言語モデルにおける日本語ストーリー内の文順序付け能力に関する評価

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）の社会実装が進む中、特にロボット等の自律システムにおいては、文脈から出来事の順序を把握する「時間的推論」能力が不可欠である。本研究では、計算資源の限られた環境で動作する「ローカル型 LLM」における時間的推論能力の向上と、その適正な評価手法の確立を目的とする 。

【背景】

~クラウド型 LLM とローカル型 LLM~

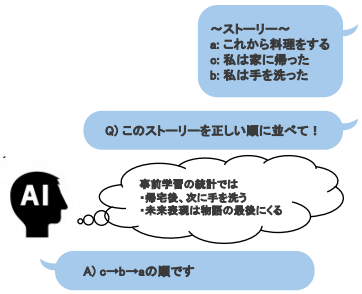
・クラウド型 LLM：インターネットを介してクラウドサーバ上でモデルを実行する。
例）Chat-GPT・Gemini 等

・ローカル型 LLM：通信不要で、ローカル環境上でモデルを実行する。
例）ELYZA・Gemma 等

→ローカル型 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

【課題】

本研究では未学習コーパス DanSto を用いた 5 文の文順序付けタスク(Sentence Ordring）において、既存のローカル型 LLM7B は正答率が約 20%と低迷しており、実用レベルに達していない。



本研究では、以下の 2 段階の手法を提案・実施した。

1. 評価データの品質担保
   使用データに「解釈の曖昧さ」が含まれているとして GPT-4 を含む高性能モデル 3 種の合議制により、正解が一意に定まるデータのみを抽出し、信頼性の高い評価基盤を構築した。
2. 教師ありファインチューニング(SFT)
   Qwen3-8B モデルに対して教師あり学習（SFT）を行った 。

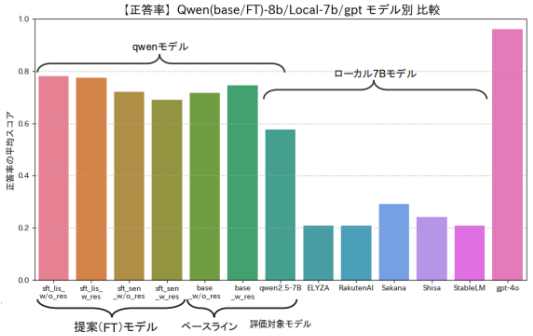
【結果】

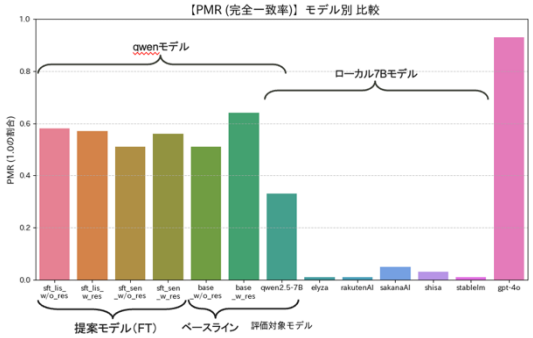
219 件のデータ(提案手法 1 で抽出されたデータ)からランダムに抽出された 100 件をテストデータとして行った結果について

1. SFT モデルは既存の日本語ローカルモデル（平均正答率約 20%）を大きく上回り、最高**正答率 78.2%**を達成した。クラウド型の GPT-4o（96.2%）に迫る実用的な性能となった
2. Qwen3 の機能である推論機能を用いた場合逆に精度が低下した。

→冗長な思考プロセスがかえってノイズとして働いた可能性

本研究により、ローカル環境でも高精度な順序推論が可能であることが示された。





【今後の展望】

・提案手法 1 について…
今回除外した人間でも人間でも判断が割れる物語に対し「難易度の高い（曖昧な）データ」へのモデルがその曖昧性を検知・指摘できるような仕組みの検討

・提案手法 2 について…
SFT に加え、「正解に近い誤答」である負例と「正解データ」である正例を用いた強化学習(DPO)によって「正解に近い誤答」である負例と「正解データ」である正例を用いた学習によって，文順序の微細な差異の識別能力向上を目指す。

【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