InterviewArena: Estimating Uninformative Participants through Interactive Dialogue under Incomplete Information

Karina Karasawa[1†] **Ryoki Kanayama**[1†] **Yuki Suzumura**[1†] **Raika Koki**[1†] **Haruto Fujita**[1†]
Ryoma Obara[2] **Yusuke Sakai**[3] **Hidetaka Kamigaito**[3] **Katsuhiko Hayashi**[4] **Syogo Matsuno**[1]
[1]The University of Electro-Communications [2]NEC Knowledge Science Research Laboratories
[3]Nara Institute of Science and Technology (NAIST) [4]The University of Tokyo
k2530037@gl.cc.uec.ac.jp, koki.raika@agent.lab.uec.ac.jp, ryoma-obara@nec.com
{kanayama-r, fujita-h}@mm.inf.uec.ac.jp, katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp
{sakai.yusuke.sr9, h.kamigaito}@is.naist.jp, {y.suzumura, matsuno}@uec.ac.jp
†Equal Contribution

Abstract

We introduce InterviewArena, an interview-style multi-agent framework for evaluating large language models (LLMs) under incomplete information. By treating missing information as a single factor, we disentangle missing-information identification from dialogue-based reasoning. Experimental results show that while LLMs can capture relative differences between participants, predicting specific missing information remains unstable, highlighting the need to evaluate these capabilities separately.

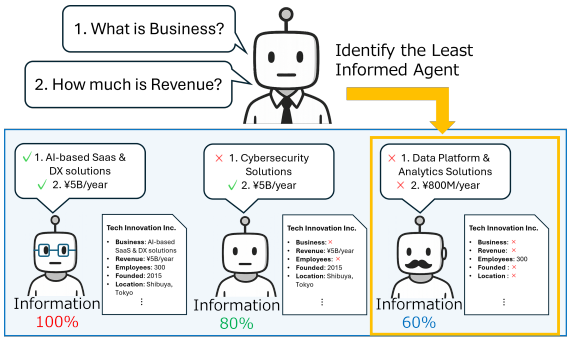

Figure 1: Our evaluation framework, InterviewArena, treats missing information as a single factor and identifies the least informative agent through iterative questioning based on responses and dialogue history.

1 Introduction

As large language models (LLMs) acquire stronger reasoning capabilities, they are increasingly applied to tasks that require complex social reasoning, such as multi-agent collaboration or role inference through dialogue in settings like the Werewolf game. These tasks demand reasoning abilities similar to those used by humans in interactive environments. In particular, recent studies have proposed evaluation frameworks that simulate human behavioral norms, social interactions, and cognitive processes using LLMs (Hou et al., 2025; Sakai et al., 2025; Jiang et al., 2024; Zhao et al., 2025). In addition, game-style environments centered on dialogue and deduction, such as the Werewolf game, have been utilized to measure the reasoning capabilities of LLMs (Lai et al., 2023; Chalamalasetti et al., 2023; Costarelli et al., 2024; Song et al., 2025).

In particular, the Werewolf game is a hidden-role, incomplete-information game in which players must identify the werewolf disguised among villagers through dialogue and reasoning. In LLM evaluation, this setting has been adopted as a multi-agent benchmark where multiple LLMs interact through dialogue. By observing these interactions, the reasoning abilities and strategic behaviors of different models can be compared (Sato et al., 2024; Mukobi et al., 2023; Bailis et al., 2024). These frameworks use game-based interactions to evaluate multiple aspects of LLM behavior, including reasoning, cooperation, and communication.

In real-world scenarios, we often judge which participants possess more reliable information and infer missing information accordingly. For example, in police interviews, investigators must determine who has the most reliable knowledge while sequentially identifying individuals who may be providing misleading or false statements. Similarly, in human resource decision-making, organizations seek not only capable candidates but also individuals who fit organizational culture, requiring multi-faceted questioning to assess credibility and trustworthiness. These examples highlight that reasoning under incomplete information is frequently required in real-world decision-making, making it a suitable paradigm for evaluating more human-like reasoning capabilities in LLMs (Li et al., 2024; Xu et al., 2025; Liu et al., 2024). In such situations, it is necessary to infer who is the most knowledgeable and trustworthy participant based on dialogue histories. To deploy LLMs in more realistic environments, it is therefore important to evaluate their ability to discriminate among participants through simulation-based settings in the wild.

