

物語文に対する大規模言語モデルの読解能力の分析

板橋康知¹ 松林優一郎^{1,2}¹ 東北大学 ² 理化学研究所

itabashi.yasutomo.q7@dc.tohoku.ac.jp y.m@tohoku.ac.jp

概要

GPT-4に代表される大規模言語モデル(LLM)は、知的に高度な人間のための資格試験においても高い能力を示している。本研究では、LLMの心情理解能力を評価する目的で、公立高校入試の国語物語文問題における、LLMの読解能力を調査した。調査対象として、ChatGPT(GPT-3.5-turbo, GPT-4)を調査対象とし、2017年から2023年の物語文読解問題から108問の4択問題をランダムに選び分析した。その結果、GPT-4はランダム選択期待値を大きく上回り、GPT-3.5-turboはわずかに上回った。両モデルは正解した問題の解答理由やテキスト参照箇所も適切であった。分析の結果、GPT-4は特に文章全体の表現の特徴を問う問題に強い一方で、局所的な部分を参照し読み解く問題においては、能力の向上が求められることが分かった。

1 はじめに

大規模言語モデル(LLM)は広範なテキストデータの学習により、複雑な言語表現を理解し、人間の言語運用を模倣する能力を高い水準で獲得している。これらのモデルは、様々な一般的タスクだけでなく、実際に人間が受験するような試験においても人間の水準に匹敵し注目を浴びている[1]。英語に対しては、米国医師免許試験(The USMLE)やアメリカの司法試験であるUniform Bar Exam(4択問題)において合格ラインかそれを大きく上回る水準に到達している[2, 3]。日本語においても、一般的な言語理解能力のベンチマークであるJGLUEにおいて、GPT-4が人間の回答に非常に近い優れた成績を得ることが報告されているほか[4]、日本の医師国家試験において合格水準を上回ったと示されている[5]。

他方、自然言語処理分野では、大規模言語モデルが心情・感情を理解する能力を持っているのかについての研究にも注目が集まっている。他者の心の状態、目的、意図、知識、信念、志向、疑念などを推

測する心の機能を、認知科学の分野で「心の理論(Theory of Mind)」と呼び、この能力は効率的かつ効果的なコミュニケーションのために重要とされている[6]。Choiらは「心の理論」は大規模言語モデルを含む現代の人工知能にとっての課題であると指摘している[7]。大規模言語モデルが人間とのコミュニケーションや人間間の相互理解を助けるような応用の場面では、「心の理論」のような社会的知能を獲得することが必要と言える。

この分野での研究例として、GPT-3は人間の3歳児に、そしてGPT-4は人間の7歳児に匹敵する「心の理論」の能力を持っていると示されている[8]。しかし、大規模言語モデルの「心の理論」に関する研究はほとんどが英語での設定で、かつ、サリーとアン課題[9]のような短文の設定で行われており、長い文脈の中から人間の心情を理解する能力を測った研究はまだほとんど行われていない。

本研究では、公立高校入試国語の物語文という多くの受験生が実際に受ける試験問題を用いて、最新の大規模言語モデルが人間にとって必要とされる心的な読解能力を備えているかを調査することを試みる。公立高校入試では多くの都道府県で出題の方針や各問のねらいが公開されており、国語の物語文については、例えば東京都では「叙述や描写などに即して、語句や文の意味、登場人物の様子、心情などを正しく理解する能力をみる」[10]、岡山県では「登場人物の言動の意味や心情を文章の展開に即して理解したり、理解したことを適切に表現したりする力を見る」としており[11]、他者の心的状態への理解力をはじめとした能力を測る問題だと言える。本研究では物語文を扱うことによって、従来研究より広範かつ複雑な文脈下におけるLLMの心情理解能力について分析する。

モデルによる出力の評価には、解答の正誤のほか、回答根拠の妥当性、受験者の正答率との比較、問いの内容別の性能などの観点から詳細な分析を与えた。結果として、問題全体としてはラン

