LLM を用いたペルソナ情報収集インタビューにおける人間・疑似対象者間の比較

長谷川遼[1] 土田陸斗[1] 園田哲也[2] 宇津呂武仁[1]
[1]筑波大学大学院システム情報工学研究群 [2]山梨大学工学部
{s2420791,s2520796}_@u.tsukuba.ac.jp, t22jm030_@yamanashi.ac.jp,
utsuro_@iit.tsukuba.ac.jp

概要

本論文は，半構造化インタビューに基づく LLM インタビューシステムを用いて，人間インタビュー対象者および LLM 疑似インタビュー対象者を対象に，ペルソナ属性推定性能を評価した．また，ペルソナ属性推定を行う LLM に few-shot プロンプトを導入した場合の性能についても評価した．実験として，3 つのドメインにわたる合計 36 件のインタビューを行った．zero-shot プロンプトでは，LLM 疑似対象者と比較して人間対象者に対するペルソナ属性推定がより困難であることが示された．また，few-shot プロンプトにより，人間対象者に対する推定性能が向上することを確認した．few-shot により，ペルソナ属性推定が改善される事例が確認された．

1 はじめに

大規模言語モデル (LLM) [2] の発展により，対話システムの活用が進んでいる．先行研究 [10] では，LLM を用いたユーザシミュレータを疑似インタビュー対象者として半構造化インタビュー [6, 7, 29] を行う場合，システムが半構造化インタビューを効果的に実施できることが示されている．これらの研究では，ユーザシミュレータは事前に定義されたペルソナ設定に基づいて応答を生成し，想定範囲外の質問に対しては「わかりません。」と応答するように設定されている．一方で，人間のインタビュー対象者は，曖昧な表現を用いたり，複数の情報を一つの発話に含めたり，インタビュアの意図した話題から逸脱したりすることが多く，LLM の疑似インタビュー対象者と比べて応答が多様になる．したがって，人間インタビュー対象者に対するインタビューを行うことは，インタビューシステムの有効性や課題を検証するうえで有用である．本論文では，図 2 に示す半構造化インタビューを用いて，人間インタビュー対象者と LLM の疑似インタビュー対象者に対するインタビューを実施する．提案するシステムでは，インタビュー対象者のペルソナ属性を推定する役割を担う LLM が，対話履歴およびスロットに記録された情報に基づき，各対話ターンごとにインタビュー対象者の属性を推定する．本システムの有効性を検証するため，人間インタビュー対象者とのインタビューと LLM の疑似インタビュー対象者とのインタビューにおけるペルソナ属性推定性能と，例示を提示する few-shot をプロンプトに導入することによるペルソナ属性推定性能を評価する (図 1)．さらに，few-shot をプロンプトに取り入れた場合に結果が改善された事例に対して，事例分析を行う．本論文の貢献は以下のとおりである．

1. 半構造化インタビューの枠組みに基づき，本システムが人間インタビュー対象者および LLM の疑似インタビュー対象者と対話する実験を実施し，両者のインタビューにおける本システムのペルソナ属性推定性能を明らかにした．
2. ペルソナ属性推定を行う LLM に対して，zero-shot および few-shot のプロンプトを導入し，few-shot プロンプトが人間インタビュー対象者との対話におけるペルソナ属性推定性能を向上させることを示した．
3. few-shot プロンプトの導入によってペルソナ属性推定の出力結果が改善された事例に対して事例分析を行い，few-shot プロンプトが有効に機能する場合を示した．

