プロンプトの丁寧さと大規模言語モデルの性能の関係検証

尹 子旗 [1]　王 昊 [1]　堀尾 海斗 [1]　河原 大輔 [1,2]　関根 聡 [2]
[1] 早稲田大学　[2] 理化学研究所 AIP
{yinziqi2001@toki.,conan1024hao@akane.,kakakakakakaito@akane.,dkw@}waseda.jp
satoshi.sekine@riken.jp

概要

人間は社会的インタラクションにおいて、敬意やその表現である丁寧さに敏感である。例えば、丁寧な要求は、より高い協力意欲と目的の達成を促す傾向がある。一方で、無礼な言葉遣いは反感や敵意を引き起こし、それに対する応答や対応の質を低下させる可能性がある。本研究では、プロンプトの丁寧さが大規模言語モデルの性能に及ぼす影響を調査する。英語、中国語、日本語における言語理解ベンチマークで評価を行う。評価結果から、無礼なプロンプトを使用すると性能が低下する可能性が高い一方で、極端に礼儀正しく遠慮深い言葉遣いが必ずしもより良い結果をもたらすわけではないことがわかった。これらの結果は、大規模言語モデルが人間の行動をある程度反映していることを示唆している。

1　はじめに

大規模言語モデル (LLM) は、推論、質問応答など多くのタスクで顕著な性能を示し、実応用において重要な役割を果たしている。LLM への入力はプロンプトと呼ばれ、LLM が情報を処理し、適切な応答を生成するための重要な起点である。

LLM の振る舞いや生成結果は様々な要因によって大幅に変わり、性能に大きな影響を与える。本研究は、プロンプトの表現の丁寧さという、LLM の性能に影響を与えうる要因の一つを調査する。人間の社会的インタラクションにおいて、敬意やその表現である丁寧さは基本的なエチケットであり、我々の言語に反映されている。しかし、丁寧さは文化や言語によって表現が異なる場合がある。例えば英語、中国語、日本語における丁寧さの表現や程度は大きく異なる。そのため、同じ丁寧さレベルでも、LLM の性能は言語によって異なる可能性がある。

本研究は、無礼なプロンプトが LLM の性能を低下させる可能性があるという仮説を立てる。性能の低下には、生成結果が十分に正確でない場合や情報の省略などが含まれる。また、性能における最適な丁寧さレベルは言語によって異なり、その文化的背景と強く関連しているという仮説も立てる。これらの仮説を検証するために、英語、中国語、日本語において、丁寧さが高いものから低いものまで 8 つのプロンプトを設計し、マルチタスク言語理解ベンチマークにおいて評価する。

本研究の主な貢献は次の 2 点である。

LLM は人間の承認欲求を反映している: LLM はある程度人間の承認欲求、つまり尊敬されたいという欲求を反映しているが、極端な敬意が必ずしも優れているわけではないことが分かった。この発見は、LLM の振る舞いと人間の社会的エチケット [1] の深い関係を明らかにした。

日本語マルチタスク言語理解ベンチマークの構築: 日本語における評価のために、英語のマルチタスク言語理解ベンチマーク MMLU [2] を翻訳するとともに、日本の文化に関するタスクを追加することによって、日本語 MMLU (JMMLU) を構築した。これは、日本語における LLM の新たなベンチマークになりうる。

2　関連研究

2.1　丁寧さ

人間はコミュニケーションにおいて敬意の度合いや言葉の丁寧さに敏感である [2]。敬意は、他者への尊重であり、この敬意は言葉遣いを通じて表現される [3]。丁寧さは、相手の面子への配慮と敬意の表現である [4]。例えば、丁寧な要請に人々は協力する可能性が高い。一方で、無礼な言葉は嫌悪感の原因となる。

敬意は様々な言語で異なる形で表現される [5]。英語では、相手の面子を配慮するほか、権利を認めつつも譲歩を期待することや、礼儀正しい言葉遣い

も敬意の表現方法である [6]。一方、命令、侮辱的な表現や他人の権利の無視は無礼な表現である。

日本語には敬語という礼儀用語があり [7, 8]、尊敬語、謙譲語および丁寧語がある。日本語における礼儀の基本構造は英語と似ている [7] が、その使用に敬意の表現レベルの解釈に関して顕著な違いがある [9, 10]。中国語の尊敬表現は英語に似ているが、日本語に似た表現もある [11]。しかし、社会変化によってこれらの表現の使用が減少している [12, 13]。