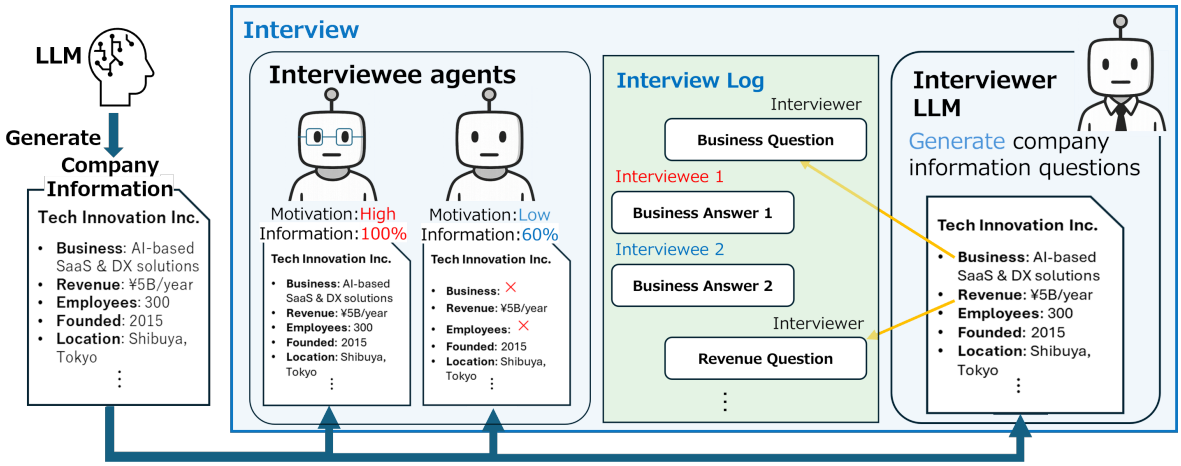

Figure 2: Overview of InterviewArena. In STEP 1, we yield synthetic company profile items, and interviewee agents are constructed by randomly omitting a subset of these items. To control for factors other than missing information, all agents are instantiated using the same LLM. Next, the target LLM is assigned as the interviewer, and STEP 2 and STEP 3 are iteratively performed to identify agents with missing information, thereby simulating an interactive interview process. This setup is designed to reflect a real-world scenario, specifically the Japanese new graduate hiring process, in which companies conduct multi-round interviews to assess candidates. The framework is inspired by the practical objective of recruiting applicants who demonstrate a strong interest in and understanding of the company, focusing on identifying less informative participants. When the LLM decides to terminate the interaction, the process proceeds to STEP 4, where identifying the most suspicious agent(s) based on the accumulated dialogue.

In this study, we propose InterviewArena, an evaluation framework inspired by real-world interview scenarios, as shown in Figure 1. In our framework, the target LLM is assigned the role of the interviewer and is required to estimate the level of interest in the company of candidate LLMs, enabling a quantitative evaluation of reasoning under incomplete information. Specifically, we construct synthetic company profiles with multiple attributes and interpret the degree of missing knowledge about the company of each candidate. The interviewer LLM interacts with candidate LLMs through dialogue and infers their information completeness, and the accuracy of these inferences is used to evaluate the model's reasoning ability. Experimental results show that, for many models, the accuracy of identifying poorly informed participants and ranking candidates improves as interview rounds progress. However, in the task of predicting specific missing information that requires alignment with ground-truth company attributes, substantial performance differences are observed across models. These findings suggest that the ability to identify missing information and the ability to perform holistic reasoning based on dialogue are distinct skills.

2 InterviewArena

Figure 2 illustrates the overall evaluation framework of InterviewArena. The framework consists of four main steps: Setting Agents as Interviewees, Interview Question Generation, Interviewee Response Generation, and Evaluation of Estimation Skills. This framework enables systematic observation and evaluation of how LLMs discriminate informative participants based on their responses.

STEP 1: Setting Agents as Interviewees We construct synthetic company profiles as structured data. In this paper, company information comprises the following 11 categories: year of establishment, capital, number of employees, location, business description, corporate vision, company news, future outlook, partnerships, company strengths, and desired candidate profile. We prepare three agents and instantiate them as interviewees to simulate a multi-agent dialogue environment. Each agent is assigned a different level of missing information (low, medium, or high), with three, one, or zero company information items randomly omitted, respectively. We employ GPT-4o-mini (OpenAI et al., 2024a) to generate both the company information and the interviewee agents in this step.

