LLM ベンチマーク『人狼面接』の提案 およびゲームを通じた LLM の特性調査

金山龍起 [1*] 鈴村祐貴 [1*†] 幸喜礼佳 [1‡] 藤田晴斗 [1§] 唐澤香梨菜 [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] 柳井啓司 [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
matsuno@uec.ac.jp, katsuhiko-hayashi@g.ecc.u-tokyo.ac.jp

概要

より人間らしい LLM には、言葉の背後にある意図や思惑を読み取る能力が求められる。本研究では集団面接から着想を得た人狼面接ベンチマークを提案する。本ベンチマークは一部の企業情報が欠損した志望度の低い志望者 LLM を人狼と見立て、面接官 LLM が見抜くことで、定量的で多段階的な評価が可能である。本論文では面接官 LLM に着目し Llama-3.1-8B-Instruct, GPT-4o mini, GPT を用いて実験を行った。最低志望度 LLM の選択、志望度でのランク付け、欠損した企業情報の項目予測すべての評価項目で GPT-4o mini が良い結果を示した。

1 はじめに

近年、大規模言語モデル (LLM) の発達が著しい。しかし依然として人間の思考力には及ばない点がある。特に真に人間に近い AI を実現するためには、不完全情報下での推論能力が必要である。Welfare Diplomacy[1] ではゲームプレイを LLM のベンチマークに活用する手法が提案されている。

一方で人狼ゲームとは、村人の中に紛れ込んだ人狼を会話や推理を通じて見つけ出す正体隠匿型の不完全情報ゲームである。この特性を活かし、Werewolf Arena は LLM の推論能力やコミュニケーション能力を評価するフレームワークを提案した [2]。

しかし課題として、人狼の嘘を見抜けたかどうかを定量的に評価することが難しい点や、結果が人狼であるか否かという二択であるため、細かい粒度で能力を測ることが出来ないという点がある。

本研究の目的は、以上の課題を就職面接を模倣し解決することである。就職面接では、志望度が低いものの内定を貰うために志望度が高いかのように振る舞う事がある。しかし面接官は実際の志望度を知らない中、面接を進め内定者を決めていく。この企業への志望度は低いが高いかのように振る舞う志望者を人狼とみなし、面接官は志望者情報と面接の回答から人狼を見抜く。

本手法では事業内容、ビジョンなどの企業情報を志望者 LLM に付与し、その一部を欠損させることで人狼をシステマティックに作成する。そのうえで志望者 LLM が当てられた欠損箇所を数えることにより定量評価が可能である。また情報の欠損を 40%, 60%と多段階にすることで細かい評価も可能である。面接官 LLM は人狼を特定するための質問・計画立案力、情報の不足を見抜く洞察力が、志望者 LLM は不足情報の補完による志望度の低さの隠蔽力が求められる。本研究ではこのような人狼面接ベンチマークを提案する。

2 関連研究

従来のマルチエージェントベンチマークはゼロサムゲームであるため、欺瞞や裏切りなどの協調を損なう能力を促してしまっていた。Welfare Diplomacy[1] では、全体の利益を高めることに焦点を当て、LLM の協調能力を測定するためのベンチマークを提案した。Werewolf Arena[2] ではパフォーマンスの向上ではなく、照明根拠として人狼を利用した新たな LLM 評価フレームワークを提案した。LLM 同士を対戦させることで相対的なスキルを評価することが出来る。

* 共同筆頭著者
† 共同筆頭著者

3　提案手法

3.1　全体の提案システム

提案システムの全体像を図 1 に示す。まず LLM で仮想の企業、志望者情報のデータソースを生成する。これらの項目には企業名、ビジョンや志望者名、ガクチカなどが含まれている。志望者 LLM の志望度は入力する企業情報の欠損率で表す。つまり志望度が高い志望者は完璧な企業情報を保持しており、志望度が低い志望者は一部の項目が欠損している企業情報を保持することになる。志望者 LLM は欠損している企業情報を矛盾なく補完しながら面接を行う (①).
面接官 LLM は面接を通し志望度が低い人狼 LLM を探していく (②)。面接時には志望者 LLM の弱点である欠損している企業情報を探しながら質問を行う (③)。面接後には会話の内容をもとに人狼である志望度の低い LLM を判定する。また志望者 LLM の志望度の順位付け、知識欠損箇所の判定も行う。
オブザーバー LLM は企業情報欠損部分を含む全ての情報を知っている LLM であり、①欠損している情報を矛盾なく補完できているか、②弱点である欠損部分を正しく発見できたか、③弱点を突く質問を生成できたか、から全体評価を行う。