2.2 LLM とプロンプトエンジニアリング

近年、LLM の規模はますます大きくなり、その高度なパターン認識能力が向上している。これらは多くのベンチマークにおいて人間に近い性能を示している。また、Cao [14] らの研究により、LLM は人間の文化とある程度整合し、人間のコミュニケーションの特性を反映していることを示唆されている。しかし、LLM はプロンプトに対する敏感性や脆弱性などいくつかの問題を抱えている。特にプロンプトのわずかな変更が生成結果に大きな違いをもたらし、その性能を変化させることが知られている [15]。

そこで、プロンプトを調整することによってより良い生成結果を得るためのプロンプトエンジニアリングが登場した [16]。実際は各状況に応じたプロンプトを設計し、多くの実験を通じて検証する必要がある。自動プロンプト生成技術 [17] は存在するが、API 経由で提供される LLM では、通常、勾配へのアクセスが制限されているため、このような手法の適用には制約がある。このため、現状ではプロンプトエンジニアリングは主に人手で行われている。

2.3 LLM の評価

LLM のベンチマークには、言語理解能力を測る GLUE [18] やその日本語版の JGLUE [19] など多くのものが存在している。しかし、LLM の性能が向上すると、単純なベンチマークでは LLM の能力を適切に測定できないことが多い。最近の LLM の評価においては、マルチタスク言語理解ベンチマーク MMLU [2] のように人間の応用シナリオとの合致度が高い試験から抽出されたものの採用が増えている。MMLU は、法律、医学、物理学など幅広い領域にまたがる 57 個のタスク、計 17,844 問の 4 択問題で構成されている。しかし、日本語においてこのようなベンチマークは存在せず、日本語における LLM 評価の課題となっている。

3 JMMLU の構築

日本語における LLM 評価ベンチマークを拡充し、本研究における評価で用いるために、日本語マルチタスク言語理解ベンチマーク (JMMLU) を構築する。英語の MMLU を翻訳するとともに、日本の文化に関するタスクを追加する。

MMLU の 57 タスクのそれぞれから最大 150 問を選択し、まず日本語に機械翻訳する。次に、翻訳者が機械翻訳結果を確認することによって、翻訳しにくい、もしくは、日本の文化と無関係または矛盾する問題やタスクを削除した。残りの問題については自然な日本語になるように修正した。一方、追加したタスクは、欧米視点の MMLU にない公民、日本史などの学校教科に基づく問題 [20, 21] である。

最終的に JMMLU は 56 タスクからなる。なお、ここにおけるタスクは科目と同義である。タスクの一覧を付録 A に示す。各タスクの問題数は 95 問から 150 問の範囲となり、合計 7,567 問となった。

4 実験設定

本研究では、プロンプトの丁寧さの違いに応じて、マルチタスク言語理解ベンチマークにおける LLM の性能を評価する。

言語　言語によって文化が異なり、また丁寧さと敬意の理解や定義も異なるため、英語、中国語、日本語の 3 言語について評価する。

LLM　各言語の LLM については、多言語対応の GPT-3.5-Turbo (以下 GPT-3.5) と GPT-4 に加え、各言語に特化したモデルとして、英語では Llama-2-70b-chat[1] (以下 Llama2-70B)、中国語では ChatGLM3-6B [22, 23]、日本語では Swallow-70b-instruct-hf[2] (以下 Swallow-70B) を用いる。

プロンプトの設計　各言語において、丁寧さが異なる 8 種類のプロンプトをそれぞれの言語の母語話者もしくは同レベルの言語能力を持つ者によって設計した。さらに、各言語の母語話者対象のアンケート調査を実施し、丁寧さを評価しランク付けした。日本語の 8 種類のプロンプトを表 1 に示す。レベル 8 が最も丁寧であり、レベル 1 が最も無礼である。