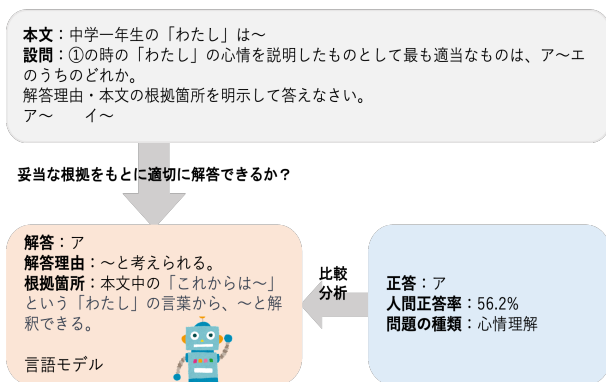


図1 実験の概要図

ダム選択期待値の25%をGPT-4は大きく上回り、GPT-3.5-turboはわずかに上回った。正答率ごと、問いの内容ごとについては、特定のカテゴリーの問題について正解率が高い又は低いといった明確な結果は見られなかった。また、両モデルは正解した問題の解答理由やテキスト参照箇所も適切であり、少なくとも正解した問題に関しては論理的に破綻した根拠は見られなかった。分析の結果、gpt4は特に文章全体の表現の特徴を問う問題に強い一方で、局所的な部分を参照し読み解く問題においては、能力の向上が求められることが分かった。

2 分析対象

データセット 分析の対象データとして、一般に公開されている公立高校入試国語の物語文の問題を利用する。実際の入試においては受験生の学力を測る指標が多岐に渡るため、国語の物語文では人物の心情を問う問題以外の問題も含まれている。本研究では、単に漢字の読み書きや慣用句の意味を問うといった読解を必要としない問題は調査対象から除外した。また、一般にLLMに入力することが困難な図表を元に解答する問題、LLMが指定字数通りに出力することが困難な記述式問題についても除外することとし、テキストのみで問題の入力が可能かつ読解を必要とする記号選択問題を扱うこととした。また、人間の正答率との比較を行う目的で、正答率が公開されている都道府県の問題を対象とした。以上の条件にもとづき、2017年度～2023年度の問題のうち、4択問題となっているものをランダムに108問抽出した。このような問題の具体例を図1の上部に示す。問題は抜粋された物語本文と設問、選択肢からなる。モデルに与えられるタスクは、設問に対し適切な選択肢の一つを選択することである。この際、我々の実験では後の分析のために回答理由や

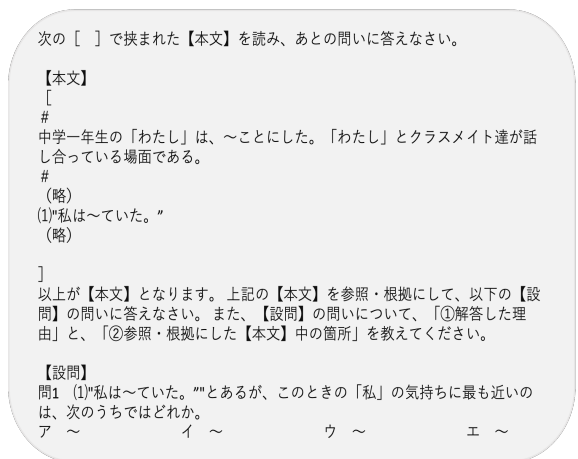


図2 プロンプトの概要。スペースの都合上、実際の問題の一部を(略)とした。

参照した本文箇所も出力させるが、実際の問題にはこのような要件は含まれてはいない。

問題の種類 モデルが出力する解答の正誤について、より詳細に分析するため、分析対象の問題を人間の正答率別、問題類型別に分類した。人間の正答率別については、10%刻みに10の区間に分類した。結果として、0%～20%の区間が0問、30%～と40%～の区間がそれぞれ5問、50%～から80%～の4つの区間がそれぞれ20問、90%～の区間が18問となった。問題の類型化に関しては、読み手が読解の中で生成するとされる推論としてよく用いられる13タイプの推論[12]に基づいて実験者が人手で行った。分類結果は、Class 1: 登場人物の行為意図や動機、目的への推論(例：Aさんが～のような行動を取った目的を答えなさい。)が1問、Class 6: 登場人物の心情に対する推論(例：～の時のAさんの気持ちを答えなさい。)が74問、Class 8: 名詞句が具体的に示すものへの推論(例：ここでの「トロフィー」はどのようなものを例えていますか。)が5問、Class 11: 状況、特性、位置、信念等への推論(例：～から伺えるAさんの刺繍に対する考えを答えなさい。)が12問、Class 13: 筆者の意図や考えへの推論(例：～によって筆者は何を表現しようとしていますか。)が12問であり、分析対象とした物語文の問題は心情理解の問題を69%程度含み、その他、意図や状況を把握させる問題を含む。