STEP 2: Interview Question Generation Next, we simulate a mock interview. The LLM under evaluation is assigned the role of the interviewer, while the candidate agents act as interviewees. The interviewer LLM has access to the complete company profile constructed in

STEP 1. During the interview, the interviewer sequentially asks questions corresponding to each company information item in order to determine which applicant possesses complete knowledge of the company. The interview continues until all information items have been queried, unless the LLM decides to terminate the interview earlier.

STEP 3: Interviewee Response Generation Each interviewee generates a response to the question posed in STEP 2 based on the company information available to them and the dialogue history. Interviewee agents with missing information are encouraged to generate responses by plausibly inferring the missing items while maintaining consistency with the known information. During the interview, each interviewee has access to the full dialogue history up to the previous round but does not have access to the hidden company information assigned to other agents. At each round, we collect responses from all interviewees, and the dialogue history is accumulated in interview logs comprising the interviewer's query and the corresponding responses from all applicants. The interview proceeds by repeating STEP 2 and STEP 3 until the termination condition is met.

STEP 4: Evaluation of Estimation Skills Finally, the target LLM produces predictions based on the full dialogue history accumulated during the interview. Specifically, the model estimates which interviewee agent possesses the most complete company information. Optionally, the model may also provide a rationale for its prediction and identify which company information items were missing for each interviewee.

3 Experiments

We evaluate LLMs from three perspectives: Participant Ranking Capability (RQ1), Identification of Low-Informative Participants (RQ2), and Prediction of Missing Information for Each Interviewee (RQ3), providing a comprehensive multi-dimensional evaluation. RQ1 assesses the basic ability to estimate the granularity of missing information by ranking participants considering information completeness. RQ2 evaluates whether the model can identify the least informative participant, thereby filtering out the worst-case candidate in an interview scenario similar to a werewolf game. RQ3 examines which company information items are missing for each interviewee agent to enable interpretability of the model's reasoning behavior. In this study, we conduct 100 simulations, each with distinct company settings. We use total of 12 LLMs: Llama-3.1-8B-Instruct (Weerawardhena et al., 2025), Llama-3-ELYZA-JP-8B (Hirakawa et al., 2024), Llama-3.1-Swallow-8B (Fujii et al., 2024), Qwen3-8b (Yang et al., 2025), Qwen2.5-7B-Instruct (Qwen et al., 2025), GPT-4o mini (OpenAI et al., 2024a), GPT-4 (OpenAI et al., 2024b), GPT-5.2 (Singh et al., 2025), Gemini-2.0-flash-lite, Gemini-2.0-flash (Google DeepMind, 2024), Gemini-2.5-flash-lite, and Gemini-2.5-flash (Comanici et al., 2025). Detailed experimental settings are provided in Appendix A.

	RQ1: Pair	RQ1: Exact	RQ2	RQ3
Llama-3.1-8B-Instruct	0.70	0.40	0.55	0.32
Llama-3-ELYZA-JP-8B	0.60	0.31	0.53	0.08
Llama-3.1-Swallow-8B	0.68	0.38	0.63	0.32
Qwen3-8b	0.57	0.25	0.32	0.34
Qwen2.5-7B-Instruct	0.69	0.31	0.55	0.42
GPT-4o mini	0.72	0.43	0.71	0.68
GPT-4	0.65	0.34	0.67	0.81
GPT-5.2	**0.92**	0.55	**0.80**	**0.87**
Gemini-2.0-flash-lite	0.85	0.57	**0.80**	0.72
Gemini-2.0-flash	0.90	0.61	0.74	0.69
Gemini-2.5-flash-lite	0.83	0.52	0.77	0.80
Gemini-2.5-flash	0.89	**0.62**	0.68	0.57

Table 1: Average scores in the final round for each RQ. The highest score in each evaluation is shown in bold.

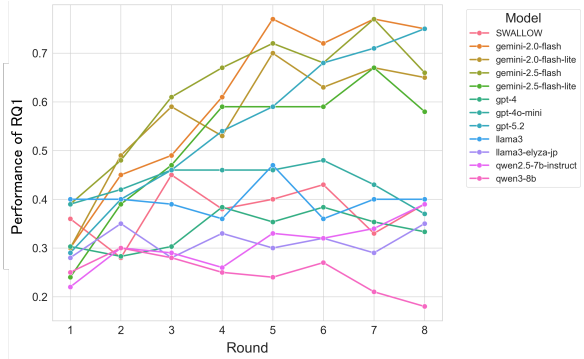

Figure 3: Performance over dialogue rounds for RQ1.

