ABCI 3.0開発加速利用 （2025年度）成果概要（公開用）

課題名：計算機の文章読解・生成能力の向上、評価に関する研究

実施時期：2025年度
所属機関名：早稲田大学
代表者氏名：河原大輔

成果概要：

日本語大規模言語モデル(LLM)の安全性、信頼性、言語理解、創作能力の向上を目指した研究を実施した。具体的には、マルチターン対話の安全性を評価する「JMT-Safety」、手順型応答の妥当性を問う「ProcedureFC」、車載対話の品質検証基盤「JaCarEval」を構築した。また、複合動詞の内部表現分析によるLLMの意味理解の分析、強化学習による日本語ラップ歌詞の生成を実現した。このように、実社会の多様な要求に応える言語処理技術に関する成果を得た。

成果のポイント：

計算機の文章読解・生成能力の向上、評価に関して、大規模言語モデル(LLM)を核とした以下の研究を実施した。

- **対話の安全性評価**: 日本語マルチターン対話の安全性評価ベンチマーク「JMT-Safety」を構築した。マルチターン対話での有害応答率がシングルターンに比べ最大約3.9倍に達する脆弱性を明らかにした。
- **手順型応答の信頼性向上**: 手順型応答に特化した自動ファクトチェック枠組み「ProcedureFC」(図1)を開発した。手順をステップやフローチャートに分解して検証する手法により、高精度なハルシネーション検出を実現した。
- **車載対話の品質検証**: 日本語車載対話アシスタントを対象とした評価フレームワーク「JaCarEval」を構築した。対話応答の適切性判定に加え、評価器自体の識別性能を定量的に検証するメタ評価基盤を確立した。
- **複合動詞の意味理解分析**: プロービング手法を用い、LLMが日本語複合動詞を内部でどう捉えているかを分析した。LLMが複合動詞を構成動詞の単純合成ではなく一つのまとまり(コンストラクション)として内部表現に保持していることを実証した。
- **強化学習によるラップ歌詞生成**: 強化学習の一手法であるGRPOを利用し、既存歌詞に依存しない日本語ラップの歌詞生成モデルを構築した。独自の報酬関数を設計することで、モデルが自律的に高度な押韻能力を獲得できることを示した。

これらの研究を通じ、日本語LLMの安全性、信頼性、および高度な言語理解と創作能力の実現に貢献した 。

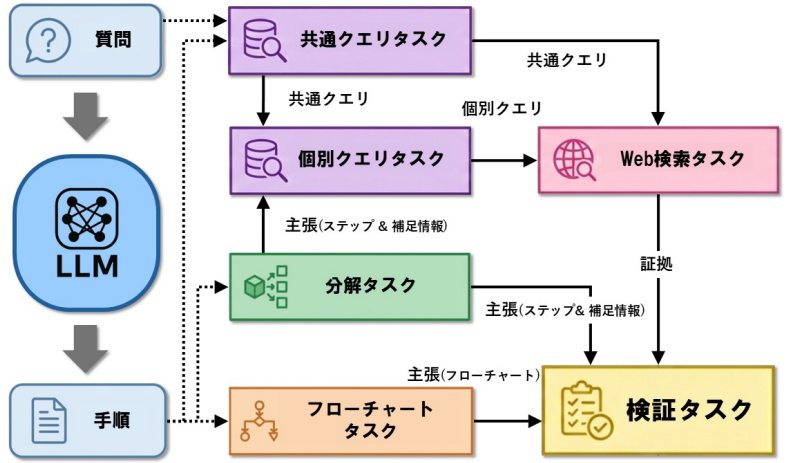

図1: ProcedureFCの概要

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：

- 五十里渚, 福田創, 高山隼矢, 綿岡晃輝, 河原大輔. JMT-Safety: 日本語マルチターン対話における安全性評価ベンチマーク. 言語処理学会第32回年次大会. 2026年3月.
- 杣谷星音, 河原大輔. 大規模言語モデルの手順型応答を対象としたファクトチェックフレームワークの構築. 言語処理学会第32回年次大会. 2026年3月.
- 藤田一颯, 織田宥楽, Sebastian Zwirner, 河原大輔. JaCarEval: 日本語車載対話に対するLLM 評価器のメタ評価フレームワーク. 言語処理学会第32回年次大会. 2026年3月.
- 小野聡, 河原大輔. 大規模言語モデルに対するプロービングによる複合動詞の意味理解の分析. 言語処理学会第32回年次大会. 2026年3月.
- 小川隼斗, 河原大輔. GRPOを用いた日本語ラップの歌詞生成モデルの構築. 言語処理学会第32回年次大会. 2026年3月.