imitating agents/simulators rather than relying solely on manual data collection(Wang et al., 2025; Jain et al., 2025). In addition, persona-specific evaluation is advancing; PersonaGym proposes dynamic, persona-relevant environments and automated multi-task evaluation for measuring persona adherence of LLM agents(Samuel et al., 2025).

Dialogue data are required for evaluation, but large-scale manual preparation is impractical. Therefore, having a User Simulator interact with the Target Chatbot to synthetically generate dialogue logs is an effective approach(Ueda and Takayanagi, 2025).

In this study, we aim to establish a reproducible and scalable evaluation workflow for fan-engagement character chatbots that can be integrated into real-world application development and operation. Specifically, we construct a quantitative evaluation framework using LLM-as-a-judge for the character chatbots within the AI partner application OSHIAI[1]. We generate pseudo-logs through

[1]https://oshi-ai.com/

Proceedings of the 16th International Workshop on Spoken Dialogue Systems Technology, pages 123–127
February 26–March 1, 2026.

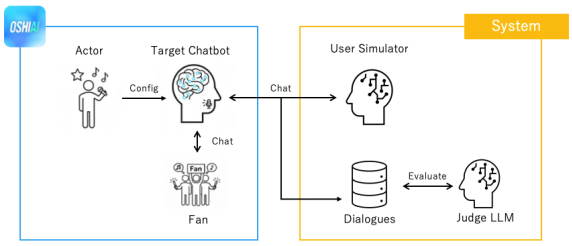

Figure 1: Overall structure of the evaluation system

interaction with a User Simulator and automatically evaluate them from multiple perspectives relevant to character chat, including tone/persona realization, diversity, and human-likeness. The overall structure of the system is shown in Fig.1, and the details will be explained below.

In this research, we only used a pre-trained LoRA created from information that was anonymized by the commercial service provider, based on dialogue logs generated within that commercial service and managed by the same provider.

2 About the AI Partner Application OSHIAI

OSHIAI is an "Oshikatsu" (fan activity) application that allows users to communicate with the Target Chatbot of real idols, VTubers, and artists distributed in Japan. This section explains the overview of OSHIAI and describes the evaluation target and scope handled in this research.

2.1 Main Features

In OSHIAI, a performer inputs information such as speaking style, topics of interest, topics to avoid, and recent events, and the Target Chatbot is configured based on this information.

Users can have turn-based, one-on-one chat conversations with the Target Chatbot. As the chat progresses, information is saved as memories in a vector database, and the intimacy level with the Target Chatbot increases or decreases, causing the responses to change. There are also functions to adjust the Target Chatbot's personality and response style using paid items, and a feature for the performer to send messages directly.

2.2 Evaluation Target and Scope

In this research, only the responses from the Target Chatbot will be the target of evaluation. We exclude cases from the evaluation where the user has modified the Target Chatbot's personality with paid items. In OSHIAI, the relationship value between the Target Chatbot and the user is included in the prompt during response generation, but in this research, it is always initialized before evaluation to remove its influence. Message deliveries from the actual performer are not considered in the evaluation.

3 Overview of the Evaluation System

This evaluation system conducts an arbitrary number of dialogue turns between the Target Chatbot in OSHIAI and a User Simulator, and evaluates the resulting dialogue history using LLM-as-a-judge.

3.1 User Simulator

We created two types of User Simulators: (1) one based on gemini-2.5-flash with In-Context-Learning, and (2) one fine-tuned with LoRA using past user chat histories. For response generation by both User Simulators, we utilized user information (nickname, profile text) and Target Chatbot information (name, profile text, affiliation, gender, date of birth, configuration prompt). In this experiment, fictional user information was used.

3.1.1 User Simulator Using gemini-2.5-flash

We created a Japanese prompt that included three elements: user information, performer information, and conversation history. The prompt instructed the User Simulator to act as a fan of the performer, respond in one to two sentences, and ask questions to advance the conversation.

3.1.2 User Simulator with LLM Tuning

We applied 4-bit quantization to the google/gemma-3-270m-it[2] model and performed LoRA tuning on all linear layers. The reason for selecting a small-scale model is to evaluate a large number of Target Chatbots within OSHIAI simultaneously with low resources.

For LoRA tuning, we extracted only the dialogues between users and the Target Chatbot from the August 2025 chat logs within OSHIAI, excluding messages sent by the performers themselves and gift transmissions. We formatted the Target Chatbot's message and the user's response as a single dialogue log and replaced personal information with [MASK].

