ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：
信頼できる大規模言語モデルの研究

実施時期：2025/07-2026/03
所属機関名：北陸先端科学技術大学院大学
代表者氏名：井之上直也

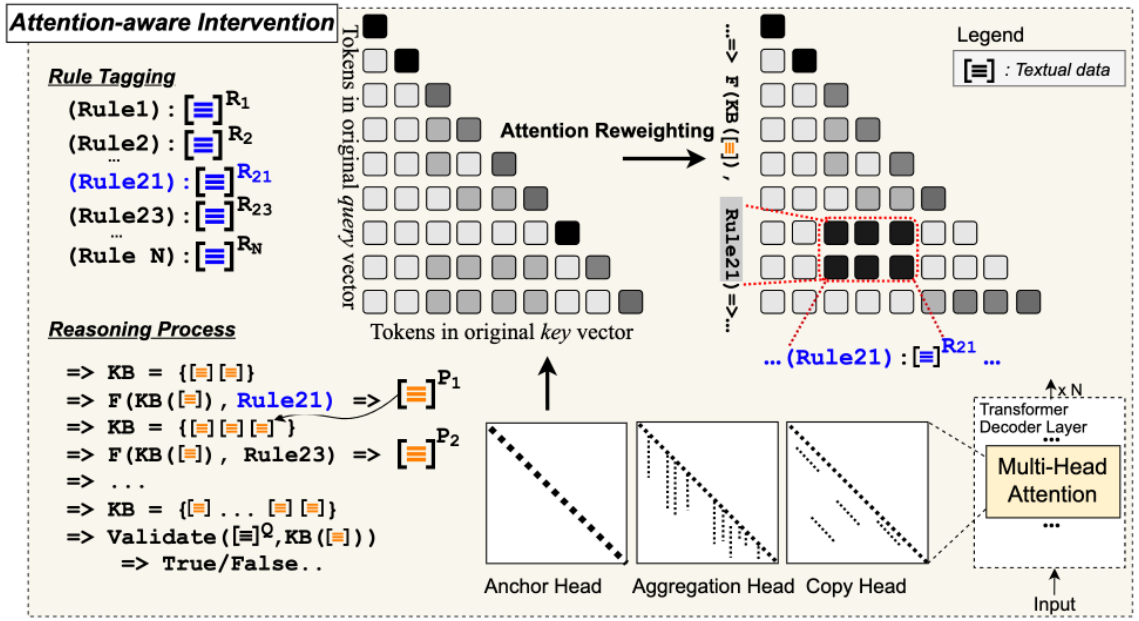
成果概要：

本課題では、信頼性の高い言語処理システムの実現を目指し、最先端の大規模言語モデル（LLM）に対して、論理推論、エージェントの指示理解の各観点から内部機序の解明に取り組んだ。具体的には、記号的情報を活用したChain-of-Thought推論の改善、指示曖昧性の内部表象解析を行い、LLMがどのように情報を表現し、推論し、指示の不確実性を内部で表現しているかに関する基礎的知見を得た。

成果のポイント：

① 論理推論における思考連鎖手法の改善
大規模言語モデル（LLM）の論理推論能力向上を目的に、思考連鎖（Chain-of-Thought）の改善に取り組んだ。記号推論を活用した非反復型の推論手法により推論効率と精度を両立した先行研究 Symbolic-Aided CoT (Nguyen et al. 2025) に対し、推論中のアテンション機構の偏りに着目し、適切な介入によって重要な前提条件の見落としを軽減する手法を提案した [1] (右図)。

② LLMエージェントにおける指示曖昧性の内部表現解析
LLM をエージェントとして活用する際に生じる「指示の曖昧性」問題に対し、モデルの内部表現を解析する手法を提案した。曖昧な指示と明確な指示とでは、モデル内部の表現空間において異なる構造が形成されることを明らかにした [2]。



成果についてより詳細な情報を提供しているWebページ、発表論文などの情報：

1. Phuong Minh Nguyen, Dang Huu-Tien, and Naoya Inoue. Improving Chain-of-Thought for Logical Reasoning via Attention-Aware Intervention. In Findings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 2917-2941, March 2026.
2. 貝出直大, 井之上直也. LLMエージェントにおける指示曖昧性の内部表現解析. 言語処理学会第32回年次大会発表論文集, 4 pages, March 2026.