言語モデルにおける日本語ストーリー内の時間的推論に関する研究

北海道大学　工学部　情報エレクトロニクス学科　メディアネットワークコース
言語メディア学研究室　横山明咲

【緒言】

大規模言語モデル（LLM）は自然言語の高次理解を示すものの、物語中の出来事の時間順序推論能力については、その詳細な能力把握が求められている。本研究は、この領域の評価と向上に焦点を当てる。

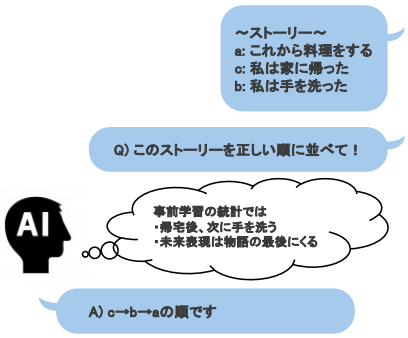
【背景】

~ローカル LLM とクラウド型 LLM~

・クラウド型 LLM：インターネットを介してクラウドサーバ上でモデルを実行する。
例）Chat-GPT・Gemini 等

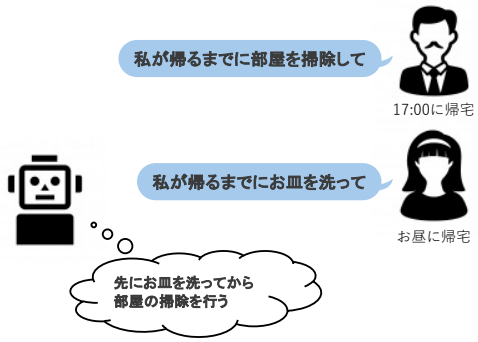
・ローカル LLM：インターネット接続せず、ローカル環境上でモデルを実行する。
例）ELYZA・Gemma 等

→ローカル LLM の方がセキュリティ面、コスト効率は良い反面、一般的にクラウド型 LLM より精度が低い



→事前学習されているデータで実験を行うと提示された文脈や意味的特徴から自律的に順序関係を推論していると言えない。
実際にヒトからされる命令は事前学習データセットに依らない。事前学習データ内の統計的パターンに依存せず、提示された文脈や意味的特徴から自律的に順序関係を推論できるかを確認することは重要である。

ローカル LLM がクラウド型 LLM と同レベルの行動の順序理解が可能になったとき、プライバシー保護、リアルタイム性、コスト効率を重視するロボットなどに実装することが期待できる。



→ロボットが人間の複雑な指示を理解し、行動の順序や因果関係を正確に把握することで、より自律的にタスクを遂行できるようになることが期待される。

【目的】

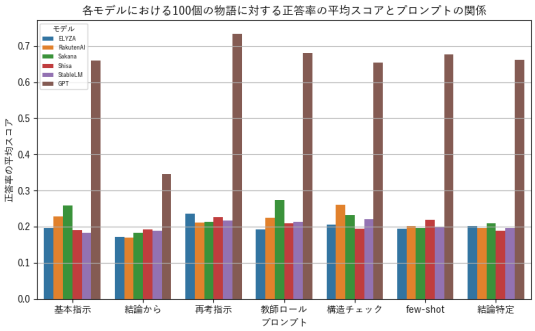
本研究では未学習コーパス DanSto を用いた 5 文の並び替えタスクにより、複数の LLM の統計的共起に依存しない順序推論能力を評価する。

特に、学習規模の小さいローカル LLM に対しプロンプトエンジニアリングを適用することで、その理解精度向上の有効性を検証する。

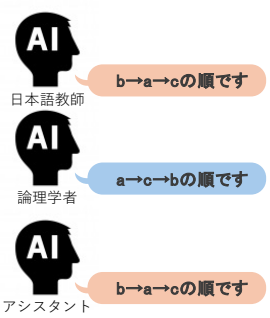
【結果】

様々な工夫を加えた複数のプロンプトに対するローカルモデルの正答率は、全て 0.3 以下となり、 GPT モデルの精度との間に大きな差が生じた。よって、プロンプトエンジニアリングによるサポートで順序関係の理解の向上は困難な可能性が高い。

また、クラウド型モデルである GPT モデルでも正答率の平均のほとんどが 0.7 以下という結果になった。これは DanSto データセットに複数の自然な順序が存在する可能性も示唆している。



【今後の展望】



ユーザープロンプトだけでなくシステムプロンプトの変更による理解度の変化の検証を行い、複数のエージェントの組み合わせによる理解度向上を目指す。

また、正解とは異なっても自然なストーリーとして成立するデータが含まれている可能性を考慮し、複数のクラウド型 LLM を使用したデータセットの精査を行う。

【補足資料】

~使用データセット(DanSto)~

DanSto…5 文からなる日本語の短いストーリーが格納されている。

例）

0) ペンで絵を描いている人が居た

1) 眺めているとその人は突然手を止めた

2) 手元を見るとペンが壊れてしまったらしい

3) その人はバッグから別のペンを取り出した

4) そのペンはインクが切れていたようだ

↓ プログラムによって文の順を複雑に入れ替える ↓

1) 眺めているとその人は突然手を止めた

2) 手元を見るとペンが壊れてしまったらしい

0) ペンで絵を描いている人が居た

3) その人はバッグから別のペンを取り出した

4) そのペンはインクが切れていたようだ

↓

正しい順にLLMに並び替えされる