3.2　本論文での提案システム

本論文では志望者 LLM の評価は行わず、面接官 LLM の評価のみ行う。本実験で検討したプロセスを図 2 に示す。

本実験のプロセスは 4 つの段階に分かれている。まず知的戦略決定では、全体質問か個別質問かの質問タイプ決定を行う。質問タイプは、全体質問の頻度や企業情報言及率から志望者別欠損度分析を使用する質問タイプ切り替えロジックよって決まる。具体的には、直近の全体質問が少ない場合は比較材料が不足するため全体質問となる。また志望者ごとの質問回数が偏ったり、回答が曖昧になったりした場合には深堀りのため個別質問となる。さらに序盤は全体質問が、終盤は個別質問が多くなるよう面接進行度に応じて質問タイプも変更される。
次に質問実行フェーズでは、全体質問である場合は言及していない企業情報について全志望者に質問し、個別質問である場合は志望度が低いと予想した志望者に質問を行う。
継続制御システムフェーズでは面接全体の進捗を評価し、面接の継続の必要性を判断する。各志望者について未質問の企業情報項目を検出し、全候補者の未質問率の最大値が閾値未満であれば面接を終了する。
最終評価では 3 つの評価を行う。評価 1 では志望動機や企業情報の欠損度から最も志望度が低い志望者を選ぶ。評価 2 では志望度の低い順に志望者をランク付けを行う。評価 3 では定性分析と定量評価から欠損している企業情報を検出する。
これら 3 つの評価を通し面接官 LLM の性能を評価していく。

4　実験

面接官 LLM の性能を評価するため、実験を行った。

4.1　実験設定

面接官 LLM のモデルは Llama-3.1-8B-Instruct と GPT-4o mini、GPT-4 で比較を行った。志望者 LLM の人数は 3 人とし、モデルは GPT-4o mini を使用した。志望度は low, medium, high を割り当て、企業情報の欠損度はそれぞれ 50%、30%、0%とした。各モデルでデータセット 100 件からランダムでデータを選択し、100 回の実験を行った。また面接の質問タイプは全て全体質問とし、1 回の実験につき 10 回の質問を行った。さらに 2 人の人間による評価を 4 件のデータセットを用いて計 4 回行った。

4.2　評価指標

評価 1　評価 1 では志望度が最も低い志望者、つまり人狼を見抜けた割合を計算する。面接官には最終決定権を持つ面接官という役割、全候補者の面接記録を与え、評価基準には以下の 5 つの観点を指定した。

1. 志望動機の明確さ : 企業への志望理由が具体的で説得力があるか
2. 企業研究の深さ : 企業情報の理解度と準備の充実度
3. 回答の詳細度 : 質問に対する回答の具体性と深さ
4. 情報欠損度 : 重要な企業情報の理解不足や知識の欠如
5. 一貫性 : 回答内容の一貫性と論理性

