LLM を用いた対話エージェントの開発とユーザ体験向上の検討

Development of a Communication Agent Using LLM and Examination of User Experience

中山裕翔 [1]　峯岸朋弥 [1]

Yuto Nakayama[1] and Tomoya Minegishi[1]

[1] 専修大学

[1] Senshu University

Abstract: 本研究では，対話を行うバーチャルエージェントにおいて，自然言語モデルを活用した発話がユーザー体験の向上に与える影響について検討する．大学内施設の案内を行うバーチャルエージェントを開発し，案内を受けるタスクを設定した．自然言語モデルから生成された文章を用いて案内を行う条件と，予め用意された文章を読み上げることにより案内を行う条件を用意し，アンケートを実施した．分析から，LLM を用いたエージェントは技術的態度，適応性，順応性の点で優れていると評価された．

1　序論

人工知能技術の進歩により，人とインタラクションを行うエージェントシステムの開発が加速しており，その応用範囲が広がりつつある．特に大規模言語モデル（以下，LLM）は，ユーザとの対話を効果的に行う手段として注目されている．これに伴い，自然言語を発話することによりインタラクションを行うエージェントシステム（以下，対話エージェント）が，顧客サービスや教育・医療分野など広く開発されている[1–3]．また応用先として例えばデジタルゲームにも LLM を利用する動きがある[4, 5]．従来においても対話エージェントは研究されており，その有効性が検証されている一方で，これらは LLM を使用しないため，対話エージェントによる返答が，制作者が予め用意した定型文になる点，制作者が想定していない対話に対応できない点の問題が挙げられる．これにより，ユーザがサービスとして対話エージェントを利用した際に感じる印象（以下，ユーザ体験）の向上が難しい課題がある．この課題を解決するために，LLM へ事前知識を膨大に与えることや対話エージェント以外の代替手段の構築が考えられるが，実際に制作する際にこのような手段をとることは時間的制約等から困難である場合が多い．

本研究は，簡易的に事前知識を与えた LLM を用いた対話エージェントが，自然言語の発話によりインタラクションを行うことがユーザ体験に影響を与えるかを検証する．具体的には，LLM を使用する条件と使用しない条件を用意し，大学内のキャンパス案内を想定したタスクを実施し，それに対するアンケートを実施した．その中で，参加者の発話文字数，質問回数，アンケートを分析し，簡易的な事前知識を与えた LLM による対話エージェントの効果を明らかにする．

2　関連研究

LLM を用いた対話エージェント研究が行われている一方で，事前に LLM に知識を簡易的に与え，これが十分に人のユーザ体験向上に有効か検証された例は少ない．Nie ら[6]は，LLM に専門知識を与えられていない課題に対処するため，検索エンジンを組み合わせたシステムを提案している．Nandy ら[7]は，対話エージェントシステムのインタフェースをリアルタイムで生成する方法を提案している．これらの研究により，ユーザ体験向上の可能性が示唆されている一方で，事前に対話エージェント開発者が準備する事柄が多く，簡易的な知識提供がユーザ体験向上に寄与するか研究された例は少ない．

3　開発システム

ユーザが LLM を用いた対話エージェントと自然言語を発話することを用いて会話を行うことを実現

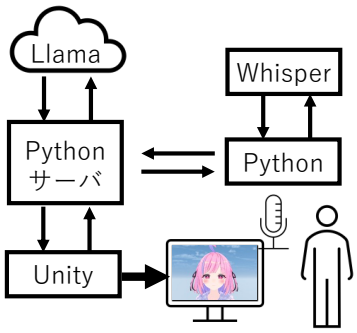

図 1　システム構成図

する．また開発する対話エージェントを，コンピュータの画面上で仮想的に動作させることを想定する．この機能の実現のために，新たに画面上で動作可能な対話エージェントを作成した．概要を図 **1** に示す．

システムはマイク入力を受け付ける Python プログラムを経由し，ユーザからの発話を OpenAI 社により提供される音声認識ライブラリである Whisper に送る．Whisper によりユーザの発話内容から文字起こしする．この文字列は Python により構築された基幹サーバに送られ，サーバは音声から認識された文字起こしからプロンプトを作成し，LLM へ送る．なお使用する LLM は，事前実験により応答速度と生成内容が高く評価された Meta 社による Llama-3.1 を使用する．本モデルは API によりクラウドサービスとして利用可能である．対話エージェントは画面上で動作させることを想定するため，3D モデルの制御を得意とする Unity を使用する．さらに，対話エージェントによるユーザへの発話のために，VOICEVOX を使用しテキストを音声ファイルに変換する．変換されたファイルは Unity で再生するが，その際に Lip Sync 技術を用いて，対話エージェントの発話内容と口の動きを同期させる．

