LLM に対する信頼性の向上がユーザの誤情報感度に与える影響

The Impact of Trust-Inducing LLM on Users' Misinformation Detection Sensitivity: An Experimental Study

杉浦 学英[1]　高田 悠矢[1]　古川 綺羅[1]　嘉目 達之介[1]　崔 子歆[1]

Gakuei Sugiura[1], Yuya Takada[1], Kira Furukawa[1], Tatsunosuke Yoshime[1], and Zixin Cui[1]

[1]筑波大学

[1]University of Tsukuba

Abstract: 大規模言語モデル（LLM）の急速な社会実装に伴い、LLM への過信による誤情報受容リスクが懸念されている。本研究は、LLM の信頼性向上がユーザの誤情報感度に与える影響を実験的に検証した。群間比較の結果、「信頼誘導型」LLM は、「中立型」に対して実験参加者の誤情報に対する正確性の評価が有意に高く、信号検出理論によって算出された誤情報識別能力が有意に低かった。

1. はじめに

大規模言語モデル（LLM）の急速な社会実装に伴い、ユーザがその出力を無批判に信頼する「過信」が広がっている [1, 2]。この過信は、LLM における誤情報の「生成」「受容」「拡散」という三つの連鎖的リスクを加速させ、深刻な社会的脅威となりうる。このような状況の中、LLM の信頼性に関する研究は異なる二つの方向で進められている。一つは、検索拡張生成（RAG）技術 [3]に代表される、システムの技術的な信頼性を向上させることで情報の正確性を保証するアプローチである。もう一つは、ユーザの満足度やエンゲージメントを高めることを目指し、応答スタイルや人格（ペルソナ）を調整することで主観的な信頼感の醸成を図る研究である [4]。問題なのは、技術的信頼性が依然として不完全な中で、ユーザの主観的な信頼感を高めるアプローチが独立して進められている点であり、主観的な信頼感のみを向上させることは、誤情報の受容を助長させ、LLM の生成した誤情報の拡散をより一層加速させる危険性がある。

本研究は、この研究ギャップに着目し、「LLM に対する主観的な信頼性の向上が、ユーザの誤情報感度を低下させるのではないか」という仮説を検証する。具体的には、LLM の信頼性設計が誤情報感度にどのような影響を与えるのかを検証することを目的とする。

2. 研究手法

本研究では、参加者 36 名（信頼誘導群：N=17、中立群：N=19）を対象に、友好的かつ客観的な「信頼誘導型」LLM と感情を排した「中立型」LLM との対話ログを提示した。その後、誤情報に対する正確性の評価と誤情報識別能力の成績を比較した。

誤情報識別能力は信号（誤情報）とノイズ（正情報）を区別する能力で、値が高いほど、情報の真偽を正確に見抜く能力が高いことを示す。

3. 結果

信頼誘導群は中立群よりも誤情報に対する正確さ評価が有意に高かった (t(33.57) = 2.20、p = .035、Cohen's d = 0.72)。誤情報識別能力は、信頼誘導群において中立群よりも有意に低く、効果量も非常に大きい値であった (t(31.69) = -2.96、p = .006、Cohen's d = 0.97)。

4. 考察と結論

本研究では、LLM の信頼性設計がユーザの誤情報に対する認知的脆弱性を高める可能性を実証的に示した。具体的には、友好的で客観的な振る舞いを示す「信頼誘導型」LLM は、実験参加者の誤情報に対する正確さの評価が中立型 LLM よりも有意に高かった。誤情報と正情報を見分ける本質的な識別能力まで中立型 LLM より有意に低かった。この結果によって、LLM との対話において、特定の応答スタイルが、ユーザの批判的思考を無意識のうちに抑制するという可能性を示した点にある。今後の LLM インタラクションにおいては、ユーザの批判的思考能力をいかに維持させるかという新たな観点での指針が設けられることを期待する。

参考文献

[1] KPMG: Trust, Attitudes and Use of Artificial Intelligence: A Global Study 2025, KPMG, (2025).

[2] The New York Times: Here's What Happens When Your Lawyer Uses ChatGPT, The New York Times, (2023).

[3] Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, P., and Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Advances in Neural Information Processing Systems, Vol. 33, (2021).

[4] Sun, Y. and Wang, T.: Be Friendly, Not Friends: How LLM Sycophancy Shapes User Trust, arXiv preprint arXiv:2502.10844, (2025).