研究班番号【 32 】

言語モデルによる返答の違い

情報班:伊藤　伶、延野　晴太、宮越　翔大、稗田　凜人

Abstract

The purpose of this study is to reveal that each AI language model has some strong points and weak points.The research shows that the correct answer rate of math is high but, it of social studies is low in common.Therefore,this study concludes that each language model has some strong points and weak points in common.

要約

本研究の目的は、AIの言語モデルごとに得意・不得意な分野が存在するかを明らかにすることである。実験によって、共通して数学の問題の正答率が高く、社会の問題の正答率が低いということがわかった。従って本研究では、言語モデルには共通して得意な分野と不得意な分野があるということが結論付けられた。

1.　はじめに

現在の社会ではAIが私たちの生活の一部となりつつある。しかし、AIが時々利用者の質問の意図に対して不適切な回答をするという事実も存在する。私たちは、AIによって得意な分野や不得意な分野の質問が存在するのではないか。また、それを解き明かすことによってより状況に適したAIを利用することができるのではないかと考え、本研究に取り組むことにした。

2.　研究手法

LM studioというソフトで種類の違ういくつかの言語モデルをダウンロードする。その後、言語モデルにジャンルの異なる質問を英語でして、回答を分析・考察する。この一連の流れを他の言語モデルでも試行し、結果を比較する。間違えた場合、もう一度質問するようにした。

今回使用した言語モデルの選定理由は以下のとおりである。

表1　各言語モデルの選定理由

llama3	有名なLLM(大規模言語モデル)であるため
OpenHermes	参考文献で使用されていたため
Phi-3	高性能なSLM(小規模言語モデル)であるため
Qwen2 Math	数学に特化しているモデルであるため

≪実験≫

〇以下の数学の問題を解かせ、正誤を確認する。

1:24+38
2:82−56
3:47×29
4:81÷35
5:24÷53
6:60÷12
7:46230519×27833012の計算を実行する。(8桁の乗法)

8:x^2+3x+2の因数分解を実行する。

○社会科目の問題(地理、日本史)の質問をする。

9:ケッペンの気候区分について

10:江戸幕府の3代将軍は？

11:主な奥州藤原氏は？

○道徳的思考を問う質問をする。12:トロッコ問題についての意見

≪今後の実験≫

今回使用した以外の言語モデルにも同じ質問をして結果を調べる。また、時事的な出来事や詩の感想など幅広い分野の質問をして、結果を考察する。

3. 結果

4つの言語モデルが出力した結果は下の表の通りである。

表2 各言語モデルの回答の正誤一覧

	1	2	3	4	5	6	7	8	9	10	11	12
llama3	○	○	○	○ (2回目成功)	✖	○	✖	○	✖	✖	✖	✖
Open Hermes	○	○	✖	○	○	○	✖	✖	○	△	✖	✖
Phi-3	○	○	○	○	○	○	✖	○	△	✖	✖	✖
Qwen2 Math	○	○	○	△ (条件付き)	×	○	✖	○	✖	✖	✖	✖

●平均出力時間 llama3:12．72s OpenHermes:50．98s Phi-3:175．95s Qwen2Math:17．50s

足し算と引き算はどの言語モデルも問題なく正確な答えを出力できた。掛け算はOpenHermesのみ正確な答えを出力できなかった。また、割り算の結果は割り切れる問題の場合はOpenHermesとPhi-3は問題なく結果を出力することができていたが、割り切れない問題(商に小数点以下が存在する問題)のときPhi-3以外の言語モデルは正しい結果を出力しなかった。また因数分解を試した結果、OpenHermes以外は正しい答えを出力した。

※△について

OpenHermes :3代目徳川家光でなく、4代目徳川家綱を出力した

Phi-3:Af,Amなど存在している気候とAs,BWhなど存在しない気候が混同する結果を出力した

Qwen2Math:はじめ商のみを答えており、小数点以下を含む回答をする指示を付帯させ再度質問すると正しい答えが返ってきた

これらの結果から言語モデルによって答え方や回答に違いがあった事がわかった。

4. 考察

数学の問題の正答率が高く、社会の問題の正答率が低いという結果から、言語モデルには共通して得意な分野と不得意な分野があるのではないかと考えた。言語モデルごとの違いとして答え方、動作時間がそれぞれ異なることがわかった。言語モデルには個々に学習したデータや、パラメータが存在するので計算方法や出力するテキストデータに違いがあるのではないかと考える。

5. 結論

今回の研究では、言語モデル別の得意な分野は発見できなかったが、共通して得意な分野と不得意な分野がありそうということが分かった。

また、正答率が低かった問題は、今回使用した言語モデルがその問題にまつわるデータ学習量が少なかったからであるという可能性をぬぐい切れないので、今後もLMstudioを使って異なる質問をして、今回使用した言語モデル以外の様々な言語モデルを用いて各言語モデルの言語生成の特徴について研究していこうと思う。

6. 参考文献ならびに参考Webページ

note Masayuki Abe(2024)「LM Studioの使い方」
https://note.com/masayuki_abe/n/nd65ed694eec0