2 関連研究

対話システムの分野において，インタビュー対話には，質問生成，対話管理，対話戦略など多様な要素が含まれており，これまでに幅広い研究が行われ

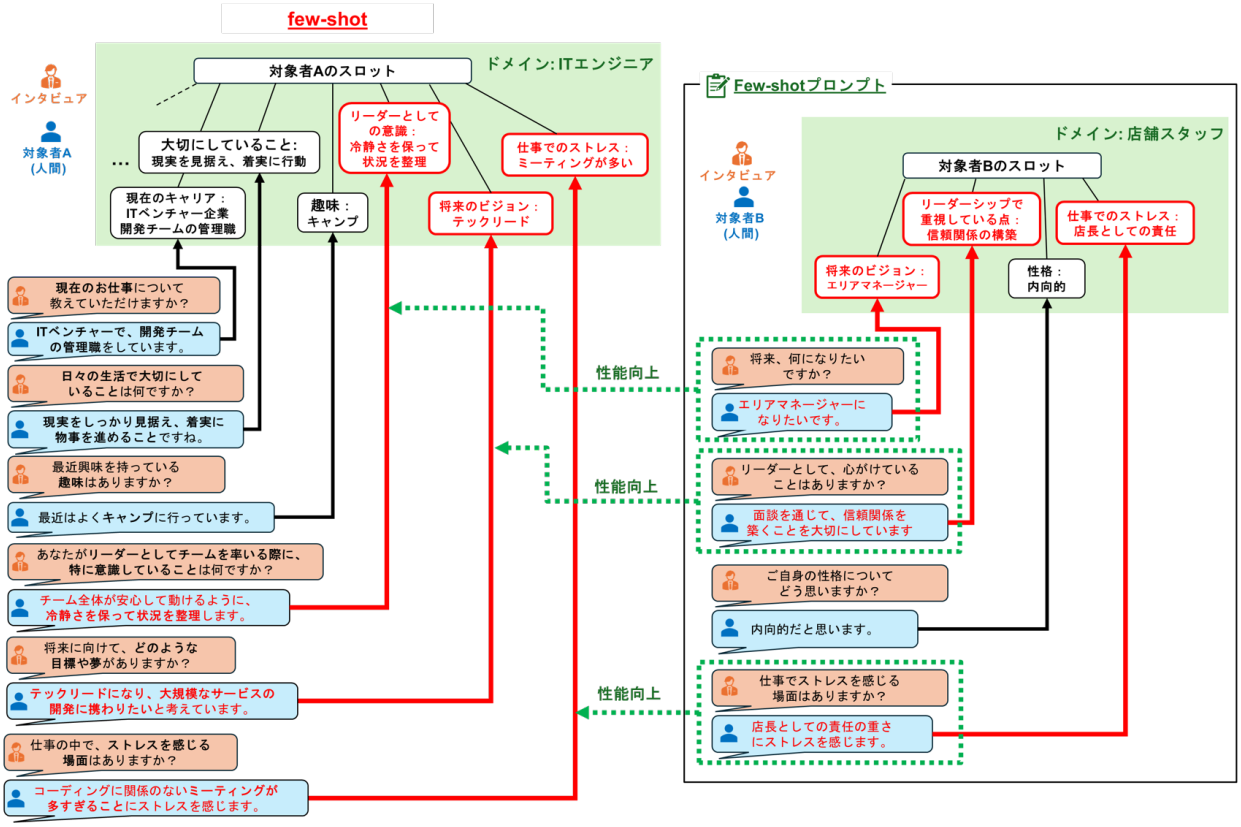

図 1　few-shot プロンプトを使用した半構造化インタビュー

てきた [30, 4, 17, 18, 1, 20, 8]. 就職面接を対象とした対話システムの研究も挙げられる [24, 25]. 文献 [15] は，アンドロイドロボットがインタビュアーを務めるシステムを開発した．文献 [32] は，知識グラフに基づいて質問を生成する料理嗜好インタビューシステムを提案している．文献 [22] は，対話戦略の最適化を目的としてユーザ応答をシミュレーションする統計的モデリング手法を提案している．近年，LLM の急速な発展により，対話システムの性能の向上が見られる [9]. 特に，LLM をスロットフィリング型アーキテクチャと組み合わせることで，高度な言語理解能力と柔軟な対話制御を両立できることが示されている [14, 16, 23, 3]. さらに，LLM の高い文脈理解能力を活用することで，状況に応じたインタビュー対話が可能となっている [26, 5, 12, 27]. また，文献 [28] は，LLM とスロットによる対話管理を組み合わせることで，対話の流れを細かく制御できることを示している．さらに，動的スロット生成に関する研究も行われている [19, 11, 31]．半構造化インタビューを対象とした研究はこれまでにも行われてきたが [21, 13]，多くは静的，あるいは人手で設計された制御に依存している．本論文は，LLM 疑似インタビュー対象者を用いたシミュレーションにとどまらず，人間インタビュー対象者を対象とした半構造化インタビューへと実験対象を広げる．

3　半構造化インタビューによるペルソナ属性推定

ペルソナは，インタビュー対象者の人物像や役割を表す情報である．インタビューを通じて収集された情報は，システム内でスロットとして保持される．各スロットは，スロット名と値の対で構成される．スロット名は抽象的なトピックや属性カテゴリを表し，値にはインタビュー対象者の発話内容が格納される．インタビューの各ターンにおいて，システムは新たなスロットの生成および，対話履歴に基づく既存スロットの値の更新を行う．本システムの構成要素の一つとして，対話履歴およびスロットに基づいて，インタビュー対象者のペルソナ属性を推定する LLM を導入する．この LLM は，ターンごとに対話履歴と現在保持しているスロットから，インタビュー対象者のペルソナ属性を推定する．推定結果である推定ペルソナ属性には，当該属性に関する質問が未実施か，あるいは既に実施済みかを示すラベルが付与されている．これにより，未質問の推定ペルソナ属性に関連する質問を優先的に生成する