3 回答モデル

本研究では、大規模言語モデルに上述の物語文に関する読解問題を解かせ、その解答や解答理由を分析することで、妥当な根拠をもとに解答できている

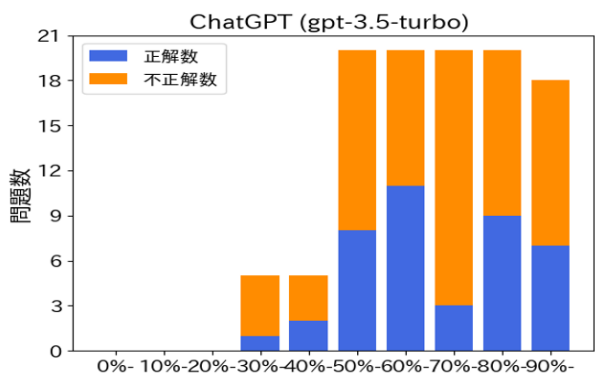


図3 人間の正答率ごとの正解・不正解数 (GPT-3.5-turbo)

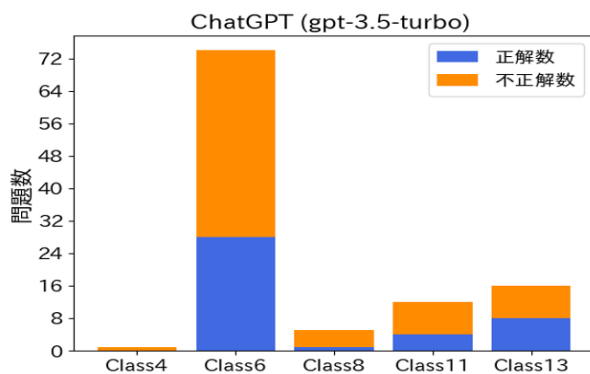


図5 問題類型ごとの正解・不正解数 (GPT-3.5-turbo)

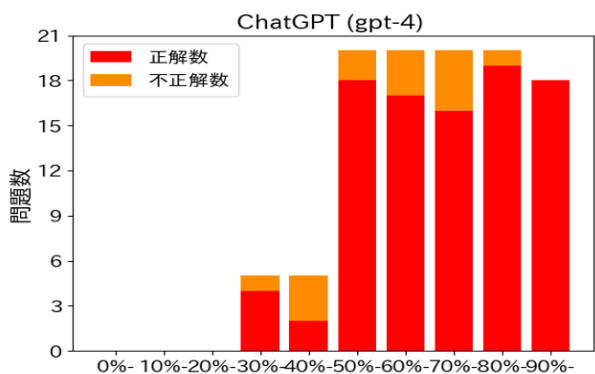


図4 人間の正答率ごとの正解・不正解数 (GPT-4)

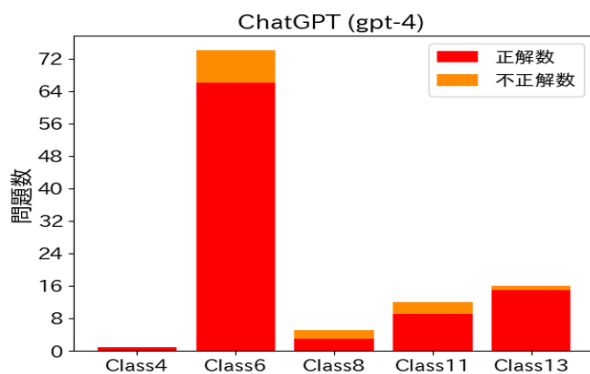


図6 問題類型ごとの正解・不正解数 (GPT-4)

かや、どのような問題に正答又は誤答するかの傾向を調べる。分析対象のモデルには、社会適用が急速に進みつつある GPT-3.5-turbo, GPT-4 を選択した。

プロンプト 抽出された問題を、ChatGPT が読み取り可能な形式に成形し、プロンプトとして入力する。具体的には、実際の入試問題では棒線部で示される部分があるが、このような表示は ChatGPT では読み取り不可能のため、鉤括弧を用いたような表示に変更する。また、ChatGPT の解答が、ランダムに生成された解答ではなく、根拠に基づいた妥当な解答であるか確認するため、解答の理由と参照した本文中の箇所を示させるようなプロンプトを追加した。プロンプトの概要を図 2 示す。

出力 上記プロンプトに対して出力された解答について、選択肢については正答と照らし合わせ、解答の妥当性については解答理由と参照箇所をもとに、許容可能か否かを著者らが人手で評価する。