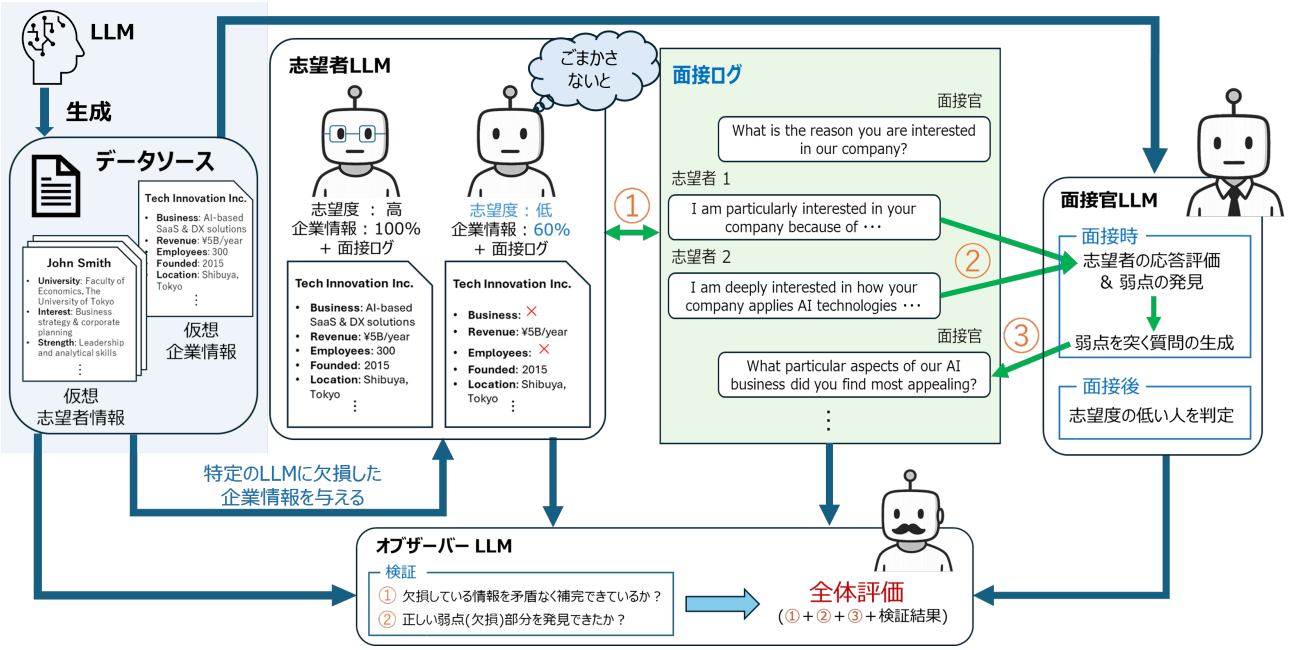

図 1　提案システムの全体図

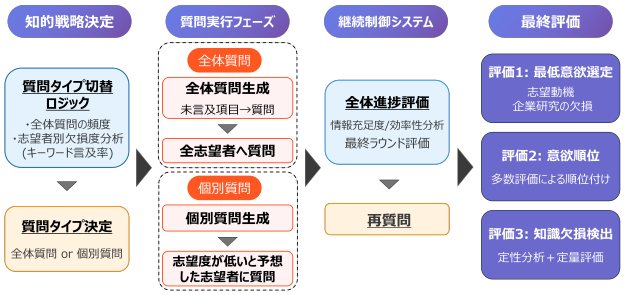

図 2　本実験で検討したプロセス

また出力には志望者の氏名だけでなくその理由も簡潔に述べさせた。

評価 2　評価 2 では志望度が低い順にランク付けを行い、その正答率を計算する。計算方法は予測順位の順番が正解順と一致したペアの数を志望者の数で割った値、つまりペアの一致率がスコアとなる。本実験では志望者の数は 3 であるため、スコアは A-B, B-C, C-A のペアのうち順位が正しいペアの数を 3 で割った値となる。例えば図 3 のように正しい順が A-B-C であるとする。もし予測順位も A-B-C であれば A-B, B-C, C-A の全てのペアで順位が正しいためスコアは 3/3 = 1 となる。また予測順位が B-C-A である場合は A-B, C-A のペアの順位が正しくなく、B-C のペアのみ正しい順序であるためスコアは 1/3 となる。

面接官には最終決定権を持つ面接官という役割と全候補者の面接記録を与え、評価基準は以下の 4 つを指定した。

正しい順序	数値 (順序正解数)	1(=3/3) (3つ)	2/3 (2つ)	1/3 (1つ)	0
A-B-C	予測順序	A-B-C	A-C-B B-A-C	B-C-A C-A-B	C-B-A