表 1　ペルソナ属性推定の評価結果 (再現率/適合率/F 値,「IT エンジニア」ドメイン)

	zero-Shot		few-Shot	
対象者	LLM	人間	LLM	人間
1 人目	1.00 (2/2) / 1.00 (2/2) / 1.00	0.86 (6/7) / 1.00 (6/6) / 0.92	1.00 (3/3) / 1.00 (3/3) / 1.00	1.00 (10/10) / 1.00 (10/10) / 1.00
2 人目	1.00 (4/4) / 1.00 (4/4) / 1.00	0.85 (11/13) / 1.00 (11/11) / 0.92	0.80 (4/5) / 1.00 (4/4) / 0.89	0.93 (13/14) / 1.00 (13/13) / 0.96
3 人目	0.82 (9/11) / 1.00 (9/9) / 0.90	0.64 (9/14) / 0.90 (9/10) / 0.75	0.89 (8/9) / 0.89 (8/9) / 0.89	0.87 (13/15) / 1.00 (13/13) / 0.93
マクロ平均	0.94 / 1.00 / 0.97	0.78 / 0.97 / 0.86	0.90 / 0.96 / 0.93	0.93 / 1.00 / 0.96

ことが可能となり，重複した質問を回避するとともに，情報収集の効率を向上させる．

4　人間のインタビュー対象者と LLM による疑似インタビュー対象者の比較

人間のインタビュー対象者は，曖昧な表現を用いたり，インタビュアの質問の意図とは完全には一致しない回答を行ったり，複数の情報を一つの発話にまとめて述べたりすることがある．それに対して，LLM を用いた疑似インタビュー対象者は，定義されたペルソナ設定に基づいて応答を生成し，ペルソナ設定の範囲外の質問に対しては，「わかりません。」と応答するよう設計されている．このような違いは，ペルソナ属性推定の難易度に大きく影響する．特に，インタビュー対象者が人間である場合には，対象者のペルソナ属性の推定がより困難となる．本論文では，人間のインタビュー対象者と LLM による疑似インタビュー対象者を対象としたインタビューについて評価を行う．提案するインタビューシステムにおいて，インタビュー対象者のペルソナ属性を推定する役割を担う LLM のプロンプトに対し，few-shot プロンプトを導入する (図 1)．zero-shot 条件では，ペルソナ属性推定 LLM のプロンプトには例示を与えず，インタビュー対象者の発話内容のみに基づいてペルソナ属性を推定する．これに対し，few-shot 条件では，ペルソナ属性推定のための少数の例示をプロンプトとしてペルソナ属性推定 LLM に与える．これにより，LLM はペルソナ属性を推定する際の判断基準が明示された状態でインタビューを行うことが可能となる．