柔軟なタスクへの対応を実現するため，開発したシステムの LLM には簡易的に事前知識教示を行った．教示内容を表 1 示す．User は対話の相手である参加者，Assistant は対話エージェントを示す．予め想定される質問と回答例を示すことにより，教示内容に類似する質問と回答を行うことができ，参加者へ柔軟な対応が可能になる．LLM を使用しない条件においては，事前教示内容を変化させず一定で発話する．

表 1　事前教示内容の具体例

User	Assistant
大学の食堂はどこですか？	大学の食堂は4号館の2階と5号館の3階にあるよ。
図書館はどこですか？	10号館の3階にあるよ。
コンビニはありますか？	3号館の2階にセブンイレブンがあるよ。
おすすめの休憩場所は？	休むなら10号館3階の広場がおすすめだよ。
眺めの良いスポットは？	10号館の屋上は街が見渡せる良いスポットだよ。

4　実験

4.1　実験方法

LLM を用いた対話エージェントを用いて，大学内に設置された対話エージェントから大学キャンパス案内を受ける場面を想定したタスクを設定した．参加者は実験実施者から，対話エージェントに対して大学に関する質問とそれ以外の任意の質問を行うよう指示される．

実験参加者からの発話内容により発話内容を変化させる対話エージェント(以下「LLM あり条件」)と，予め実験者が用意した音声ファイルを発話する対話エージェント(以下「LLM なし条件」)の 2 条件を用

表 2　アンケート項目

ID	質問内容
Q1	このエージェントを利用することは良いアイデアだと思う
Q2	このエージェントは人生をより面白いものにしてくれそうだと思う
Q3	このエージェントを活用することは良いことだと思う
Q4	このエージェントは私の必要とすることに応じることができると思う
Q5	このエージェントは私が必要と思う時に助けてくれるだろうと思う
Q6	このエージェントは私がその時必要と思うことだけをしてくれると思う
Q7	このエージェントに話しかけられるのは楽しい
Q8	このエージェントと一緒に何かすることは楽しい
Q9	このエージェントは面白い
Q10	このエージェントは魅惑的だ
Q11	このエージェントは退屈だ
Q12	このエージェントは私にとって役に立つと思う
Q13	このエージェントを所有することは私にとって便利になると思う
Q14	このエージェントは多くのことで私を助けてくれると思う
Q15	このエージェントがアドバイスをくれたら、私はこのエージェントを信頼するだろう
Q16	私はこのエージェントがくれたアドバイスに従うだろう

意して実験を行った．LLM あり条件は事前に簡易的な教示を行った LLM を用いて発話内容を確定させる．LLM なし条件は，事前に実験者が用意した音声ファイルを再生するのみとし，LLM を一切使用しない条件である．参加者が実験者の想定しない質問を行った場合「その質問にはお答えできません」と回答する．

実験の参加者は，専修大学に在籍する 15 名の学生である．各参加者には，LLM あり条件と LLM なし条件の両方を経験してもらい，それぞれの経験後にアンケートに回答してもらった．カウンターバランスを考慮するため，参加者により実施条件の順序を入れ替えた．

4.2 仮説

事前に簡易的な知識を教示した LLM を用いることにより，参加者の質問へ柔軟な対応が可能になることが考えられる．これにより参加者は対話エージェントに興味を示し，好意的な印象を抱く可能性がある．また興味を示すことにより，対話エージェントの使用可数が増加する可能性がある．よって本研究では以下の仮説を設定し検証した．

H1. LLM あり条件においてユーザ体験が向上する

H2. LLM あり条件において対話エージェントの使用頻度が増加する

4.3 評価項目

設定した仮説を検証するため，本研究ではユーザ体験としてアンケート，ユーザの対話エージェント使用頻度として会話文字数，質問回数を条件間で比較する．アンケート項目は Heerink ら[8]の項目を一部抜粋し使用する．参加者は 16 項目の質問に対して，5 段階(1: 全くそう思わない〜5:とてもそう思う)により評価する．実際に使用したアンケート項目を表 2 に示す．本アンケートにより，技術的態度，適応性，楽しさ，有用性，信頼性の 5 つを評価可能である(Q1~Q3: 技術的態度，Q4~Q6: 適応性，Q7~Q11: 楽しさ，Q12~14: 有用性，Q15~16: 信頼性)．

表 3 アンケート項目の比較
(* は $p < .05$, ** は $p < .01$,
*** は $p < .001$ を示す)