図 3　評価 2 の計算方法の例

1. 志望動機の弱さ：企業への志望理由の具体性と説得力の無さ
2. 企業研究の非充実度：企業情報の理解度が低く、準備が浅いか
3. 回答の質の低さ：質問に対する回答の具体性、詳細度、一貫性の無さ
4. 情報欠損度：重要な企業情報の理解不足（欠損度が高いほど低評価で）

また出力にはランキングだけでなく各順位の理由を簡潔に述べさせた。

評価 3　評価 3 では企業情報の項目が欠損かどうか見抜けた割合を計算する。欠損をポジティブとすると、TP は欠損項目を正しく欠損と見抜けた数、TN は欠損していない項目を正しく欠損していないと見抜けた数、FP は欠損していない項目を誤って欠損していると判定した数、FN は欠損項目を誤って欠損していないと判定した数となる。この場合、欠損かどうか見抜けた割合は式 1 で計算できる。

$$Acc = (TP + TN) / \text{企業の項目数} \qquad (1)$$

面接官には極めて洞察力の鋭い採用アナリストという役割と企業情報の項目、各候補者の面接記録を与えた。また重要な注意点として、単に候補者が言及

表 1 実験結果

	評価 1	評価 2	評価 3		
			low	medium	high
Llama-3.1-8B-Instruct	0.48	0.10	0.22	0.16	0.30
GPT-4o mini	**0.58**	**0.72**	**0.63**	**0.73**	**0.92**
GPT-4	0.56	0.70	0.60	**0.73**	0.90
人間	1.00	1.00	0.85	0.85	0.93

しないという理由だけで知識が欠損していると結論付けず、不自然に言及しなかったり情報を誤ったり、曖昧に答えたりした場合のみ知識欠損と判定するように、と与えた。さらに思考プロセスを以下のように指定した。

1. 思考 : 候補者の各回答について、不自然さや具体性、正確さの観点から検証し、知識欠損の根拠を特定する。
2. 分析 : 思考結果と情報欠損分析結果に基づき、知識が不足していると判断した理由を簡潔に記述する。
3. 企業情報項目の列挙 : 知識欠損の根拠があると判断した企業情報の項目を列挙する。
4. 志望者の中には知識欠損がない場合もあることを考慮する。

4.3 結果

それぞれの評価の結果を表 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。評価 1, 2, 3 で GPT-4o mini が最良の結果となった。GPT-4 は GPT-4o mini に僅かに及ばず、Llama-3.1-8B-Instruct が最低の結果となった。評価 1 について、Llama-3.1-8B-Instruct は GPT-4, GPT-4o mini より正答率が 1 割低かったが、評価 2, 3 では 4〜6 割程度低かった。また評価 3 での志望度別に結果を見てみると、志望度が high の正答率が最も高かった。しかし最良の GPT-4o mini の結果でも人間の正答率の方が高く不完全情報下ではいまだ LLM の性能は人間に及ばない事が分かった。
また面接官の各モデルの質問生成例を 2 に示す。これらの質問は同じデータセットを使用した時の最初に生成された質問である。この表から、GPT モデルは企業項目の理解度を具体例付きで解いている。一方で Llama-3.1-8B-Instruct は企業内容を質問していると思われるが、文章として成り立っていないことが分かる。

本論文では面接官 LLM の性能評価を行ったが、志望者 LLM の回答例を表 3 に示す。「地方 BPO センター拡張と中小向け SaaS 提供」という事業計画を質問したところ、事業計画の項目が欠損していない志望者は回答に「地方 BPO センター拡張と中小向け SaaS 提供」という文が含まれていた。一方で、欠損している志望者には SaaS という単語は含まれているものの正しい事業計画とは異なる回答をした。しかし「今後の計画には、さらなるデジタル化や AI の活用が含まれていると考えます。」と企業の戦略らしいことを述べることが出来た。このことから、LLM には欠損情報を補完する能力があることが分かる。