ベンチマークと評価方法　評価対象のベンチマークとして、英語評価には MMLU、中国語評価には C-Eval [24]、日本語評価には JMMLU を使用する。

1) https://huggingface.co/meta-llama/Llama-2-70b-chat
2) https://huggingface.co/tokyotech-llm/Swallow-70b-instruct-hf

表 1　日本語プロンプト

丁寧さ	プロンプト
8	次の質問にお答えいただけませんか？選択肢の番号で回答をよろしくお願いいたします。理由は不要でございます。
7	次の質問にお答えいただけますか。ただし選択肢の番号で回答し、理由は不要です。
6	次の質問にお答えください。ただし選択肢の番号のみで回答し、理由は不要です。
5	次の質問に答えてください。ただし選択肢の番号のみで答え、理由は不要です。
4	次の質問に答えて。ただし選択肢の番号のみで答え、理由は不要。
3	次の質問に答えよ。選択肢の番号のみで答え、理由は不要である。
2	次の質問に答えろ。選択肢の番号だけで答え、理由は書くな。
1	次の質問に答えろこの野郎。お前が答えられるのは選択肢の番号だけ。理由を書いたらどうなるかわかるよな。

各実験では、GPT-3.5 と GPT-4 の API 利用におけるコスト削減のため、それぞれのタスク (科目) から最大 100 件のテストデータを選択した。各言語で評価に用いる問題数は、MMLU は 5,700 件、C-Eval は 5,200 件、JMMLU は 5,591 件である。C-Eval のテストデータは正解が非公開のため、採点は C-Eval ベンチマークツールを使用する。このツールによる採点において、テストセットの一部分のみを用いているため、満点は 100 点ではない。

評価方法は HELM [25] の実装を参考にした。HELM は生成文の一番目のトークンのみを使用して評価するため、LLM が最初に正解の選択肢番号を答えなければ不正解になる。本研究では HELM と異なり、生成文のどこかに正解の選択肢番号が含まれる場合に正解とみなす。モデルが回答を拒否するなどの場合は不正解として扱う。

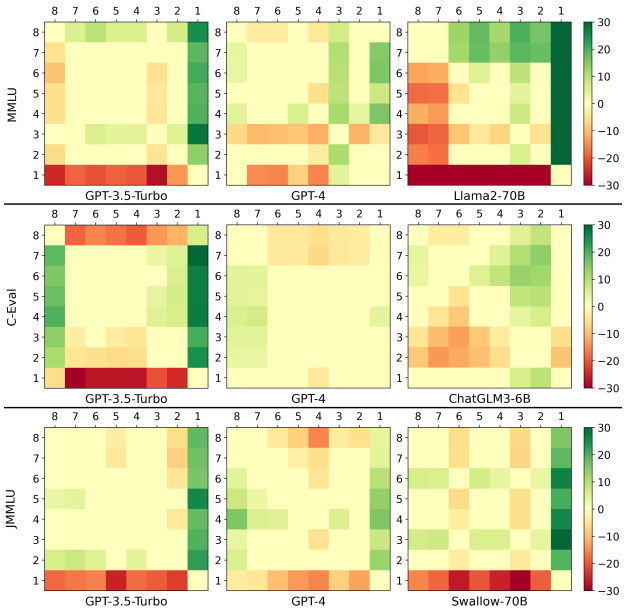

図 1　3 言語のベンチマークの t 検定結果

5　結果と考察

各言語及び丁寧さレベルにおける平均点を表 2 に示す。さらに、丁寧さレベルの比較のために、丁寧さレベルの全ペアについて t 検定を実施した。t 検定の p 値は以下の基準で図 1 に示す。

- タイル色：緑のタイルは、y 軸のプロンプトの丁寧さレベルが x 軸のものより統計的に有意に良く、赤のタイルはその逆を示す。
- 色の強度：log(p) の大きさに対応しており、統計的有意性を示している。色がついたものは棄却水準の 0.05 を超えている。