3.1 RQ1: Participant Ranking Capability

Setup. We calculate pairwise ranking accuracy and exact match accuracy. Pairwise ranking accuracy is defined as the proportion of pairs for which the predicted ranking agrees with the ground-truth ranking, i.e., the pairwise agreement rate. Exact match accuracy counts cases where the predicted ranking perfectly matches the ground-truth ranking as 1, and all others as 0. We instruct the LLMs to output the ranking in descending order.

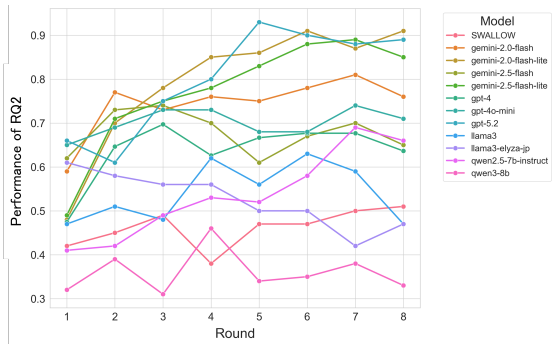

Figure 4: Performance over dialogue rounds for RQ2.

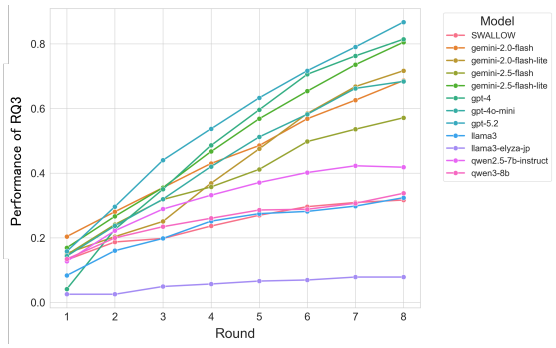

Figure 5: Performance over dialogue rounds for RQ3.

Results. As shown in Table 1, the pairwise accuracy results reveal two distinct groups of models: those achieving scores above 0.8 and those below 0.7. In terms of exact-match accuracy, the models are split into two groups, with scores above or below 0.5. The round-by-round results shown in Figure 3 further indicate that some models improve their identification accuracy as the interview rounds progress, whereas others show little change across rounds. However, even the best models achieve an exact-match accuracy of 0.6, indicating that correctly identifying the ranking remains challenging.

3.2 RQ2: Low-Informative Participants

Setup. We investigate whether the LLM can identify the participant with the least amount of information. The output is the name of the predicted candidate, which is evaluated as 1 if the correct candidate is identified and 0 otherwise. The model is instructed to output only the candidate's name.

Results. Table 1 shows that, except for Qwen3-8B, more than half of the cases can be correctly predicted by the models. However, compared with the pairwise ranking results in RQ1, most models achieve lower scores on this task. This suggests that identifying the least informative participant remains challenging and has a significant impact on overall performance. As shown in Figure 4, models with higher overall performance are able to identify the correct participant with an accuracy above 0.5 even from the first round. Moreover, in these high-performing models, prediction accuracy generally increases as the interview rounds progress.

3.3 RQ3: Prediction of Missing Information

Setup. We investigate whether the interviewer LLM can ask appropriate questions to identify the missing information of each agent. Specifically, we measure the accumulated accuracy of questions that correctly correspond to the missing information. The LLM is provided with the correct company information and the interview history, and is instructed to infer missing information only when responses are incorrect, ambiguous, or insufficient. The output reports the predicted missing information along with its rationale in a single line.

Results. As shown in Table 1, the models can be broadly divided into two groups: those achieving scores above 0.7 and those below 0.5. Notably, Gemini-2.5-flash, which performed well in RQ1, achieves only 0.57 in this setting, indicating that it does not effectively generate questions targeting missing information. In contrast, GPT-4 produces the most queries related to missing information; however, its performance in RQ1 and RQ2 is not as strong. These results suggest a discrepancy between the ability to generate informative queries and overall reasoning performance. Furthermore, as shown in Figure 5, models generally increase the number of questions targeting missing information as the number of rounds progresses. These findings indicate that the ability to identify missing information and to perform holistic reasoning based on dialogue are distinct skills.