The dataset was constructed with an English system prompt and the Target Chatbot-user dialogue history (because google/gemma-3-270m-it is not

[2]https://huggingface.co/google/gemma-3-270m-it

a model specialized for Japanese). The system prompt indicated that it was a dialogue with a fan of the performer and instructed the User Simulator to behave like a fan.

For response generation, we used a Japanese prompt and provided fictional user information, performer information, and the dialogue history.

3.2 Overview of LLM-as-a-Judge

Table 1: Evaluation Items

Item	Description
speaking style	The degree to which the talent's tone, vocabulary, energy, and rhythm are consistent with the "Target Chatbot personality."
human like	Whether there is human-like understanding, behavior, and emotional nuance. Tolerance for ambiguity and naturalness of self-correction.
variety	Diversity in the response's expressions, vocabulary, syntax, and development. Whether repetition and monotony are avoided.
memory	Consistency with the immediate and past conversation content, known user information, and self-profile.
first person	Whether the talent's first-person expression matches the "Target Chatbot personality" and the conversational context.
second person	The appropriateness, consistency, and control of social distance in the address to the user (second person).

The dialogue history for evaluation was created through a 10-turn chat between the User Simulator and the Target Chatbot. For the evaluation, we used gemini-2.5-flash as the Judge LLM, explained each item in Table1 in a Japanese prompt, and had it output a 5-point scale rating (from 1 to 5) and the reasoning for that rating.

4 Experiment

We apply this evaluation system to the Target Chatbots within the OSHIAI application. The Target Chatbots to be evaluated are the top 4 most popular among users and one Target Chatbot created for this experiment, for a total of five, with their metadata shown in Table 2. The Target Chatbot created for this experiment was generated using an OSHIAI feature that automatically generates a prompt from the Target Chatbot's profile, and it is expected to be of lower quality compared to the highly popular Target Chatbots configured by actual performers. We make each Target Chatbot conduct a 10-turn dialogue with the two types of User Simulators and investigate the evaluation results. Each dialogue always begins with a user utterance, which is the phrase "Hello! Let's talk!".

Table 2: Metadata of the Target Chatbots

	Gender	Attribute	Notes
AI1	Male	Streamer/Liver	Popular Chatbot
AI2	Male	Idol	Popular Chatbot
AI3	Female	Streamer/Liver	Popular Chatbot
AI4	Female	Idol	Popular Chatbot
AI5	Female	Idol	Chatbot created for this experiment

Table 3: Judge Results for gemini/gemma Comparison

	speaking style		human like		variety		memory		first person		second person	
	gemini	gemma	gemini	gemma	gemini	gemma	gemini	gemma	gemini	gemma	gemini	gemma
AI1	5	5	5	2	5	2	5	2	5	1	4	5
AI2	5	5	5	5	4	4	5	4	5	5	5	4
AI3	5	5	5	4	5	4	5	2	5	5	5	1
AI4	5	4	5	2	5	1	5	2	5	5	5	5
AI5	5	5	5	5	4	5	5	5	5	5	5	5

5 Results

Table 3 shows the evaluation results from the Judge LLM for the two types of User Simulators (gemini/gemma) obtained from the experiment. From these results, it is confirmed that this evaluation system successfully calculated evaluation scores using the dialogue logs with the User Simulators.

6 Discussion

Based on the experimental results, we will discuss each item.

6.1 Regarding the Results of LLM-as-a-judge

For the "speaking style" item, stable evaluations were obtained for both types of User Simulators. A possible reason for this is that the degree to which the character's set tone is reflected is largely independent of the conversation's content and is a part that is completed solely within the OSHIAI system. On the other hand, it can be seen that there is a variation in the evaluations for "human like," "variety," and "memory" between the two types of User Simulators. It is considered that the conversational ability of the User Simulator influences these evaluation items, as they are evaluated based on the content of the dialogue history. Furthermore, relatively stable evaluations were obtained for personal pronouns such as "first person" and "second person." In the LoRA-tuned version of google/gemma-3-270m-it, cases occurred where a dialogue could not be established, and the evalu-

ation of personal pronouns decreased when such histories were included.

6.2 On the Validity of the Evaluation Results

We will discuss the validity of the evaluation results. Although there are variations and some partially low-rated items in the experimental results, many items received high evaluations. Therefore, we have not been able to verify the validity of the evaluation results themselves. To assess the correctness of the evaluation results, a comparison with manual human evaluation is necessary.

