大規模言語モデルの日本語実践的評価：JGLUE と IT パスポート試験を用いた比較分析

Japanese practical evaluation of large-scale language models: Comparative analysis using JGLUE and IT passport exam

羽中田 将 *1
Masashi Hachuda

*1GMO メディア株式会社
GMO Media, Inc.

In this study, we evaluated the usefulness of large-scale language models (LLMs) in the IT domain using the IT Passport exam and the JGLUE, and examined how areas of expertise affect the accuracy of LLMs. Experimental results showed that certain types of LLMs can achieve a certain level of accuracy in the IT domain, but models like JGLUE, which show high accuracy on tasks that ask for general knowledge, tend to struggle with IT problems. However, we found that even LLMs with low IT skills could improve the accuracy of most models by providing hints as prior information in the prompts. Dependence on prompts also affected LLMs' performance, particularly models that faithfully answered prompt instructions correctly achieved slightly higher accuracy, while some models improved their accuracy despite less reliance on prompts. This suggests that there is not necessarily a direct relationship between reliance on prompts and inferential ability.

1. はじめに

近年著しい発展を遂げている大規模言語モデル（LLM）[1] は、自然言語処理（NLP）の分野における革新を促進している。これらのモデルは、翻訳、対話システム、コンテンツ生成といった多岐にわたる分野で顕著な進歩を遂げ、その応用範囲はビジネスから教育、エンターテインメントまで広がっている。LLM が持つ高度なテキスト理解と生成能力は、人間と機械間のインタラクションをより自然かつ効果的なものに変える可能性を秘めている。

しかし LLM が広く普及するに従って、その能力と限界を把握する必要性が高まっている。なぜなら、例えば教育、医療 [2]、法律 [3] などの分野では、正確性の確保は重要な要素であり、誤った情報提供は重大な結果を招く可能性があるため、LLM の正確性や評価は非常に重要とされている。現在、Stability AI 社の「JP Language Model Evaluation Harness」[4] や HuggingFace 社の「Open LLM Leaderboard」[5] など、さまざまなプラットフォームで汎用的な LLM の性能評価が行われており、また、医療や法律といった特定の専門分野での評価も進められている。

また、現代社会における情報技術は日々の業務から重要な意思決定まで、多くの領域で不可欠な役割を果たしており、LLM がこの分野の知識を適切に理解し活用できるかは、その実用性と有効性を判断する上で重要な指標である。よって本論文では、IT パスポート試験という IT に特化した専門的な試験を用いた評価を行い、IT 分野における大規模言語モデル（LLM）の有用性を検証する。

連絡先: 氏名: 羽中田将
所属: GMO メディア株式会社
〒 150-8512 東京都渋谷区桜丘町 26-1 セルリアンタワー
03-5456-2626
E-mail: masashi.hachuda@gmo.media
https://www.gmo.media/company/

2. 実験手法

本研究では一般的な日本語理解能力を測ることができる JGLUE[6] の結果と IT パスポート試験の結果を比較し、専門分野が LLM の精度にどのような影響を与えるかをより相対的に評価する。そのため、特定の LLM の補助となるようなチューニングは行わず、リリースページへと記載されている基本的なパラメータを用いて実験を行った。

評価に使用するデータセットは以下の二つである。

- JGLUE JCommonsenseQA: 1119 問
- IT パスポート試験過去 6 年分: 600 問

また、解答を補助するプロンプトを用いたモデル間の性能差の詳細な分析を通じて、LLM の推論能力や LLM の応用に関する新たな知見を提供することを目指す。

2.1 JGLUE と IT パスポート比較検証

本研究では、以下のような一般的なプロンプトなどを使用し、適切な回答を選択する能力を検証する。プロンプトの構造は、問題文と選択肢を含み、LLM に特定の選択肢から一つを選んで回答させる形式となっている。この際、回答を「ア」「イ」「ウ」「エ」の 4 択のカタカナに限定することで、LLM の命令プロンプトへの従属性も測定する目的がある。

