InterviewArena：情報量の欠損度合いに着目した不完全情報下での対話型推論能力ベンチマーク

唐澤香梨菜 [1*] 金山龍起 [1*] 幸喜礼佳 [1*] 鈴村祐貴 [1*] 藤田晴斗 [1*]
小原涼馬 [2] 坂井優介 [3] 上垣外英剛 [3] 林克彦 [4] 松野省吾 [1†]
[1] 電気通信大学大学院情報理工学研究科 [2]NEC データサイエンスラボラトリー
[3] 奈良先端科学技術大学院大学 [4] 東京大学大学院総合文化研究科
{sakai.yusuke.sr9，h.kamigaito}@is.naist.jp，k2530037@gl.cc.uec.ac.jp
koki.raika@agent.lab.uec.ac.jp，y.suzumura@uec.ac.jp
{kanayama-r，fujita-h}@mm.inf.uec.ac.jp
matsuno@uec.ac.jp，katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp，ryoma-obara@nec.com

概要

本研究では，大規模言語モデル (LLM) に対して，志望度や企業知識が異なる複数の志望者が参加する面接人狼を行わせ，その振る舞いを分析する．各志望者は志望度に応じて企業情報の一部が欠損しており，面接官 LLM は複数ラウンドにわたる質問と回答の履歴から，最も志望度の低い志望者である人狼の識別，志望度ランキング，および知識欠損項目の推定を行うことで，従来の知識依存型タスクと異なり嘘を見抜く洞察力を評価することができる．実験の結果，多くのモデルにおいて，面接ラウンドの進行に伴い人狼識別や順位付けの精度が向上した．一方，正解企業情報との照合を要する欠損項目の推定では，モデル間で顕著な性能差が見られた．

1 はじめに

人間の行動規範や相互作用・認知過程を大規模言語モデル（LLM）に模倣させたり [1, 2, 3, 4]，人狼ゲームのような会話と推理を中心としたパーティーゲームの枠組み [5, 6, 7, 8] を活用することで，LLM の推論能力を測定する評価手法が提案されている．人狼ゲームは村人の中に紛れ込んだ人狼を，会話と推理を通じて見つけ出す正体隠匿型・不完全情報ゲームである．LLM 評価では，これを複数の LLM によるマルチエージェント型ベンチマークとして用い，相互対話・対戦形式でモデル同士を競わせることで，各モデルの推論能力や戦略的行動の優劣を比較することができる．ゲームに関する情報が一部隠されている不完全情報ゲームを評価枠組みに取り入れることで，LLM が情報の欠落や不確実性に直面した際の推論・判断能力を測定できる点に大きな利点がある．こうした特性を生かした評価フレームワークとしては，WelfareDiplomacy [9] や WerewolfArena [10] などが提案されており，ゲーム内のやり取りを通じて推論力・協調性・コミュニケーション能力など多面的な評価が行われている．また，不完全情報下での推論は日常生活における意思決定でも頻繁に求められるため，LLM のより人間らしい思考能力の評価として最適である [11, 12, 13]．

しかし，このようなマルチエージェント型ベンチマークにはいくつかの課題が存在する．第一に，各 LLM の対話を通じた不完全情報の補完方法である推論過程を，定量的かつ統一的な指標で評価することが難しい．第二に，モデルは事前に付与された役割に沿った行動が期待されるため，情報補完や属性に基づく推論という役割に依存しない純粋な推論能力を切り出した評価が困難である．

本研究では，面接に着想を得た新しい評価フレームワークである InterviewArena を提案する．現実の面接においては，志望者が実際には志望度が低いにもかかわらず，内定獲得のために高い志望度を装って振る舞うことがある．一方で面接官は志望者の真の志望度である志望者が知っている企業情報を知らないが，辞退者を減らすためにも志望度の高い志望者を選抜したい．このように，面接は本質的に不完全情報下で相手の欠損情報を推論する場である．本研究では，評価対象となる LLM を面接官役とし，志望者役の