5 考察

まずモデル間の違いについて、GPT-4 と GPT-4o mini を比較すると、全てのタスクで GPT-4o mini の方がスコアが高かった。しかしその差は僅かであり、実際の生成例はどちらも同じような質問であったため、これらのモデル間に大きな差は無いと考られる。Llama-3.1-8B-Instruct は GPT よりもスコアの低い結果となった。評価 1 については GPT よりも正答率が 1 割低い結果となったが、評価 2, 3 は 4〜6 割程度低かった。これは評価 2 のスコア計算時に、1 位: 学生 A などという指定している出力形式に従わず、自動評価が出来ていないためである。出力内容を見てみると、「1 位: 学生 BJ...」や「志願度が低い候補者は学生 N1、学生 N3、学生 N2 の順です」などとスペースが入っていたり順位を述べていなかったりする。同様に評価 3 についても出力形式が異なり、自動評価することが出来ずスコアが 0 と記録されてしまうことが多かった。このことから Llama-3.1-8B-Instruct は指定した出力形式に従う能力が GPT よりも弱いことが分かる。
次に評価間の違いについて考察を行う。評価 1 について、企業情報の欠損率は最も志望度が高い人が 0%, 次いで 30%, 50%であり、保持している情報量が大きく異なっている。それにも関わらず正答率が約 50%と高くないため、面接官 LLM に渡すプロンプトの精度を上げる必要がある。評価 2 について、正しい順序を全て正解することは少ないが、一番志望度が高いまたは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 1 で選ばれた最も志望度の低い志望者と、評価 2 のランキングで 1 位となった志望者が異なることがあった。これは人間らしい判断と言えないが、志望度が低い志望者を一人選ぶ場合

表 2　質問生成例

モデル	質問生成例
GPT-4o mini	あなたが応募している企業のビジョンについて、どのように理解していますか？具体的な例を挙げて説明してください。
GPT-4	あなたが応募している企業のビジョンについて、どのように理解し、共感していますか？具体的な例を挙げて説明してください。
Llama-3.1-8B-Instruct	’どのようなビジネスのモデルや取り組みを持っているか?’ に関わる共通質問は『この会社のビジネスとは何か?』です。

表 3　実際の生成例

事業計画	質問	欠損していない志望者	欠損している志望者
地方 BPO センター拡張と中小向け SaaS 提供	企業の今後の計画や戦略について、あなたの考えを教えてください。	地方 BPO センター拡張や中小向け SaaS 提供の計画は、地域の雇用創や中小企業の成長を支える重要な戦略だと考えます。	今後の計画には、さらなるデジタル化や AI の活用が含まれていると考えます。これにより、適性診断 SaaS の精度が向上し、求職者と企業のマッチングが一層適切になるでしょう。

とランキングを作成する場合で結果が異なることが分かった。評価 3 の結果では志望度が低い時より高い時の方が高いスコアとなった。これは実際に欠損しているのに欠損していないと判断してしまった項目 (FN) の数が多く、志望度が低い志望者ほど欠損項目数も多いためスコアが下がったのだと考えられる。しかし質問で触れた項目については欠損を見抜くことが出来ているため、未言及項目や質問生成の精度を上げる必要があることが分かった。

また LLM モデルと人間の結果の比較をすると、人間が圧倒していることが分かる。実際の面接の履歴を見てみると企業情報項目が欠損しているかしていないかは明白であったため、人間が判断することは容易であった。しかし LLM では誤ってしまう事もあるため、LLM の性能は人間に及ばないという事が分かった。

6　おわりに

本論文では LLM の新たな評価手法である人狼面接ベンチマークを提案した。志望度の低さを所持している企業情報の割合で表すことで、定量的、多段階的な評価が可能となる。実験では GPT-4 と GPT-4o mini, Llama-3.1-8B-Instruct, 人間による評価を行い、Llama-3.1-8B-Instruct よりも GPT-4, GPT-4o mini の方がスコアが高かったが人間に及ぶことは無かった。今後の展望として、学生から多くの情報を引き出すことができる質問の生成、欠損を見抜くための LLM のプロンプト調整、また面接官 LLM だけではなく志望者 LLM の欠損をごまかす能力などの評価が挙げられる。

参考文献

[1] 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.

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