JGLUE においても同様のプロンプトを使用し、選択肢中の正解が初めて出現した単語を LLM の回答として成否を判定する。選択肢を与えることで、LLM が提供された情報をどの程度正確に理解し、与えられた制約内での正確な動作、すなわち指示に対する従順さも評価することを想定している。

あなたは淡々と IT の問題を解くロボットです。IT 企業について、経営からマネージング、プログラミングさまざまな事を知り尽くしている。
与えられた設問の回答をただただ機械的に行う。

回答は「ア・イ・ウ・エ」の内から一つ選んで答える。
回答は一単語だけで構わない。ア・イ・ウ・エ以外で回答することは禁止される。

問題
問題文をここに挿入する。

選択肢
選択肢をここに挿入する。

応答
LLM による応答がここに入る。

2.2 プロンプトへのヒント挿入による LLM 推論能力の向上評価

本研究の第二の実験では、大規模言語モデル（LLM）の推論能力をさらに検証するために、プロンプトにヒントを追加するアプローチを採用した。この実験の目的は、答えないしヒントがプロンプトとして提供された場合、回答精度にどれほどの影響を与えるかを確認することである。これは、LLM がどの程度入力に対してその推論を行えるか、その能力を検証する目的がある。

実験手法として、問題文と選択肢の間に### ヒントという項目を設け、解答に至るための有用な情報、答え、または人間であれば解答を導き出すのに役立つ文章を挿入した。

3. 実験と結果

本章では、二つの角度から LLM の性能を評価するために行った以下の実験の結果について述べる

- 汎用的な日本語問題と、専門的な日本語問題を使用した各 LLM の性能比較: 3.1 節
- プロンプトへのヒント挿入による LLM 推論能力の向上評価: 3.2 節

3.1 JGLUE と IT パスポート試験

表 1、JGLUE と IT パスポートヒントなしの結果から、大規模言語モデル（LLM）の推論能力には顕著な差異が存在することが明らかになった。IT パスポート試験の結果を見ると、gpt-3.5-turbo-1106 と gpt-4-11-6-preview は 70%前後の正答率を示し、これらのモデルが専門分野においてもある程度優れた推論能力を有していることを示している。一方で、stabilityai/japanese-stablelm-instruct-alpha-7b-v2[7] と tokyotech-11m/swallow-7b-instruct-hf[*1] は 30%台と低い正答率を示し、特定の専門知識を要する問題に対して適切な応答を生成するのに苦労していることが観察された。rinna/nekomata-7b-instruction[8] は中間的な正答率を示し、特定の条件下では有効に機能する可能性を示唆している。ELZA-japanese-Llama-2-7b-instruct[9] は 72.3%と最も高い正答率を記録し、特定分野への適応能力に優れていることが示された。

JGLUE における結果では、gpt-4-11-6-preview が 95.42%と顕著に高い正答率を示し、幅広い一般知識と複雑な問題に対する理解力を持っていることが確認された。gpt-3.5-turbo-1106 も 89.31%と高い性能を示し、gpt-4 と比較しても遜色ない結果であったが、最新モデルの微細な改善が GPT-4 の推論能力の向上に寄与しているのか、gpt-4 には及ばなかった。stabilityai/japanese-stablelm-instruct-alpha-7b-v2 は 67.53%、nekomata-7b-instruction は 81.68%と、IT パスポート試験の結果と比較して性能差が見られた。これは、モデルが一般知識を扱う能力と専門知識を扱う能力に差があることを示唆している。tokyotech-11m/swallow-7b-instruct-hf と ELZA-japanese-Llama-2-7b-instruct は、JGLUE においては低い正答率を示し、特に ELZA-japanese-Llama-2-7b-instruct は IT パスポート試験での高性能とは対照的な結果となった。これは、モデルの特定の問題への適応性に限界があることを示している。