	LLMあり		LLMなし		
	Mean	SE	Mean	SE	p
技術的態度	4.22	0.09	3.64	0.14	0.012 *
適応性	3.73	0.13	3.18	0.14	0.029 *
楽しさ	4.21	0.08	3.32	0.12	4.24e-07 ***
有用性	3.78	0.11	3.24	0.11	0.005 **
信頼性	3.43	0.16	3.3	0.15	2.587

表 4 発話文字数・質問回数の比較

	LLMあり		LLMなし		
	Mean	SE	Mean	SE	p
文字数	238.64	53.63	206.64	21.80	0.734
質問回数	15.36	2.30	14.43	1.54	0.954

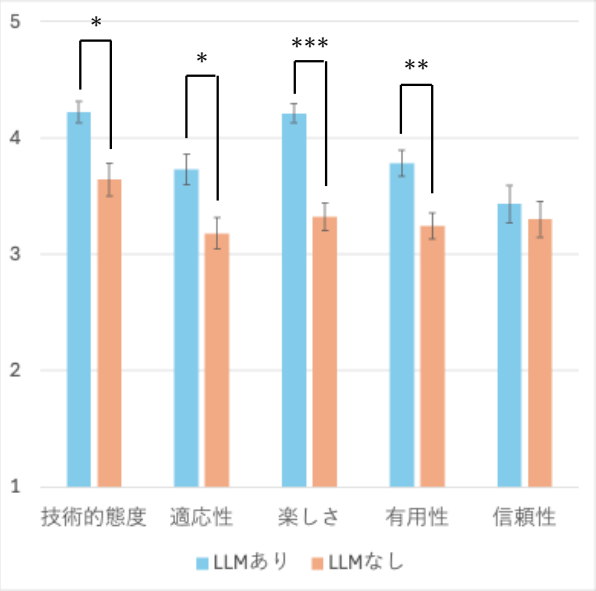

図 2 アンケートの分析結果（エラーバーは Mean±SE を示し，* は $p < .05$, ** は $p < .01$, *** は $p < .001$ を示す）

5 分析結果

アンケートについて条件間で比較を行うため，Wilcoxon の符号順位和検定を行い，Bonferroni 法により p 値を補正した．結果を表 3 と図 2 に示し，以下より平均値を Mean，標準誤差を SE と示す．

分析の結果，技術的態度において，LLM あり条件のほうが有意に高く評価された（LLM あり条件：Mean = 4.22, SE = 0.09, LLM なし条件：Mean = 3.64, SE = 0.14, $p < .05$）．また，適応性において，LLM あり条件のほうが有意に高く評価された(LLM あり条件：Mean = 3.73, SE = 0.13, LLM なし条件：Mean = 3.18, SE = 0.14, $p < .05$)．楽しさにおいて，LLM あり条件のほうが有意に高く評価された（LLM あり条件：Mean = 4.21, SE = 0.08, LLM なし条件：Mean = 3.32, SE = 0.12, $p < .001$）．さらに，有用性において，LLM あり条件のほうが有意に高く評価された(LLM あり条件：Mean = 3.78, SE = 0.11, LLM なし条件：

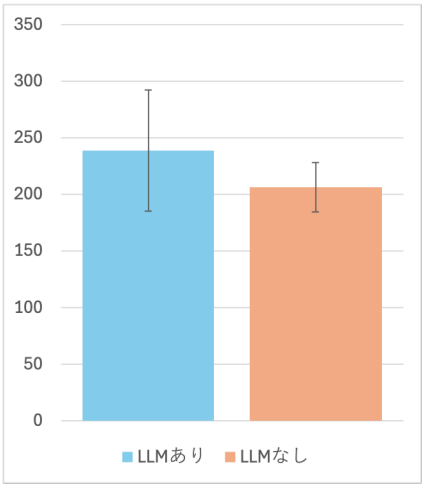

図 4　発話文字数の分析結果
（エラーバーは Mean±SE を示す）

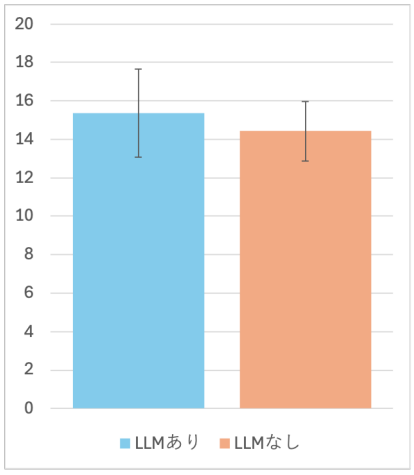

図 4　質問回数の分析結果
（エラーバーは Mean±SE を示す）

Mean = 3.24, SE = 0.11, $p < .01$).