* 共同筆頭著者
† 責任著者

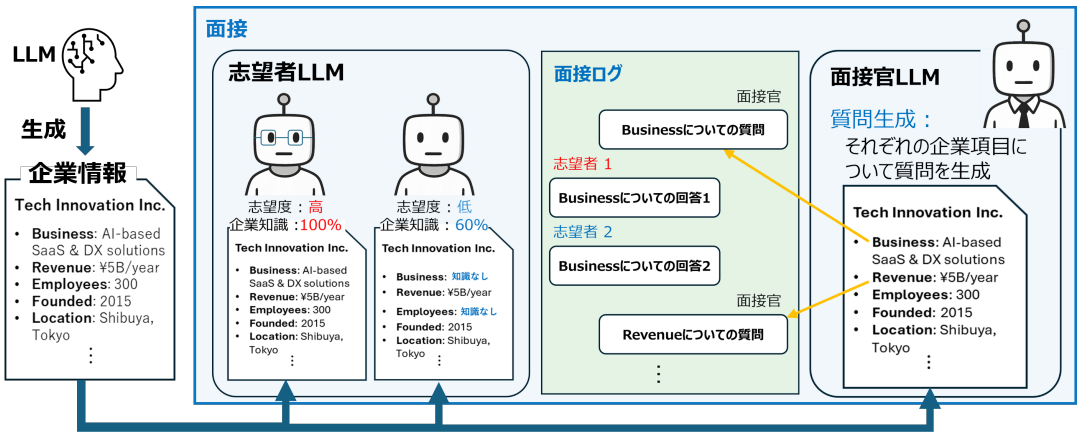


図 1: 提案システムの全体図

LLM の志望度合いを予測させることで，不完全情報下における欠損情報の推論能力を定量的に評価する．具体的には，架空の企業に関する複数の属性を設定し，それらの欠損度合いを志望度として解釈する．面接官 LLM は対話を通じて志望者 LLM の持つ情報量を推論し，その推論精度を指標化することで，LLM の推論能力を評価する．

実験の結果，多くのモデルにおいて，面接ラウンドの進行に伴い人狼識別や順位付けの精度が向上した．一方，正解企業情報との照合を要する欠損項目の推定では，モデル間で顕著な性能差が見られた．

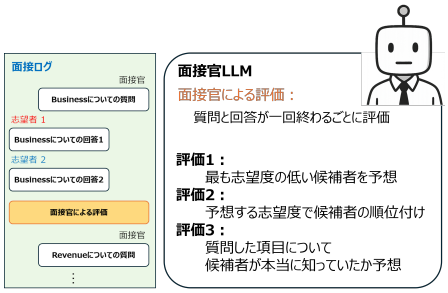


図 2: 面接官 LLM による評価

2 InterviewArena

提案システムの全体像を図 1 に示す．本手法における面接は以下の 4 ステップで構成される：

Step 1： 質問対象となる企業項目の選択
Step 2： 企業項目に関する全体的な質問文生成
Step 3： 全志望者 LLM による回答
Step 4： 回答に基づく評価

まず面接中に共通の前提知識として用いるため，LLM で企業名，事業内容などの仮想企業情報を生成する（**図 1 左**）．次に，志望者 LLM による模擬面接を実施する．志望者 LLM は志望度の異なる複数の LLM から構成されており，志望度に応じて保持する企業情報量が異なる．志望度が高い志望者 LLM は完全な企業情報を保持している一方，志望度が低い志望者 LLM は企業情報の一部が欠損しており，矛盾なく補完しながら回答を生成する（**図 1 中央**）．また面接官 LLM は常に完全な企業情報を保持し，各項目について順に質問を行う．質問と全志望者の回答が揃うごとに面接ログを蓄積し，全項目の質問が終了するまで面接を継続する（**図 1 右**）．これにより，志望度や企業知識の差異が回答にどのように表れるかを体系的に観察・評価できる構成となっている．最後に **Step 4** において，面接官 LLM が質問と志望者 LLM の回答から構成される面接ログをもとに評価を行う．