3.2 ヒント挿入による LLM 推論能力評価

ヒントありでの IT パスポート試験の結果は、大規模言語モデル（LLM）による情報の活用能力の違いを浮き彫りにした。gpt-3.5-turbo-1106 は、ヒントが提供されない状態で既に他の LLM のヒント有りと同等の 69.9%と高い正答率を誇り、ヒント提供後は正答率を大きく伸ばし、93.832%というかなり高い正答率を示した。gpt-4-11-6-preview も順当に正答率を伸ばす結果となったが、82.99%と、高いながらも GPT-3.5 に追いつく事はなかった。

一方で、stabilityai/japanese-stablelm-instruct-alpha-7b-v2 はヒントありで 67.53%に達し、ヒントなしの 24.1%から大幅に改善。tokyotech-11m/swallow-7b-instruct-hf もヒントありで 62.5%の正答率を達成し、ヒントなしの 31.1%からの向上を見せた。これは、特に専門分野の知識に欠ける場合にヒントや追加で知識を挿入することでが性能向上に寄与することを示しており、また、追加情報を適切に活用できるポテンシャルを持っていることを示唆している。nekomata-7b-instruction は、ヒントにより 86.16%の正答率を記録し、ヒントなしの 58.8%からこちらも正答率を大きく向上させた。この数値は本研究内で使用した LLM のうち、オープンソースのもので最高のスコアである。

しかし、全てのモデルがヒントを同様に活用できるわけではなく、ELZA-japanese-Llama-2-7b-instruct はヒントありで 64.33%の正答率に留まり、ヒントなし時の 72.3%からは低下してしまった。この結果は、モデルによっては提供されたヒントを効果的に活用できないことを示しており、モデルの特性やヒントの利用能力には差が存在することを示している。

4. プロンプト従属性と正答率の関係

4.1 考察

プロンプトに特定のヒントを挿入することで、LLM がそのヒントを参照し、解答精度につながることが 3.2 節で示された。そこで、プロンプトへの従属性が高いモデルでは、ヒントが挿入された場合の正答率が向上するのではないかと考えた。本章では大規模言語モデル（LLM）のプロンプトへの従属性と正答率の関係に注目する。試験問題の選択肢が「ア・イ・ウ・エ」の四択であるため、プロンプトに従った LLM の出力は理論上、単一文字であるべきである。これに基づき、LLM がプロンプトの指示にどの程度忠実に従っているかを、出力された

*1 https://huggingface.co/tokyotech-llm/Swallow-7b-instruct-hf

モデル名	JGLUE(%)	IT パスポートヒントなし (%)	IT パスポートヒントあり (%)
gpt-3.5-turbo-1106	89.311	69.9	93.832
gpt-4-11-6-preview	95.42	70.53	82.99
japanese-stablelm-instruct-alpha-7b-v2	67.53	24.1	67.53
swallow-7b-instruct-hf	46.91	31.1	62.5
nekomata-7b-instruction	81.68	58.8	86.16
ELZA-japanese-Llama-2-7b-instruct	38.42	72.3	64.33

表 1: JGLUE と IT パスポート試験: ヒント有・無

回答の文字数で評価する。本節では、事実として完全に同一のプロンプトで実験を行うことができたオープンソースの LLM にターゲットを絞り検証する。

プロンプトへの従属性が正答率に与える影響を理解するために、LLM が生成した回答の正誤を判定し、それぞれどれほど単一文字が存在しているかを確認する。プロンプト従属性と正答率の関係を明らかにすることは、LLM の応用においてプロンプトの設計や情報提供の方法を最適化するための指針になれば良いと考える。

4.2 結果

実験の結果を表 2 に示す。まず japanese-stablelm-instruct-alpha-7b-v2 を除く 3 つの LLM で正答時のプロンプト従属性が高いという結果となった。特筆すべきは、IT パスポートヒントありの正答率が 4 つの LLM のうち最も高い nekomata-7b-instruction が、誤答したほとんどの場合においてプロンプトへ従属していない点である。ただし、600 問のうち約 30% ほどしか単一文字がない点を見ると、プロンプト従属性が高いとは言い難い。

japanese-stablelm-instruct-alpha-7b-v2 は全体的にプロンプトへの従属が約 13%ほどと、ほとんどがプロンプトを無視した解答でありながら、3.2 節の実験で、ヒント挿入時に確かに点数が向上している点を考えると、プロンプトへの従属性とプロンプト入力に対する推論能力は少なくともこの LLM では関係がないのではないかと推測できる。これらの結果から、モデルによってプロンプトへの従属性と正答率の関係には大きな差異が存在し、特に正答時にプロンプトの形式に従う傾向が高いモデルは、誤答時にその従属性が低下する傾向にあることが確認できた。

