2024 年度年次報告書
次世代 AI を築く数理・情報科学の革新
2024 年度採択研究代表者

上田 亮

東京大学 大学院情報理工学系研究科
大学院生

# マルチモーダル表現学習としての創発的 LLM 間コミュニケーション

## 研究成果の概要

研究計画初年次であることや、研究者にとって比較的新しい試みを含む研究テーマであることから、本年度は主に先行研究の調査に取り組んだ。特に、人間と大規模言語モデル（LLM）が共生できる社会の実現のための研究に取り組むことを目的として、LLM がゲーム理論的な枠組みの中で協調関係を創発できるかどうかについての研究について調査を行った。

結果として、Li & Shirado (2025) の研究事例に着目するに至った。Li & Shirado (2025) では、Chain-of-Thought (CoT) Prompting や Relfection のような、LLM の最先端の推論能力の基礎となるような技術を用いた場合において、LLM が利己的になるという驚くべき結果を示している。CoT 等の手法を用いて LLM に「熟考」させるのは、LLM を賢くふるまわせるための定石として認識されているが、これによって、かえって LLM が利己的になってしまうのであれば、将来の LLM 利用の場面においてリスクになりうる。予備的な実験において Li & Shirado (2025) の実験の一部の再現を試みたところ、たしかに同様の傾向がみられた。次年度においては、どのようにすれば熟考するタイプの LLM であっても利他的に振る舞わせ、協調関係を築かせることができるのかについて検討する予定である。

参考文献

Li, Y., & Shirado, H. (2025). Spontaneous Giving and Calculated Greed in Language Models. *arXiv preprint arXiv:2502.17720*. Retrieved from https://arxiv.org/abs/2502.17720