参加者が実験中に発話した文字数と質問回数について条件間で比較を行うため，Wilcoxon の符号順位和検定を行った．結果を表 4 に示す．分析の結果，発話文字数において有意な差は認められなかった (LLM あり条件：Mean = 238.64, SE = 53.63, LLM なし条件：Mean = 206.64, SE = 21.80, $p = .734$, n.s.)．質問回数において条件間で有意差が見られなかった (LLM あり条件：Mean = 15.36, SE = 2.30, LLM なし条件：Mean = 14.43, SE = 1.54, $p = .954$, n.s.).

6　考察

技術的態度, 適応性, 楽しさ, 有用性において LLM あり条件のほうが高く評価された．実験中の記録を確認したところ，LLM あり条件において，参加者による質問意図に合致する回答を発話できた場面が多く行うことができていた．また，LLM あり条件の方が，発話する回答内容の情報量が多い場面が存在した．LLM あり条件において，生成された予測不能な回答を参加者が楽しんでいる様子が見られたことから，参加者が返答を予測できないという点が楽しさに寄与したと考察する．参加者が求める回答以上の内容を含め発話できたことが有効に作用したと考えられる．

一方で，信頼性に関する評価に有意差が見られなかった．生成内容の記録を確認したところ，LLM が誤った情報を含む内容を生成する場合があることが確認された．開発した対話エージェントは簡易的な事前教示を行っているが，これでは不十分な可能性がある．以上より H1 は一部支持された．

文字数，質問回数には有意差が見られなかった．実験の設計上，2 条件において同程度の質問を行うべきであるという無意識的なバイアスを参加者に与えた可能性がある．実際に実験後のインタビューにおいて，「もっと話してみたい」と回答する参加者が存在した．よって，仮説 H2 は支持されなかったが，これは実験の設計をもう一度見直す必要があると考えられる．

7　まとめ

本研究では，簡易的に事前知識を与えた LLM を用いた対話エージェントがユーザ体験に与える影響を検討した．実験の結果，LLM を用いる条件において，技術的態度，適応性，楽しさ，有用性の点で有意に高く評価された．LLM により，参加者が求める質問に対して適切な回答を生成することや，付加情報を提供できることが，ユーザ体験の向上に寄与したと考えられる．一方で，信頼性に関する評価には有意差が見られず，実験の記録から，LLM が生成する内容に誤情報が含まれる場合もあった．信頼性は，ユーザ体験の向上に必要な要素である．発話文字数や質問回数に有意差が見られなかった．これらの結果は，実験設計に関する見直しが求められることを示している．参加者からのフィードバックでは「もっと話してみたい」という意見もあり，今後はより自由度の高いインタラクションが促進できる設計を検討する．さらに，発話内容の生成にも工夫が必要である．

参考文献

[1]　B. Li *et al.*, “Conversational AI in health: Design

considerations from a Wizard-of-Oz dermatology case study with users, clinicians and a medical LLM," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, May 2024, pp. 1–10. doi: 10.1145/3613905.3651891.

[2] J. He *et al.*, "Frontiers of Large Language Model-Based Agentic Systems - Construction, Efficacy and Safety," *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 5526–5529, 2024, doi: 10.1145/3627673.3679105.

[3] M. Rinott and O. Shaer, "Temporal Aspects of Human-AI Collaborations for Work," *ACM Int. Conf. Proceeding Ser.*, no. 1, 2024, doi: 10.1145/3663384.3663397.

[4] B. Bateni and J. Whitehead, "Language-Driven Play: Large Language Models as Game-Playing Agents in Slay the Spire," in *ACM International Conference Proceeding Series*, New York, NY, USA: ACM, May 2024, pp. 1–10. doi: 10.1145/3649921.3650013.

[5] 村上一真, 森直樹, "ユーザの嗜好を反映した人工知能キャラクターとの共同生活シミュレーションシステム", 日本デジタルゲーム学会 夏季研究発表大会予稿集, 2024 夏季研究発表大会, p. 19-24, 2024.

[6] G. Nie *et al.*, "A Hybrid Multi-Agent Conversational Recommender System with LLM and Search Engine in E-commerce," in *18th ACM Conference on Recommender Systems*, New York, NY, USA: ACM, Oct. 2024, pp. 745–747. doi: 10.1145/3640457.3688061.

[7] P. Nandy *et al.*, "Bespoke: Using LLM agents to generate just-in-time interfaces by reasoning about user intent," in *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, New York, NY, USA: ACM, Nov. 2024, pp. 78–81. doi: 10.1145/3686215.3688372.

[8] M. Heerink, B. Kröse, V. Evers, and B. Wielinga, "Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere Model," *Int. J. Soc. Robot.*, vol. 2, no. 4, pp. 361–375, Dec. 2010, doi: 10.1007/s12369-010-0068-5.