5　評価

本論文では，提案するインタビューシステムをインタビュアとして用い，インタビュー対象者として LLM の疑似インタビュー対象者または人間対象者のいずれかを設定した．システム内のすべての LLM には，GPT-4o[1)]（gpt-4o-2024-11-20）を使用した．LLM および人間のインタビュー対象者の双方には，再現性確保のためあらかじめ作成したペルソナ設定を事前に与えた．ペルソナは，「IT エンジニア」,「店舗スタッフ」,「先生」の 3 つのドメインに分類した．実験では，各ドメインから 3 種類のペルソナを選択し，インタビューを実施した．LLM の疑似インタビュー対象者は，ペルソナ設定に記載されている内容に関連する質問には記載内容に基づいて応答し，ペルソナ設定に記載されていない話題について質問された場合には，「わかりません。」と回答するよう設計した．対話制御の実装には，LangGraph[2)] を用いた．まず，ペルソナ属性を推定する LLM のプロンプトに zero-shot プロンプトを用いて，人間および LLM のインタビュー対象者の双方に対するインタビューを順番に行った．3 つのドメインそれぞれについて 3 名ずつ順番にインタビューを行い，合計 18 件のインタビューを実施した．次に，few-shot プロンプトを用いて，同一のインタビュー対象者に対して再度インタビューを順番に行った．zero-shot 条件および few-shot 条件を合わせて合計 36 件のインタビューを行った．few-shot プロンプトに含める例示は，評価対象のインタビューと同一になることを避けるため，評価対象のドメインとは異なるドメインから作成した．few-shot 例示は，本実験の前に実施した予備実験に基づいて作成した．評価指標として，再現率，適合率，および F 値を用いた．評価結果を表 1，表 3 に示す．まず，zero-shot 条件における人間インタビュー対象者と LLM の疑似インタビュー対象者の比較から，人間対象者との対話ではペルソナ属性推定が難しくなる傾向が見られた．これらの結果は，人間対象者に対するインタビューにおいて，ペルソナ属性を推定することが難しくなることを示唆している．次に，人間インタビュー対象者では，全ドメインで few-shot 条件が zero-shot 条件を上回った．few-shot プロンプトの導入が人間の

1）https://openai.com/index/hello-gpt-4o/
2）https://github.com/langchain-ai/langgraph

表 2　few-shot 導入前後におけるペルソナ属性推定の改善例 (「IT エンジニア」ドメイン，3 人目)

(a) 実験で使用した few-shot プロンプト例の抜粋 (「先生」ドメイン)

	発話	推定ペルソナ属性
インタビュア	教育現場で、どのような指導方針を大切にされていますか？具体的なエピソードがあればぜひ教えてください。	
対象者	学力だけでなく、社会に出てから通用する礼儀や規律を生徒に身につけさせたいと考えています。	教育現場での指導方針
インタビュア	教育現場で指導を行う際に、どのような方針や考えを大切にされていますか？	
対象者	英語を話せる楽しさを、一人でも多くの生徒に伝えたいと考えています。	教育現場での指導方針

(b) zero-shot 条件と few-shot 条件において行われた対話

	zero-shot 条件の会話	few-shot 条件の会話
インタビュア	あなたがリーダーとしてチームを導く際に、どのようなスタイルやアプローチを大切にされていますか？	あなたがリーダーとしてチームを導く際に、どのようなスタイルやアプローチを大切にされていますか？
対象者	チーム全体が安心して動けるように、冷静さを保って状況を整理します。	しっかりメンバーの面倒を見ることです