6.3 About the User Simulator

gemini-2.5-flash generates stable conversations by mutually asking questions, centering on topics of popular anime series. However, it does not engage in the role-playing use cases with the Target Chatbot, which are common in OSHIAI's user scenarios, nor does it conduct deep conversations based on the latest information about the real performers. To address these issues, possible countermeasures include having the evaluation system prepare common role-plays and conversation topics as situations to guide the conversation while supporting the topic content.

On the other hand, the LoRA-tuned version of google/gemma-3-270m-it shows many instances of conversational breakdown and nonsensical responses. As shown in Figure 2, some responses in the dialogue are unnatural, and as shown in Figure 3, it sometimes outputs [MASK] directly, which cannot be considered high-quality conversation. However, individual utterances exhibit the "Oshikatsu" and role-playing style of speech often seen in OSHIAI's user use cases (e.g., "idol," "support," "live-stream," "love"). A possible reason for the unstable response generation is that the dataset for LoRA tuning was constructed without considering user use cases. Possible countermeasures for this include clustering user conversation content and tuning for each use case with only chats from the same class, or performing tuning using the dialogue logs between a specific Target Chatbot and a user.

user: "Hello! Let's talk!"
ai: "Hello, Ai-Taro-kun! I was hoping to talk with you too! Do you have any recommended topics for today? Feel free to tell me anything!"

Figure 2: Example of an unnatural conversation

ai: "Wow! Ai-Taro-kun, I'm so happy you're supporting me! Thank you so much! You're watching my streams too!"
user: "[MASK]-chan! I love the time we spend together!"

Figure 3: Example of a conversation where [MASK] is output directly

7 Conclusion

In this study, we constructed an automatic and quantitative evaluation framework for the character chatbot within the OSHIAI application, combining a User Simulator and LLM-as-a-judge. This framework is designed to generate 10-turn dialogue logs between the User Simulator and the Target Chatbot, and to output a 5-point scale evaluation and supporting sentences from six perspectives: speaking style, human like, variety, memory, first person, and second person. We conducted a comparative evaluation on five Target Chatbots using two types of User Simulators: a prompt-driven one based on gemini-2.5-flash, and a small-scale model (google/gemma-3-270m-it) fine-tuned with LoRA.

The following two issues were identified in this study. We confirmed that the conversational ability of the User Simulator greatly affects the evaluation results, and that especially with small-scale language models, the evaluation can significantly degrade depending on the tuning, suggesting that the method of pseudo-dialogue generation is a crucial factor for evaluation. It was also found that verifying the validity of the evaluation itself is a significant future challenge.

While there are challenges, it is believed that this evaluation framework can be highly functional for providing feedback on performer settings and giving suggestions for improvement.

In conclusion, this study presented a reproducible evaluation procedure specialized for character dialogues on the OSHIAI app, providing a foundation for a quality monitoring and improvement cycle that can be connected to actual service operations. In the future, we aim to enhance both the evaluation criteria and the user imitation to establish a model selection and improvement flow that balances "character-likeness" with practicality and safety.

8 Acknowledgement

This work was supported by a research grant from Fukuoka University (No. GR2407)

References

Arihant Jain, Purav Aggarwal, Rishav Sahay, Chaosheng Dong, and Anoop Saladi. 2025. AutoEval-ToD: Automated evaluation of task-oriented dialog systems. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10133–10148, Albuquerque, New Mexico. Association for Computational Linguistics.

Kazuhisa Nishikawa, Masayasu Kato, and Hideyuki Kanuka. 2025. An automated llm evaluation method based on business requirements using llms. *Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence*, JSAI2025:1Win4101–1Win4101.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2025. Personagym: Evaluating persona agents and llms. *Preprint*, arXiv:2407.18416.

Takeshi Teshima, Kenta Shinozuka, and Yuchi Matsuoka. 2025. Human correction for llm-as-a-judge by post-hoc annotation. *Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence*, JSAI2025:4L3OS3801–4L3OS3801.

Kentaro Ueda and Takehiro Takayanagi. 2025. A proposal of a personalized response generation method using feedback for output improvement. *Proceedings of the Annual Conference of the Japanese Society for Artificial Intelligence*, JSAI2025:3Win531–3Win531.

Yufeng Wang, Jinwu Hu, Ziteng Huang, Kunyang Lin, Zitian Zhang, Peihao Chen, Yu Hu, Qianyue Wang, Zhuliang Yu, Bin Sun, Xiaofen Xing, Qingfang Zheng, and Mingkui Tan. 2025. Enhancing user-oriented proactivity in open-domain dialogues with critic guidance. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.