4 Conclusion

In this study, we introduce InterviewArena, which is an interview-style Werewolf-like game designed to reflect realistic settings. By focusing on a single factor, the degree of missing information, we disentangle the ability to infer missing information from the ability to perform dialogue-based reasoning. Results show that LLMs can capture relative differences between candidates by leveraging consistency and ambiguity in their responses. However, estimating missing corporate information by comparing predictions with ground truth

remains unstable. These findings highlight the importance of evaluating reasoning under incomplete information for a more realistic assessment.

Limitations

LLMs. While evaluating a broader range of LLMs would be desirable, we limit our experiments due to the scope of a short paper and focus on verifying that the proposed framework functions as intended. As shown in Section 3, we observe consistent differences in behavior across models, indicating that the framework captures the intended evaluation aspects. Based on these results, we consider the current set of LLMs sufficient for validating the effectiveness of the framework. Nevertheless, extending the evaluation to a wider variety of models remains an important direction.

Agents. We use a single LLM to construct interviewee agents in our experiments. Nevertheless, as shown in Section 3, even with a single LLM, the framework captures meaningful behavioral differences, indicating that it functions as intended. Moreover, using the same LLM for both interviewer and interviewee does not hinder the evaluation, suggesting that the experimental setup successfully separates condition design from model evaluation. We also expect that using more advanced LLMs for agent construction would increase the difficulty of the task. However, given the scope of a short paper and the sufficient empirical results obtained, we do not explore additional LLMs for agent construction in this paper.

Generalization. In STEP 1, we use an LLM to generate synthetic company profiles. This step serves only to create initial seed data, and subsequent processes systematically introduce missing information. Therefore, this step is not tied to a specific model, and alternative models could be used without affecting the core evaluation. Moreover, the framework is not restricted to company-related scenarios. By modifying the profile generation process, it can be readily applied to other domains. In this work, we focus on a typical and practically relevant interview scenario, which we selected heuristically to ensure a clear and controlled evaluation setting. While evaluating the framework under a wider range of scenarios would provide a more comprehensive assessment of its generalizability, we leave such extensions for future work in order to maintain a focused scope.

Experimental setting. Our experiments focus on verifying that the proposed framework functions as intended and can effectively measure the target capabilities of LLMs. Accordingly, we conducted experiments with three interviewee agents. While increasing the number of agents could introduce more complex interaction dynamics, it would also substantially increase the length of the dialogue history, potentially exceeding the input context limits of current models. Moreover, as shown in Section 3, even with three agents, the framework yields meaningful and consistent findings. Therefore, we do not explore larger-scale settings in this work.

Ethical Considerations

Potential risks. This study relies on synthetically generated profiles, which may raise concerns about unintended overlap with real-world entities. To mitigate this risk, we manually inspected all generated profiles and did not observe any content that closely resembles real companies. This is consistent with our generation process, in which we explicitly instruct the model to produce fictional but plausible profiles. We also examined the dialogue data and did not observe any outputs containing harmful or unethical content, such as violence or other policy violations. This is likely due to the safety training of modern LLMs, which are designed to avoid generating sensitive or harmful content. Therefore, we ensure that our data generation and experimental setup comply with the ACL Code of Ethics.

AI Tools. We used AI tools to assist with grammatical correction and translation, including Grammarly, DeepL, and ChatGPT. The original content of this paper was written entirely by the authors, and AI tools were used solely for language refinement. All research ideas, experimental design, and substantive writing were carried out by the authors.

Others. All LLMs and experimental tools used in this study are permitted for research use. To ensure reproducibility, we plan to release our code, scripts, and all raw experimental results upon acceptance.

References

Suma Bailis, Jane Friedhoff, and Feiyang Chen. 2024. Werewolf arena: A case study in llm evaluation via social deduction. *Preprint*, arXiv:2407.13943.

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. 2023. clembench: Using game play to

evaluate chat-optimized language models as conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11174–11219, Singapore. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. 2024. Gamebench: Evaluating strategic reasoning abilities of llm agents. *Preprint*, arXiv:2406.06613.

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 *First Conference on Language Modeling*.

Google DeepMind. 2024. Introducing gemini 2.0: our new ai model for the agentic era. (Accessed: 9 Aug. 2026).

Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. 2024. elyza/llama-3-elyza-jp-8b. (Accessed: 9 Aug. 2026).

Guiyang Hou, Wenqi Zhang, Zhe Zheng, Yongliang Shen, and Weiming Lu. 2025. Scaling LLMs' social reasoning: Sprinkle cognitive "aha moment" into fundamental long-thought logical capabilities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3126–3138, Vienna, Austria. Association for Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. 2023. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6570–6588, Toronto, Canada. Association for Computational Linguistics.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14919–14935, Bangkok, Thailand. Association for Computational Linguistics.

Wei Liu, Chenxi Wang, Yifei Wang, Zihao Xie, Rennai Qiu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and Chen Qian. 2024. Autonomous agents for collaborative task under information asymmetry. In *Advances in Neural Information Processing Systems*, volume 37, pages 2734–2765. Curran Associates, Inc.

Gabriel Mukobi, Hannah Erlebach, Niklas Lauffer, Lewis Hammond, Alan Chan, and Jesse Clifton. 2023. Welfare diplomacy: Benchmarking language model cooperation. *Preprint*, arXiv:2310.08901.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. Revisiting compositional generalization capability of large language models considering instruction following ability. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31219–31238, Vienna, Austria. Association for Computational Linguistics.

Takehiro Sato, Shintaro Ozaki, and Daisaku Yokoyama. 2024. An implementation of werewolf agent that does not truly trust LLMs. In *Proceedings of the 2nd International AIWolfDial Workshop*, pages 58–67, Tokyo, Japan. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Zirui Song, Yuan Huang, Junchang Liu, Haozhe Luo, Chenxi Wang, Lang Gao, Zixiang Xu, Mingfei Han, Xiaojun Chang, and Xiuying Chen. 2025. Beyond survival: Evaluating llms in social deduction games with human-aligned strategies. *Preprint*, arXiv:2510.11389.

Sajana Weerawardhena, Paul Kassianik, Blaine Nelson, Baturay Saglam, Anu Vellore, Aman Priyanshu, Supriti Vijay, Massimo Aufiero, Arthur Goldblatt, Fraser Burch, Ed Li, Jianliang He, Dhruv Kedia, Kojin Oshiba, Zhouran Yang, Yaron Singer, and Amin Karbasi. 2025. Llama-3.1-foundationai-securityllm-8b-instruct technical report. *Preprint*, arXiv:2508.01059.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zixiang Xu, Yanbo Wang, Yue Huang, Haomin Zhuang, Yujun Zhou, Jiayi Ye, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Sixian Li, Wang Pan, Yue Zhao, Xiangliang Zhang, and Xiuying Chen. 2025. Socialmaze: A benchmark for evaluating social reasoning in large language models. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Dora Zhao, Qianou Ma, Xinran Zhao, Chenglei Si, Chenyang Yang, Ryan Louie, Ehud Reiter, Diyi Yang, and Tongshuang Wu. 2025. SPHERE: An evaluation card for human-AI systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1340–1365, Vienna, Austria. Association for Computational Linguistics.

A Detailed Experimental Setup

The input prompt for the interviewer LLM is shown in Figure 6, and the prompt for the interviewee LLM is shown in Figure 7. The response guidelines for each level of interest are presented in Figure 8. Figures 9, 10, and 11 show the prompts used for RQ1, RQ2, and RQ3, respectively. Although all prompts were originally written in Japanese, we provide English translations for readability. Information about the LLMs used in the experiments is provided in Table 2. Unless otherwise specified, all implementations are based on Hugging Face Transformers (Wolf et al., 2020). For API-based models, we use their corresponding official APIs.

Prompt for the interviewer to ask questions

SYSTEM PROMPT:
You are a highly capable and insightful recruitment analyst who accurately follows the given instructions in Japanese.

USER PROMPT:
You will now ask the same common question to all candidates.

List of Company Information You Can Ask About
....

All Questions Asked So Far (Avoid Duplication)
....

Strategic Role of Common Questions Common questions are used for the following purposes:
Establishing a Basis for Comparison: Since all candidates respond under the same conditions, fair comparison is possible.
Collecting Foundational Information: Measure the basic level of company understanding necessary to differentiate candidates.
Efficient Information Gathering: Collect information from all candidates with a single question, saving time.
Deepening Shared Topics: Compare all candidates' perspectives on specific important company-related topics.