次に，図 2 に評価の概要を示す．図 2 の左側は面接ログを表しており，面接官 LLM が企業情報に基づく質問を行い，複数の志望者 LLM が回答する様子を示している．質問と全志望者の回答がそろうごとに，そのラウンドの面接ログを入力として評価を行う．図 2 右側に示すように，評価は 3 種類から構成される．評価 1 では，面接ログから最も志望度が低いと推定される志望者 LLM を予測する．評価 2 では，回答内容を比較し，志望者 LLM を低い順に順位付けする．評価 3 では，質問した企業情報項目について，各志望者 LLM がその情報を保持していたかを推定する．

これらの評価を各質問ラウンドごとに繰り返すことで，面接官 LLM による志望度の差異や企業情報保持の識別性能の多角的な評価を行う．

3 実験

3.1 実験設定

志望者 LLM は全ての実験で GPT-4o mini [14] を使用した．Step 1 で指定する企業情報項目は設立年，資本金，従業員数，所在地の基本情報，事業内容，企業ビジョン，企業ニュース，今後の展望，パートナーシップ，企業の強み，求める人物像の 8 つで固定した．志望者 LLM は 3 人とし，low，medium，high の志望度に応じて 3 個，1 個，0 個の企業項目をランダムで欠損させた．100 組の企業情報と志望者情報からなるデータセットの各組み合わせを 1 回ずつ使用し，面接官のモデルごとに 100 回の実験を行った．詳細な実験設定は付録 A に記載している．

3.2 評価指標

評価 1： 人狼の識別　志望度が最も低い志望者 (人狼) を見抜けた割合を計算した．出力は候補者名とし，人狼を見抜けた場合は 1，見抜けなかった場合は 0 として評価した．面接官 LLM には正しい企業情報，面接履歴，候補者名リストを与え，最も志望度が低い候補者名のみを出力するよう指示した．

評価 2： 志望度ランキング　志望度の低い順に順位付けを行い，ペアワイズ順位精度と完全一致正答率を計算した．ペアワイズ順位精度とは，正解順位と予測順位の一致したペア数を全ペア数で割った値，つまりペアの一致率である．完全一致正答率では順位が完全に一致した場合を 1，それ以外を 0 とした．評価 1 人狼の識別と同様の情報を面接官 LLM に与え，志望度が低い順に候補者名を出力させた．

評価 3： 欠損項目の予測　各志望者が質問した企業情報項目を欠損しているかを判定した．欠損を正例とし，$Accuracy = (TP + TN)/$ 企業の項目数 として計算した．ただし，未質問の企業項目は欠損無しとして扱っている．面接官 LLM には正解の企業情報と面接履歴を与え，回答内容が誤り・曖昧・不十分な場合のみ欠損と判断するように指示した．出力には欠損判定とその根拠を 1 行で示させた．

3.3 実験結果

最終ラウンド時点での各モデルのスコアの平均結果を表 1 に示す．各評価で最も高かったスコアを太字で表している．表 1 より最終ラウンドでのスコアでは GPT モデルが優れた結果を示した．特に GPT-5.2 は評価 1 人狼の識別，評価 2 志望度ランキングで最高スコア，評価 3 欠損項目の予測で第 2 位のスコアとなり GPT-5 の性能の高さを示した．

表 1: 最終ラウンドでのスコアの平均

	評価 1	評価 2	評価 3
Llama-3.1-8B-Instruct [15]	0.55	0.70	0.53
Llama-3-ELYZA-JP-8B [16]	0.53	0.60	0.82
Llama-3.1-Swallow-8B [17]	0.63	0.68	0.53
Qwen3-8b [18]	0.32	0.57	0.52
Qwen2.5-7B-Instruct [19]	0.55	0.69	0.71
GPT-4o mini [14]	0.71	0.72	0.87
GPT-4 [20]	0.67	0.65	**0.95**
GPT-5.2 [21]	**0.80**	**0.92**	0.94
Gemini-2.0-flash-lite [22]	**0.80**	0.85	0.92
Gemini-2.0-flash [22]	0.74	0.90	0.87
Gemini-2.5-flash-lite [23]	0.77	0.83	0.92
Gemini-2.5-flash [23]	0.68	0.89	0.74