(c) LLM による推定結果と人手で付与した正解

zero-shot 条件のペルソナ属性推定結果		few-shot 条件のペルソナ属性推定結果	
人手で付与したペルソナ属性 (正解ペルソナ属性 14 個)	LLM によるペルソナ属性推定結果 (推定ペルソナ属性 10 個)	人手で付与したペルソナ属性 (正解ペルソナ属性 15 個)	LLM によるペルソナ属性推定結果 (推定ペルソナ属性 13 個)
印象に残っている思い出、エピソード	思い出、エピソード	現在のキャリア	現在のキャリア
やりがいを感じる瞬間	やりがい	在宅ワーク	在宅ワーク
生活で大事にしていること	価値観	SNS	SNS
チームリーダとして意識していること	将来の目標	大切にしていること	価値観
将来の目標	性格	リーダーシップスタイル	リーダーシップスタイル
性格	趣味	昇進・転職	昇進・転職
趣味	現在のキャリア	感情のコントロール	感情のコントロール
現在のキャリア	リスク管理	過去のキャリア	過去のキャリア
リスク管理の意識していること	チームメンバーの育成	家族	家族
...	課題解決のアプローチ	...	...

発話に対するペルソナ属性推定の性能向上に寄与することが示唆される．一方，LLM の疑似インタビュー対象者では，店舗スタッフドメインにおいて few-shot 条件によってマクロ平均の F 値の改善が見られた一方で，「IT エンジニア」および「先生」ドメインではマクロ平均の F 値が僅かに低下する場合も確認された．さらに，人間インタビュー対象者に対する few-shot 導入による改善は，特に再現率の向上として現れる傾向が確認された．表 2 は，「IT エンジニア」ドメインにおけるペルソナ属性推定結果について，zero-shot 条件と few-shot 条件を比較した事例を示している．zero-shot 条件下の会話では，インタビュー対象者は「チーム全体が安心して動けるように，冷静さを保って状況を整理する」と回答しているものの，LLM はこの発話から「チームリーダとして意識していること」に該当するペルソナ属性を推定できていない．一方，few-shot 条件では，同様にリーダーとしての姿勢を示す発話が得られており，この場合には，LLM が「リーダーシップスタイル」という対応するペルソナ属性を正しく推定できていることが分かる．さらに，「店舗スタッフ」ドメインにおいても，few-shot プロンプトを用いることで，将来志向の発話がペルソナ属性「将来のキャリアプラン」と適切に推定できた事例を 2 例確認した．この結果は，few-shot によって，LLM がどのような発話がどのペルソナ属性に対応するかという判断基準を事前に与えられたことで，適切な属性推定が可能になったことを示唆している．

6　おわりに

本論文では，半構造化インタビューに基づく LLM インタビュアーシステムを実装し，対話履歴およびスロット情報からインタビュー対象者のペルソナ属性を推定する性能を評価した．評価の結果，LLM 疑似インタビュー対象者に比べ，人間インタビュー対象者との対話では，ペルソナ属性推定が難しくなることが確認された．一方で，人間対象者に対しては few-shot プロンプトにより，すべてのドメインで推定性能が向上し，特に再現率の改善として現れる傾向が得られた．他方，LLM の疑似インタビュー対象者では few-shot の効果が一貫せず，改善が限定的または低下する場合も見られた．また，事例分析から，例示によりペルソナ属性推定が改善される事例を確認した．

謝辞

本論文は，一部，科研費 25K03416 の支援を受けたものである．

参考文献

[1] Pooja Rao S B, Manish Agnihotri, and Dinesh Babu Jayagopi. Automatic follow-up question generation for asynchronous interviews. In **Proc. IntelLanG**, pp. 10–20, 2020.