4 結果と考察

4.1 結果

全体としての正解率は、GPT-4 が 108 問中 94 問正解となり、正解率 87.0 % とランダム選択によ

る期待値 25 % を大きく上回った。GPT-3.5-turbo は 108 問中 41 問正解となり、正解率 37.9 % となり、ランダム選択による期待値をやや上回る程度の結果となった。解答の理由や本文の参照箇所を人手評価した結果、両モデルとも正解した問題については、明らかに誤った理由や参照箇所を出力するケースはなく、すべてのケースで許容範囲内の合理的な理由を持って正答していると判断された。

図 2, 図 3 には、人間の正答率別の正解数を示す。いずれのモデルについても、問題の難易度と正答率の間に意味のある相関は見られず、誤答はどの難易度にも傾向なく分布していた。したがって、現状の LLM に関しては、人間にとっての難易度の高い問題と、LLM にとっての難易度の高い問題は必ずしも一致しないことがうかがえる。

図 4, 図 5 には、問題類型別の正解数を示す。問題類型ごとについても同様に、特定の類型の問題について正解率が高いまたは低いといった明確な傾向はいずれのモデルにも見られなかった。

4.2 考察

より高い成績を示した GPT-4 において、どのような性質の問題を正解できたのか、またどのような問

題を間違える傾向にあるのかを分析した。まず、どの程度の問題まで正解できたのかについてであるが、今回扱った問題の中で最も人間正答率が低かったのは30%–台の問題の計5問であった。GPT-4はそのうちの4問に正しく解答することが可能であった。正しく解答できた具体例を図6に示す。図6における問7は、本文全体を通して文章の表現の特徴を答える問題であり、本文の中に示したような、特徴が現れているいくつかの箇所を適切に参照して解答することができている。このような文章全体の表現の特徴を答える問題については図6の例以外にもGPT-4はすべて正解しており、解答が得意な種類の問題だと考えられる。

次に、どのような問題を間違える傾向にあるのかについてであるが、GPT-4が解答を誤った問題は計14問あり、代表的なものをいくつか挙げる。例えば、図7の問は棒線部より少し前の発言を根拠にして解答すべき問題であるが、gGPT-4が誤って選んだ解答のイに含まれる「他の仕事との兼業までして店を運営していこうと思っている「私」と言うのは、棒線部より後の発言の内容であり、物語の時間軸的にも整合性が合わない。文章全体を通じた理解に得意さを持つことが示唆された一方で、対照的に本来棒線部直前のみを参照すべき問題に関しては、棒線部の後ろまたは更に前まで参照したことで結果解答を誤っており、印象的な結果となった。この結果は、言語モデルが物語の時系列や物事の因果の関係性について、十分な理解を得られていない可能性を示唆する。

5 おわりに

本研究では、物語文における大規模言語モデルの読解能力について分析を行った。結果として、人間正答率別と問題類型別の両方において、特定の問題についての解答能力に明らかな差は見られなかった。一方で、GPT-4の解答に着目すると、文章を広く捉えることに強みをもつ一方で、反対に棒線部直前といった局所的な箇所を参照できず解答を誤る場合があるということが示され、今後のより詳細な分析を方向づける示唆が得られた。今後は、大規模言語モデルの出力に対しての妥当性を評価する手法についてより定性的なアプローチを行うことに加え、大規模言語モデルが問題に応じて柔軟に対応し解答能力が向上するような改善手法の追及について検討したい。

【本文】

(略)

まだ白い実をむしりとっている銀河と同じ年頃の子どもに、恵介は熟した赤い実を渡してやる。銀河と、少し前の銀河が四十三人集まっていると思えば、怒ったりはできなかった。

(略)

ふと恵介は思った。そういえば、銀河にはまだ一度も、俺の採りたて苺を食べさせてないな、と。

(略)

銀河と同じだ。何かを口に入れる時にはまず、口を食べ物と同じ大きさに開く。

(略)