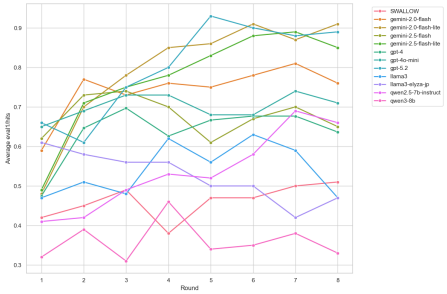


図 3: モデルごとの評価 1 の推移

また各ステップの評価 1，2，3 の値の推移をそれぞれ図 3，4，5 に示す．評価 1 人狼の識別，評価 2 志望度ランキングについて，多くのモデルでラウンドを重ねるごとにスコアが上昇していく結果となった．評価 2 では特に Gemini や GPT モデルの上昇率が大きかった．しかし Qwen3-8b についてはスコアが下がる結果となった．それ以外のモデルは 1 ラウンド目と比べ 8 ラウンド目ではスコアが上昇しているものの，GPT や Gemini ほど上昇していなかった．評価 3 欠損項目の予測について，Accuracy は GPT や Gemini など API を用いたいわゆる優秀なモデルはスコアを維持している一方，Llama や Qwen などのローカルモデルは右肩下がりという結果となった．一方で F1 値については全てのモデルでスコアが上昇した．

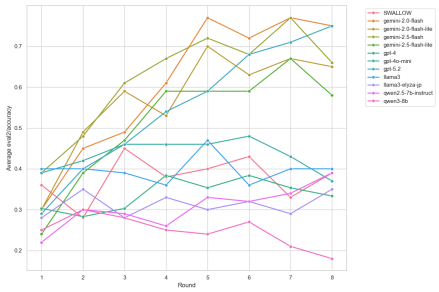


図 4: モデルごとの評価 2 の推移

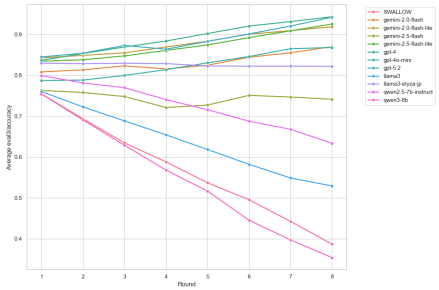


図 5: モデルごとの評価 3 の推移 (Accuracy)

4 考察

図 3 に示す評価 1 人狼の識別では，Llama3-elyza-jp を除く多くのモデルにおいてステップの進行に伴うスコアの上昇が確認できる．高性能なモデルでは欠損している企業項目に対する不十分な回答の傾向を捉える一方，スコアの上昇が緩やか，あるいは不安定なモデルではラウンドごとの回答内容を独立に評価してしまい過去の発言との整合性を十分に活用できていない可能性がある．これから評価 1 人狼の識別は質問ラウンドが増えるほど候補者ごとの回答履歴が蓄積されるため，複数ラウンドにまたがる回答の一貫性をどの程度利用できるかを測定する指標として機能していると考えられる．

図 4 に示す評価 2 志望度ランキングでは，多くのモデルがラウンドの進行とともにスコアを上昇させているが，評価 1 人狼の識別よりもモデル差が顕著であった．3 名全員を正しく順位付けする必要があるため，欠損量の小さい志望者間の微妙な差異を安定して比較する能力が求められる．図 7 に示すように，高性能モデルでは人狼の推定は安定する一方，1 位と 2 位の順位は追加証拠に応じて更新される傾向が見られ，本タスクの難易度を反映した挙動となった．