[2] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. In **Proc. 33rd NeurIPS**, pp. 1877–1901, 2020.

[3] Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. Span-ConveRT: Few-shot span extraction for dialog with pretrained conversational representations. In **Proc. 58th ACL**, pp. 107–121, 2020.

[4] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. Simsensei kiosk: a virtual human interviewer for healthcare decision support. In **Proc. 13th AAMAS**, p. 1061–1068, 2014.

[5] Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. Towards LLM-driven dialogue state tracking. In **Proc. EMNLP**, pp. 739–755, 2023.

[6] NG Fielding, editor. **Interviewing**, Vol. I. Sage Publications, 2003.

[7] NG Fielding, editor. **Interviewing**, Vol. II. Sage Publications, 2003.

[8] Yubin Ge, Ziang Xiao, Jana Diesner, Heng Ji, Karrie Karahalios, and Hari Sundaram. What should I ask: A knowledge-driven approach for follow-up questions generation in conversational surveys. In **Proc. 37th PACLIC**, pp. 113–124, 2023.

[9] Friedrich Geiecke and Xavier Jaravel. Conversations at scale: Robust ai-led interviews with a simple open-source platform. **Social Science Research Network**, pp. 1–73, 2024.

[10] Ryo Hasegawa, Yijie Hua, Takehito Utsuro, Ekai Hashimoto, Mikio Nakano, and Shun Shiramatsu. A dialogue system for semi-structured interviews by LLMs and its evaluation on persona information collection. In **Proc. IWSDS 2025**, pp. 39–59, 2025.

[11] Ekai Hashimoto, Mikio Nakano, Takayoshi Sakurai, Shun Shiramatsu, Toshitake Komazaki, and Shiho Tsuchiya. A career interview dialogue system using large language model-based dynamic slot generation. In **Proc. 31st COLING**, pp. 1562–1584, 2025.

[12] Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In **Proc. 61st ACL**, pp. 936–950, 2023.

[13] Jiaxiong Hu, Jingya Guo, Ningjing Tang, Xiaojuan Ma, Yuan Yao, Changyuan Yang, and Yingqing Xu. Designing the conversational agent: Asking follow-up questions for information elicitation. **Proc. ACM Hum.-Comput. Interact.**, Vol. 8, No. CSCW1, April 2024.

[14] Vojtěch Hudeček and Ondrej Dusek. Are large language models all you need for task-oriented dialogue? In **Proc. 24th SIGDIAL**, pp. 216–228, 2023.

[15] Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. Job interviewer android with elaborate follow-up question generation. In **Proc. ICMI**, p. 324–332, 2020.

[16] Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. "do you follow me?": A survey of recent approaches in dialogue state tracking. In **Proc. 23rd SIGDIAL**, pp. 336–350, 2022.

[17] Michael Johnston, Patrick Ehlen, Frederick G. Conrad, Michael F. Schober, Christopher Antoun, Stefanie Fail, Andrew Hupp, Lucas Vickers, Huiying Yan, and Chan Zhang. Spoken dialog systems for automated survey interviewing. In **Proc. 14th SIGDIAL**, pp. 329–333, 2013.

[18] Takahiro Kobori, Mikio Nakano, and Tomoaki Nakamura. Small talk improves user impressions of interview dialogue systems. In **Proc. 17th SIGDIAL**, pp. 370–380, 2016.

[19] 駒田啓伍, 阿部香央莉, 守屋彰二, 鈴木潤. 柔軟な応対が求められる対話タスクでの動的スロットを用いた対話整合性向上. 第 38 回人工知能学会全国大会論文集, 2024.

[20] Fuminori Nagasawa, Shogo Okada, Takuya Ishihara, and Katsumi Nitta. Adaptive interview strategy based on interviewees' speaking willingness recognition for interview robots. **IEEE Transactions on Affective Computing**, Vol. 15, No. 3, pp. 942–957, 2024.

