2023 年度年次報告書
社会課題を解決する人間中心インタラクションの創出
2023 年度採択研究代表者

高木 優

大阪大学 大学院生命機能研究科
助教

大規模言語モデルとヒト脳の相互理解と新たなインタラクション創出

研究成果の概要

ChatGPT をはじめとする大規模言語モデルと人類とのインタラクションが驚異的な速度で普及している中、その中身を人間中心に理解することや、理解をもとにした新たなインタラクションを構築することは急務である。本研究では、大規模言語モデルとヒト脳との対応を取ることで、相互の理解を促進し、新たな脳・大規模言語モデルインタラクションのあり方を創出することを狙う。それぞれに関して以下の計画で予備的な実験を行った。

1. 大規模言語モデルとヒト脳活動をインタラクションさせることによる脳の理解
音声、画像、言語といったマルチモーダルな刺激を提示中に fMRI で取得されたヒト脳活動データを用い、ヒトの脳活動に表現されている情報を大規模言語モデルによって読み取る研究を行った。本年は、GPT-2 や Llama2、その派生モデルを用いて、刺激に内在する言語表現に関する大規模言語モデル内部の潜在表現をどの程度読み取れるかを検証した。その結果、サイズの大きなモデルほど脳活動をよく読み取れる傾向を確認した。

2. インタラクション中の大規模言語モデルの内部状態を、ヒトの脳を介して理解
ヒトに提示されている刺激に関連する様々な刺激を言語情報に変換し、それらの言語情報を入力として条件づけられた（インタラクション中の）大規模言語モデルの内部表現から、ヒトの脳活動を予測することを行う。大規模言語モデル内部の様々な“脳部位“からヒト脳を全域にわたって予測することで、大規模言語モデルの脳内がヒトの脳内とどのように対応しているかを明らかにすることを狙った。本年は、GPT-2 や Llama2、その派生モデルを用いて、脳活動の予測精度を比較した。その結果、モダリティによって異なる脳部位がよく予測できることや、モデルサイズが大きなほど脳活動がよく予測できる傾向を確認した（Nakagi et al., 2024 プレプリント）

【代表的な原著論文情報】

1) Yuko Nakagi, Takuya Matsuyama, Naoko Koide-Majima, Hiroto Yamaguchi, Rieko Kubo, Shinji Nishimoto, Yu Takagi, The Brain Tells a Story: Unveiling Distinct Representations of Semantic Content in Speech, Objects, and Stories in the Human Brain with Large Language Models, *bioRxiv*, 06.579077, 2024