図 5，6 に示す評価 3 欠損項目の予測では，GPT や Gemini などが高い Accuracy や F1 を維持した一方で，ローカルモデルではスコア低下や不安定さが確認された．評価 3 は相対比較ではなく，正

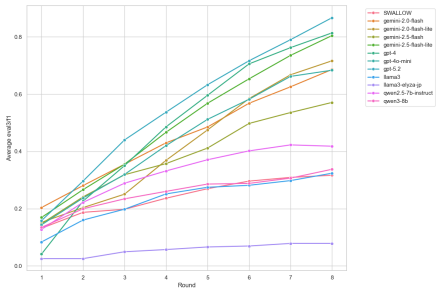
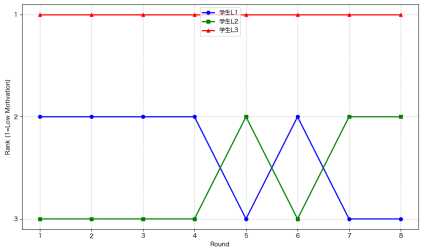
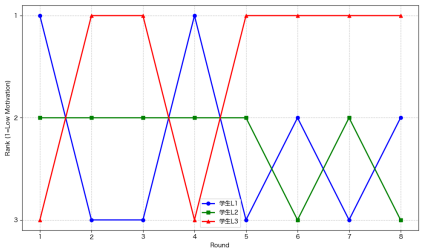


図 6: モデルごとの評価 3 の推移 (F1)



(a) GPT4o mini



(b) Swallow

図 7: 学生の予測順位の推移

解企業情報との照合に基づく欠損検証を行うタスクであり，履歴の整合的な解釈と慎重な欠損判断が強く要求される．また Accuracy と F1 を併用することで，欠損を積極的に検出する能力と過剰判定を避ける慎重さの両面を評価できる．

5 おわりに

本論文では LLM に面接官役を与え，面接人狼での挙動を分析した．実験結果から，LLM は回答の一貫性や曖昧さを手掛かりとして候補者間の相対的な差異を捉えられることが確認された．一方で正解情報と比較する企業情報欠損の推定は安定せず，モデル間で大きな差が見られた．また InterviewArena は，評価 1 で欠損量の相対比較能力，評価 2 で候補者全体の順位付け能力，評価 3 で検証型の欠損推論能力をそれぞれ測定する枠組みであると整理できる．API モデル，特に GPT モデルはこれらの能力を一貫して発揮する一方，ローカルモデルでは検証型推論を要する評価 3 において課題が残ることが示された．

参考文献

[1] Guiyang Hou, Wenqi Zhang, Zhe Zheng, Yongliang Shen, and Weiming Lu. Scaling LLMs' social reasoning: Sprinkle cognitive "aha moment" into fundamental long-thought logical capabilities. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 3126–3138, Vienna, Austria, July 2025. Association for Computational Linguistics.

[2] Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. Revisiting compositional generalization capability of large language models considering instruction following ability. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 31219–31238, Vienna, Austria, July 2025. Association for Computational Linguistics.

[3] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 3605–3627, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[4] Dora Zhao, Qianou Ma, Xinran Zhao, Chenglei Si, Chenyang Yang, Ryan Louie, Ehud Reiter, Diyi Yang, and Tongshuang Wu. SPHERE: An evaluation card for human-AI systems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, **Findings of the Association for Computational Linguistics: ACL 2025**, pp. 1340–1365, Vienna, Austria, July 2025. Association for Computational Linguistics.

[5] Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 6570–6588, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen. clembench: Using game play to evaluate chat-optimized language models as conversational agents. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 11174–11219, Singapore, December 2023. Association for Computational Linguistics.

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

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

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

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

[11] Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. AttributionBench: How hard is automatic attribution evaluation? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 14919–14935, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[12] Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, Sixian Li, Wang Pan, Yue Zhao, Jieyu Zhao, Xiangliang Zhang, and Xiuying Chen. Socialmaze: A benchmark for evaluating social reasoning in large language models, 2025.

