ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：大規模な深層学習モデルの解釈性向上のための研究開発

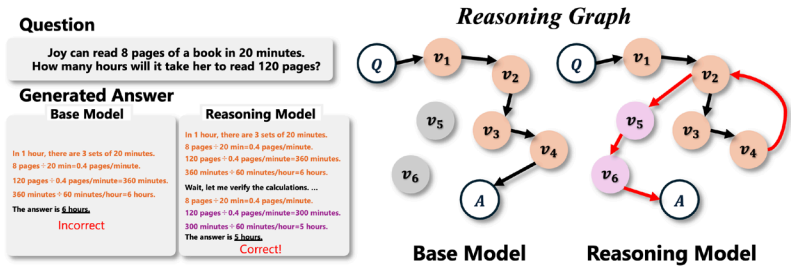
実施時期： 2025/5/1~2025/10/1
所属機関名：東京大学大学院工学系研究科
代表者氏名：峰岸剛基

成果概要：

大規模言語モデル（LLM）の推論能力について，内部挙動からの分析を行なった．また，事後学習を行いデータのクオリティについても検証した．

成果のポイント：

本研究では，大規模推論モデル（Large Reasoning Model；DeepSeek-R1等）の推論能力向上の内部メカニズムを，モデルの隠れ状態から抽出した「推論グラフ（Reasoning Graph）」の構造的性質という観点から明らかにした．



【手法】

推論中の各ステップの隠れ状態をK-meansでクラスタリングしてノードを定義し，ステップ間の遷移を有向エッジとしてグラフを構築．**サイクル数**，**直径**，**スモールワールド指数**の3つのグラフ理論的指標で構造を定量化した（図1）．

【主な発見】

- 推論モデル（DeepSeek-R1-Distill-Qwen-32B）はベースモデル（Qwen2.5-32B）と比較し，**1サンプルあたり平均約5個のサイクル**を持ち，**スモールワールド指数は約6倍**に達する．
- これらの構造的優位性は**タスク難易度（GSM8K → MATH500 → AIME 2024）およびモデルサイズと共に顕著化**し，精度と正の相関を示す．特にAIME 2024ではサイクル検出率がほぼ100%に達した．

【SFTデータ品質との関係】

s1データセットによる教師ありファインチューニングにおいて，**高品質データ（s1-v1.1）ほど推論グラフの直径が拡大し，性能向上と整合**することを確認．

【意義】

ブラックボックス的な大規模推論モデルの性能向上機構を，**解釈可能なグラフ構造**として説明する枠組みを提示．

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：

【発表論文1】 **Gouki Minegishi**, Hiroki Furuta, Takeshi Kojima, Yusuke Iwasawa, Yutaka Matsuo. "Topology of Reasoning: Understanding Large Reasoning Models through Reasoning Graph Properties." *39th Conference on Neural Information Processing Systems (NeurIPS 2025).*
arXiv: https://arxiv.org/abs/2506.05744
GitHub: https://github.com/gouki510/Topology_of_Reasoning
【発表論文2】 Kohsei Matsutani, Shota Takashiro, **Gouki Minegishi**, Takeshi Kojima, Yusuke Iwasawa, Yutaka Matsuo. "RL Squeezes, SFT Expands: A Comparative Study of Reasoning LLMs." *The Fourteenth International Conference on Learning Representations (ICLR 2026).*
arXiv: https://arxiv.org/abs/2509.21128