【設問】

問7 この文章の表現の特徴について説明したものととして最も適当なのは、ア～エのうちではどれですか。一つ答えなさい。

ア 恵介と園児たちの短い会話を連続させることによって、園児たちに恐る恐る声をかける恵介のどこか不機嫌な態度を表現している。

イ 園児たちを銀河と重ね合わせて描写することによって、園児たちを息子同様に温かく見守っている恵介の様子を表現している。

ウ 比喩表現を使わないで具体的に説明することによって、園児たちのかわいらしい表情やしぐさをより生き生きと表現している。

エ 物語中の出来事を常に恵介の視点で語ることによって、園児たちに対する恵介の行動や気持ちの変化を客観的に表現している。

【解答】

問7：適当な答えは「イ」です。【本文】では、恵介が園児たちを銀河と重ね合わせる描写が何度もあります。例えば、園児たちが苺を食べるシーンでは、恵介は銀河にも同じ体験をさせたいと考えるなど、彼の息子に対する思いやりが園児たちにも向けられています。これにより、恵介が園児たちを息子のように見守っている様子が表現されています。

(正答：イ 正答率：36.0%)

図7 GPT-4が正解した正答率の低い問題の例。スペースの都合上、実際の問題および解答の一部を(略)とした。

【本文】

(略)

「おばあちゃんが作る雑貨は、それと似てるの。その雑貨を手にした人が勇気を出したり、自信を持ったりできる、スイッチなんだ。『ゆうづつ堂』は、そんなスイッチを人に分けてくれる店なの。私は、お客さんの背中を押す雑貨を作りたい。誰かの支えになりたい。その思いを活かせる場所、『ゆうづつ堂』しか知らないんだ。」

話しているうちに想いをかき出して、止まらなくなる。お母さんは黙って聞いていた。しばらくして、2"お茶を手に取り、ため息を吹きかける"。

「あんたはまた、そんな夢心地な……。楽しいだけじゃダメなもの、店が好きって気持ちだけじゃどうにもならないのも、わかっているんじゃないの?」

「うん。だから、考えたんだ。もっとお店が繁盛するように、来年からは即売会にたくさん出店しようかなって。それと、遠くの人にも知ってもらえるように、ネット販売始めてみようかなとか!」

私が言うと、お母さんもお父さんも驚いた顔をした。

「それでも不安だったら、副業の会社に就職して、店との兼業も考えてる。昼は会社、夕方以降と会社が休みの日は『ゆうづつ堂』の店番をするの。」

(略)

【設問】

2"お茶を手に取り、ため息を吹きかける"における「お母さん」の心情を説明したものととして最も適切なものを、次のア～エから一つ選び、記号で答えなさい。

(略)

イ 他の仕事との兼業までして店を運営していこうと思っている「私」をひそかに心配する気持ち。

ウ 自分の思いを優先して非現実的ども思える希望を述べている「私」に少しあきれている気持ち。

(略)

【解答】

(略)

イ お母さんが「私」の計画を心配していることは本文から読み取れます。特に、「お茶を手に取り、ため息を吹きかける」という行動は、心配や緊張の表れと解釈できます。

ウ 「私」の非現実的な希望に対して、お母さんがあきれているという内容は本文には明確には記述されていません。お母さんの反応は心配や懸念の方向に傾いているように見えます。

(略)

以上の分析から、最も適切な答えは「イ」です。本文中には、「私」が店を守るために様々な計画を立てていること、およびそれに対するお母さんの心配の様子が描かれています。特に「お茶を手に取り、ため息を吹きかける」行動は、お母さんの「私」に対するひそかな心配を表しています。

(正答：ウ 正答率：78.2%)

図8 GPT-4が解答を誤った問題の例。スペースの都合上、実際の問題および解答の一部を(略)とした。

参考文献

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.
- [2] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. **PLoS digital health**, Vol. 2, No. 2, p. e0000198, 2023.
- [3] OpenAI. Gpt-4 technical report. **ArXiv**, Vol. abs/2303.08774, , 2023.
- [4] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.
- [5] Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations, 2023.
- [6] Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms, 2023.
- [7] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020.
- [8] Michal Kosinski. Theory of mind might have spontaneously emerged in large language models, 2023.
- [9] Simon Baron-Cohen, Alan M Leslie, and Uta Frith. Does the autistic child have a “theory of mind” ? **Cognition**, Vol. 21, No. 1, pp. 37–46, 1985.
- [10] 東京都教育委員会. 令和 5 年度東京都立高等学校入学者選抜学力検査結果に関する調査報告書. 東京都教育委員会, 2023.
- [11] 岡山県教育委員会. 一般入学者選抜 [第 I 期] のための学力検査の概要. 岡山県教育委員会, 2023.
- [12] Arthur C Graesser, Murray Singer, and Tom Trabasso. Constructing inferences during narrative text comprehension. **Psychological review**, Vol. 101, No. 3, p. 371, 1994.