ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：大規模言語モデル等の透明性・信頼性の確保に向けた研究開発	実施時期：2025年7月～2026年3月 所属機関名：国立情報学研究所 代表者氏名：黒橋　禎夫

成果概要：信頼性の高い大規模言語モデル（LLM）を実現することを目的として、データ構築、モデル構築、事後学習、推論制御、評価、内部解析を一体的に推進した。具体的には、指示チューニングデータの構築・選別、選好学習、BG-MCTSによる推論最適化、MoEモデルの効率的構築を通じて、LLMの能力向上と計算効率の改善に取り組むとともに、BuzzerQAにより事実性評価の基盤を整備し、プロセス報酬モデルにより数学的推論過程の検証・改善を図った。
さらに、記憶機構、LoRA・Transformer・VLMの内部挙動、第二言語・音声学習における人間との類似性と差異を分析することで、LLMの振る舞いに関する理解を深め、より透明性・信頼性の高い次世代基盤モデルの開発に資する知見を得た。

成果のポイント：大規模言語モデル（LLM）の透明性・信頼性・効率性の向上を目的として、学習データ構築、事後学習、推論制御、評価基盤、内部機構解析、モデル構築に関する研究を総合的に推進した。まず、LLMの能力向上を支える基盤として、LMSYS-Chat-1Mからユーザ発話を抽出し、Llama-3.1およびGemma-2により応答を合成することで、指示チューニング用データセットを構築・公開した。さらに、指示チューニングに有効なデータを選別することで、データ量を大幅に削減しても同等の性能を達成できることを示した。また、最小ベイズリスクに基づく生成結果を選好学習に活用し、外部モデルに依存せず単一モデルを改善できる可能性を示した。
限られた計算資源を有効活用する観点からは、推論手法およびモデル構築手法の改善に取り組んだ。推論段階では、所与の予算制約下で解答品質を最大化する探索型デコーディング手法 Budget-Guided MCTS（BG-MCTS）を提案し、残予算に応じて探索幅と深さを適応的に制御することで、高品質な解の生成を可能にした。また、Mixture of Experts（MoE）モデルについて、upcyclingによる学習効率化を検討し、一部パラメータのみを初期化するdrop upcyclingが通常手法を上回る性能を示すことを明らかにした。さらに、エキスパート数の増加が記憶タスクと論理推論タスクに異なる影響を与えることを確認した。
LLMの信頼性を高めるため、出力や推論過程を評価・検証する研究も進めた。人間向けクイズを模した日本語QAベンチマーク「BuzzerQA」を構築し、最新LLMに対して十分な難易度を持つことを確認した。また、数学的推論過程を高効率に検証するため、アンサンブル蒸留と学習ベース集計を組み合わせたプロセス報酬モデル（PRM）を構築し、難関タスクで従来手法を上回る精度を達成した。加えて、人間向け試験問題のデータセット整備やマルチエージェント翻訳システムの評価を通じて、LLMの実用的能力を多面的に測定する基盤を整備した。
さらに、LLMの振る舞いを理解するため、内部機構や人間との類似・相違に関する分析を行った。学習データの記憶機構、4次元幾何問題における人間と類似した正答率低下、LoRAファインチューニング時の固有次元変化、VLMやTransformerの内部挙動を分析し、モデルの出力や学習過程を説明するための知見を得た。また、英語混入を除去した日本語Webコーパスを用いて1.8Bパラメータの日本語特化型LLMを訓練し、日本人英語学習者に近い振る舞いを再現できることを確認した。音声言語モデルについても、自発音声データや第二言語獲得を模した実験により、人間とLLMの学習傾向の差異を明らかにした。
以上の成果は、LLMの能力向上だけでなく、学習効率、推論効率、評価可能性、説明可能性、人間との比較可能性を高めるものである。データ構築と事後学習、推論制御、評価基盤、内部解析、モデル構築を一体的に進めることで、より信頼性の高い次世代基盤モデルの開発に貢献した。

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：
https://llmc.nii.ac.jp/achievements/