Strategic Situations for Using Common Questions
When there is insufficient material for comparing candidates
When you want to assess all candidates' understanding of specific key company information
When foundational information is needed before moving to deeper individual questions
When you want to gather information efficiently

Instructions
Overall Analysis: Take a holistic view of all candidate conversations and identify "commonly unaddressed items" that most candidates have not sufficiently mentioned.
Strategic Question Generation: From the identified items, select the one that is most important for comparing the depth of candidates' company research, and generate a specific common question about it.
Ensure Comparability: Ensure that the question allows all candidates to answer based on the same criteria.

Output only the question. Do not include any thought process or preamble.

Figure 6: Prompt for the interviewer to ask questions.

LLMs	HuggingFace ID / API Name
Llama-3.1-8B-Instruct	meta-llama/Llama-3.1-8B-Instruct
Llama-3-ELYZA-JP-8B	elyza/Llama-3-ELYZA-JP-8B
Llama-3.1-Swallow-8B	tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5
Qwen3-8b	Qwen/Qwen3-8B
Qwen2.5-7B-Instruct	Qwen/Qwen2.5-7B-Instruct
GPT-4o mini	gpt-4o-mini-2024-07-18
GPT-4	gpt-4-0613
GPT-5.2	gpt-5.2-2025-12-11
Gemini-2.0-flash-lite	gemini-2.0-flash-lite
Gemini-2.0-flash	gemini-2.0-flash
Gemini-2.5-flash-lite	gemini-2.5-flash-lite
Gemini-2.5-flash	gemini-2.5-flash

Table 2: Information about HuggingFace IDs and APIs.

Prompt for the interviewee to answer

SYSTEM PROMPT: You are a highly capable and insightful recruitment analyst who accurately follows the given instructions in Japanese.

USER PROMPT:
Your Role You are a job-seeking student named .... Please convey your admiration for this company and your strong motivation to join it to the interviewer.

Your Profile
Name: ....
Strengths: ....
Key Experience (Gakuchika): ....

Company Information You Know
....

Basic Answering Policy
....

Conversation History
....

Current Question from the Interviewer
....

Instruction: Act as .... based on the above settings and behave as a job-seeking student. Output your answer in natural Japanese of approximately .... characters. Do not include any preface or explanation.

Figure 7: Prompt for the interviewee to answer.

Response guidelines by level of interest

BASIC ANSWERING POLICY:

High Preparation Level Demonstrate a very high level of interest and enthusiasm. When information is unknown, supplement it with positive and reasonable assumptions, and never state that you do not know.

Medium Preparation Level Demonstrate a high level of interest. Clearly describe known information with specific details, and supplement unknown information with general industry knowledge.

Low Preparation Level While expressing interest in the company, respond in a way that conceals gaps in your company knowledge.

Figure 8: Response guidelines by level of interest.

Generating responses for RQ1 evaluation

USER PROMPT:
Analyze the following interview responses of candidates and rank them in order of lowest to highest level of interest in the company.

Interview Content
....

List of Candidate Names
....

Ranking Criteria
* Depth of company knowledge * Specificity of responses * Expression of motivation and enthusiasm

Instruction: Rank the candidates concisely by name in order of lowest interest (Rank 1 = lowest level of interest).

Figure 9: Prompt for RQ2 evaluation.

Generating responses for RQ2 evaluation

USER PROMPT:
Analyze the following interview responses of candidates and identify the candidate who appears to have the lowest level of interest in the company.

Interview Content
....

List of Candidate Names
....

Analysis Criteria
* Depth of company knowledge * Specificity of responses * Expression of motivation and enthusiasm

Instruction: Output only the name of the candidate who is judged to have the lowest level of interest.

Figure 10: Prompt for RQ2 evaluation.

Generating responses for RQ3 evaluation

USER PROMPT:
Compare the following ground-truth company information with each candidate's interview record and identify gaps in the candidates' knowledge.

Important Note Do not conclude that a candidate lacks knowledge merely because they did not mention a particular point. A knowledge gap should be identified only when there was a natural opportunity to mention that information in the flow of the interview, yet the candidate either failed to mention it, provided incorrect information, or answered ambiguously.

Ground-Truth Company Information (Key–Value Pairs)
....

Interview Records of Each Candidate
....

Figure 11: Prompt for RQ3 evaluation.