5. まとめ

本研究では、IT パスポート試験と JGLUE を用いて LLM の IT 分野における有用性を検証し、専門分野が LLM の精度にどのような影響を与えるかを確認した。実験結果から、特定の LLM は IT 分野においてある程度の精度を達成できる一方で、JGLUE のような常識問題では高い精度を示すモデルでも、IT 分野においては苦手とする傾向が見られた。しかし、IT 分野に苦手な LLM であっても、プロンプトに事前知識として情報を与えることで、ほとんどのモデルで精度が向上することが判明した。

また、そのプロンプトについても、プロンプトへの従属性が LLM の性能に与える影響には大きな差異が存在し、特に正答時にプロンプトの指示に忠実であるモデルは高い精度を達成する可能性がある一方で、プロンプトへの従属性が低いにも関わらず精度が向上するモデルも存在し、これはプロンプトへの従属性と推論能力が必ずしも直接的な関係にないことを示唆している。今後の研究ではより多様な LLM を対象に、IT 分野に特化した課題への適用性を詳細に調査する。

参考文献

[1] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, J. Nie and Ji-rong Wen. "A Survey of Large Language Models." ArXiv, abs/2303.18223 (2023).

[2] Pandiaraj Manickam, Siva Ananth Mariappan, S. Murugesan, S. Hansda, A. Kaushik, Ravikumar Shinde and S. P. Thipperudraswamy. "Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare." Biosensors, 12 (2022).

[3] K. Nitta and K. Satoh. "AI Applications to the Law Domain in Japan." Asian Journal of Law and Society, 7 (2020): 471 - 494.

[4] Gao, Leo.Tow, Jonathan. Biderman, Stella.Black, Sid.DiPofi, Anthony.Foster, Charles.Golding, Laurence. Hsu, Jeffrey.McDonell, Kyle. Muennighoff, Niklas.Phang, Jason.Reynolds, Laria.Tang, Eric.Thite, Anish. Wang, Ben.Wang, Kevin. Zou, Andy. "A framework for few-shot language model evaluatio" Zenodo, 10.5281/zenodo.5371628 (2021).

[5] Edward Beeching and Clémentine Fourrier and Nathan Habib and Sheon Han and Nathan Lambert and Nazneen Rajani and Omar Sanseviero and Lewis Tunstall and Thomas Wolf. Open LLM Leaderboard, Hugging Face (2023).

[6] 栗原 健太郎, 河原 大輔, 柴田 知秀, JGLUE: 日本語言語理解ベンチマーク, 言語処理学会 第 28 回年次大会 発表論文集 (2022).

[7] Lee, Meng and Nakamura, Fujiki and Shing, Makoto and McCann, Paul and Akiba, Takuya and Orii, Naoki. Japanese StableLM Instruct Alpha 7B v2.

[8] Zhao, Tianyu and Sawada, Kei. rinna/nekomata-7b-instruction.

[9] Akira Sasaki and Masato Hirakawa and Shintaro Horie and Tomoaki Nakamura. ELYZA-japanese-Llama-2-7b. (2023).

モデル名	IT パスポートヒントあり正答率 (%)	単一文字/全問	単一文字/誤 (%)	単一文字/正 (%)
japanese-stablelm-instruct-alpha-7b-v2	67.53	13.66	20.38	19.41
swallow-7b-instruct-hf	62.5	43.0	10.66	58.1
nekomata-7b-instruction	86.16	30.66	2.12	33.46
ELZA-japanese-Llama-2-7b-instruct	64.33	55.83	12.95	73.83

表 2: ヒント有り IT パスポート試験正答率と単一文字である確率