[21] Angelina Parfenova. Automating the information extraction from semi-structured interview transcripts. In **Proc. WWW**, p. 983–986, 2024.

[22] Jost Schatzmann and Steve Young. The hidden agenda user simulation model. **IEEE Transactions on Audio, Speech, and Language Processing**, Vol. 17, No. 4, pp. 733–747, 2009.

[23] A.B. Siddique, Fuad Jamour, and Vagelis Hristidis. Linguistically-enriched and context-awarezero-shot slot filling. In **Proc. WWW**, p. 3279–3290, 2021.

[24] Ming-Hsiang Su, Chung-Hsien Wu, and Yi Chang. Follow-up question generation using neural tensor network-based domain ontology population in an interview coaching system. In **Interspeech**, pp. 4185–4189, 2019.

[25] Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang, Qian-Bei Hong, and Huai-Hung Huang. Follow-up question generation using pattern-based seq2seq with a small corpus for interview coaching. In **Interspeech**, pp. 1006–1010, 2018.

[26] Guangzhi Sun, Shutong Feng, Dongcheng Jiang, Chao Zhang, Milica Gasic, and Phil Woodland. Speech-based slot filling using large language models. In **Findings of ACL**, pp. 6351–6362, 2024.

[27] 鈴木順大, 石垣龍馬, 宿里晃太郎, 藤本拓真, 河窪大介, 酒造正樹, 前田英作. ただ一つのプロンプトによるタスク指向型対話システムの実現. 言語処理学会第 30 回年次大会論文集, pp. 2720–2725, March 2024.

[28] Nicolas Wagner and Stefan Ultes. On the controllability of large language models for dialogue interaction. In Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani, editors, **Proc. 25th SIGDIAL**, pp. 216–221, September 2024.

[29] Tom Wengraf. **Qualitative Research Interviewing**. Sage Publications, 2001.

[30] Jie Zeng, Yukiko Nakano, and Tatsuya Sakato. Question generation to elicit users' food preferences by considering the semantic content. In **Proc. 24th SIGDIAL**, pp. 190–196, 2023.

[31] Yuki Zenimoto, Mariko Yoshida, Ryo Hori, Mayu Urata, Aiko Inoue, Takahiro Hayashi, and Ryuichiro Higashinaka. Automated administration of questionnaires during casual conversation using question-guiding dialogue system. In **Proc. SEMDIAL**, pp. 115–124, 2025.

[32] 曽傑, 中野有紀子. 知識と話題の埋め込み表現に基づく質問生成と対話システムへの適用—料理嗜好インタビューシステムに向けて—. 自然言語処理, Vol. 28, No. 2, pp. 598–631, 2021.

A　構造化インタビュー・半構造化インタビュー

構造化インタビューと半構造化インタビューのイメージ図を図 2 に示す.