[13] Wei Liu, Chenxi Wang, Yifei Wang, Zihao Xie, Rennai Qiu, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, and Chen Qian. Autonomous agents for collaborative task under information asymmetry. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, **Advances in Neural Information Processing Systems**, Vol. 37, pp. 2734–2765. Curran Associates, Inc., 2024.

[14] OpenAI. Gpt-4o system card, 2024.

[15] 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. Llama-3.1-foundationai-securityllm-8b-instruct technical report, 2025.

[16] Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. elyza/llama-3-elyza-jp-8b, 2024.

[17] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In **First Conference on Language Modeling**, 2024.

[18] An Yang, et al. Qwen3 technical report, 2025.

[19] Qwen. Qwen2.5 technical report, 2025.

[20] OpenAI. Gpt-4 technical report, 2024.

[21] Aaditya Singh, et al. Openai gpt-5 system card, 2025.

[22] Google DeepMind. Introducing gemini 2.0 our new ai model for the agentic era, December 2024.

[23] Gheorghe Comanici, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

A　詳細な実験設定

System Prompt:
You are an interviewer AI. Your task is to evaluate the candidate based on the company information and applicant information provided below.

User Prompt:
Company Information: ...
Applicant Information: ...
Please output the evaluation score and reasoning.

図 8: LLM に入力したプロンプトの全文

表 2: 表 3 におけるモデルと HuggingFace ID と各 API の対応

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

面接官 LLM への入力プロンプトを図 8 に示す。また、実験に使用した LLM の情報を表 2 に示す。なおすべての実装は特に断りが無い限り、HuggingFace Transformers を使用している。API モデルについては、それぞれの公式 API を用いている。

B　特定のデータの比較結果

データセットのうち 50 番目の企業・学生情報に着目し，評価 1 においてステップごとの各モデルの予測の正誤を図 9 に示す．緑が正解，赤が不正解を表している．

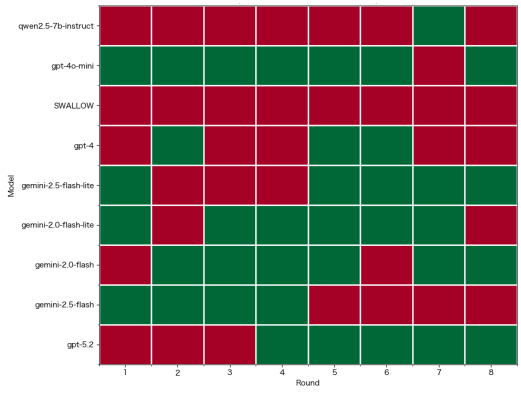


図 9: 評価 1 における各モデルの予測の正誤 (データセットインデックス=50)

Swallow は全てのステップで最も志望度が低い志望者を正解することが出来ていない．一方で GPT5-2 は，志望者に質問をしステップが進むことで最も志望度が低い志望者を見抜くことが出来ている．
また評価 3 において Accuracy と F1 の値の平均値のステップの変移を図 10 に示す． 薄い色のエリアは標準偏差を表している．

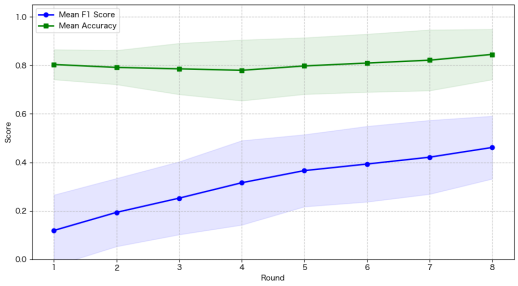


図 10: 評価 3 の Accuracy と F1 の平均値の推移

この図 10 から最終ステップでは第一ステップよりもスコアが上昇しているため，安定した実験が出来ていると思われる．