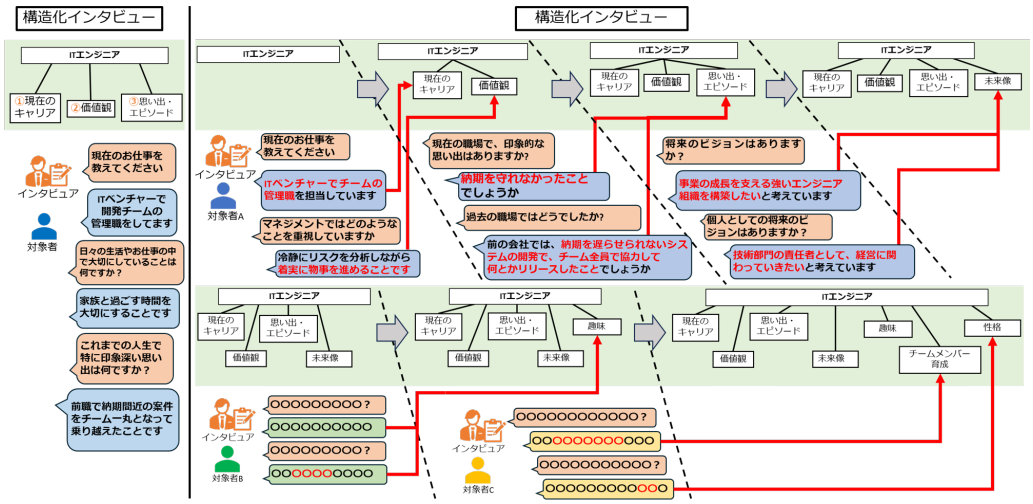

図 2　構造化インタビュー・半構造化インタビューのイメージ図 [10]

B　LLM による半構造化インタビュー

本システムは，複数の LLM から構成されており，それぞれの LLM が異なる処理を担当する．具体的には，各モジュールは，質問生成，新たなスロットの生成，インタビュー対象者の発話内容に基づくスロットの値の記録・更新といった機能をそれぞれの LLM が行う．本システムではインタビューを通じて得られた情報をスロットとして保持し，各ターンごとに更新することで，これまでに取得された情報を管理する．また，一連のインタビューを通じてスロットを蓄積する設計とすることで，蓄積されたスロットを後続のインタビューで利用できるようにする．

C　評価結果の詳細

表 3 は，「店舗スタッフ」ドメインと「先生」ドメインにおける，LLM の疑似対象者と人間対象者に対するペルソナ属性推定の再現率・適合率・F 値を示したものである．

表 3　ペルソナ属性推定の評価結果 (再現率/適合率/F 値)

(a) 「店舗スタッフ」ドメイン

	zero-Shot		few-Shot	
対象者	LLM	人間	LLM	人間
1 人目	0.71 (5/7) / 1.00 (5/5) / 0.83	0.56 (5/9) / 1.00 (5/5) / 0.71	0.80 (5/7) / 1.00 (5/5) / 0.89	1.00 (4/4) / 1.00 (4/4) / 1.00
2 人目	0.86 (6/7) / 1.00 (6/6) / 0.92	0.58 (7/12) / 0.88 (7/8) / 0.70	0.88 (7/8) / 1.00 (7/7) / 0.93	1.00 (5/5) / 0.83 (5/6) / 0.91
3 人目	0.75 (6/8) / 1.00 (6/6) / 0.86	0.91 (10/11) / 0.77 (10/13) / 0.83	1.00 (9/9) / 1.00 (9/9) / 1.00	0.88 (7/8) / 1.00 (7/7) / 0.93
マクロ平均	0.77 / 1.00 / 0.87	0.68 / 0.88 / 0.75	0.89 / 1.00 / 0.94	0.96 / 0.94 / 0.95

(b) 「先生」ドメイン

	zero-Shot		few-Shot	
対象者	LLM	人間	LLM	人間
1 人目	1.00 (3/3) / 1.00 (3/3) / 1.00	0.78 (7/9) / 1.00 (7/7) / 0.88	1.00 (5/5) / 1.00 (5/5) / 1.00	1.00 (3/3) / 0.75 (3/4) / 0.86
2 人目	0.86 (6/7) / 1.00 (6/6) / 0.92	0.91 (10/11) / 0.91 (10/11) / 0.91	0.89 (8/9) / 1.00 (8/8) / 0.94	1.00 (11/11) / 1.00 (11/11) / 1.00
3 人目	0.90 (9/10) / 1.00 (9/9) / 0.95	0.77 (10/13) / 0.83 (10/12) / 0.80	0.85 (11/13) / 1.00 (11/11) / 0.92	1.00 (11/11) / 0.92 (11/12) / 0.96
マクロ平均	0.92 / 1.00 / 0.96	0.82 / 0.91 / 0.86	0.91 / 1.00 / 0.95	1.00 / 0.89 / 0.94