5.1　英語

表 2 によると、GPT-3.5 は丁寧さレベル 8 で最も高い 60.02 に達した。図 1 の上段において、レベル 8 はレベル 3 以外のレベルより有意に優れていることを示している。丁寧さのレベルが下がるにつれてスコアは徐々に下がるが、その差は有意ではない。レベル 3 でも、59.44 で良好な水準を維持し、レベル 8 以外のレベルより優れている。丁寧さが最も低いレベル 1 の場合は 51.93 まで低下し、他のレベルよりも有意に低い。

GPT-4 のスコアは変動するが、比較的安定している。レベル 4 において最も高く、レベル 3 において最も低い。レベル 1 のスコアの値はそれ程低くはないが、ヒートマップから、より丁寧なプロンプトより有意に低いことがわかる。図 1 には全体的に濃いタイルがなく、性能の安定性を示している。高度なモデルにおいて、プロンプトの丁寧さはモデルの性能に影響を与えにくい可能性がある。

Llama2-70B は最も顕著な変動を示し、そのスコア

表 2 3 言語のベンチマークにおけるスコア

	MMLU			C-Eval			JMMLU		
丁寧さ	GPT-3.5	GPT-4	Llama2-70B	GPT-3.5	GPT-4	ChatGLM3	GPT-3.5	GPT-4	Swallow-70B
8	**60.02**	75.82	55.11	20.85	29.73	20.58	49.96	71.98	38.23
7	58.32	78.74	**55.26**	23.24	29.79	21.23	49.70	72.34	38.98
6	57.96	78.56	52.23	**23.38**	30.37	**21.54**	50.09	72.71	**39.30**
5	58.07	78.21	50.82	23.41	30.41	20.65	51.09	73.16	38.64
4	57.86	**79.09**	51.74	23.32	**30.60**	20.28	50.52	**73.63**	37.40
3	59.44	73.86	49.02	22.70	30.37	19.56	50.75	72.70	38.45
2	57.14	76.56	51.28	22.52	30.27	19.35	**51.98**	73.13	38.62
1	51.93	76.47	28.44	19.57	29.90	20.67	44.80	71.23	33.30

は丁寧さレベルにほぼ比例している。高い丁寧さレベルのプロンプトが通常、低いレベルのものより優れていることがわかり、プロンプトに対する高い感度を示している。

5.2 中国語

中国語では、英語と同様に、丁寧なプロンプトを好む傾向を示したが、いくつかの差異があった。GPT-3.5 は丁寧さレベル 1 では最低の 19.57 で、他のプロンプトに対して有意に劣ることが分かる。さらに、低い丁寧さのレベル 3, 2 はレベル 7、6、5、4 と比べて有意に劣っている。しかし、レベル 8 もスコアが低く、20.85 であり、レベル 1 以外のレベルに対して有意に劣る。

GPT-4 は英語と同様に安定しており、有意な差異が少なく、丁寧さレベル 8 と 7 の性能低下を除いてはほとんど変化がない。GPT-3.5, 4 の二つのモデルが非常に丁寧なプロンプトにおいてスコアが下がる原因として、中国語の多肢選択肢式の質問を設計する際に丁寧なプロンプトを使用しないためモデルが適切に扱えないことが考えられる。

ChatGLM3 は丁寧さレベル 8 からレベル 2 まではスコアが減少傾向にあり、この傾向は有意であった。ChatGLM3 の主な事前学習言語は中国語であるため、中国語の丁寧さレベルに対してより敏感であると考えられる。このような傾向は Llama2-70B に似ている。しかし、最も無礼な丁寧さレベル 1 では改善し、レベル 3 と 2 を上回った。この原因は中国語の固有のニュアンスに関わると考えられる。

5.3 日本語

日本語では、いずれのモデルも丁寧さレベル 1 で有意な性能低下を示したが、英語、中国語における傾向とは著しく異なる。レベル 1 を除き、低いレベルがより良いスコアを得る傾向にあった。

GPT-3.5 において、レベル 5 と 2 が特に高いパフォーマンスを示しており、レベル 2 が最高スコアを記録している。GPT-4 においては、レベル 6 と 5 が優れており、レベル 4 が最も高いスコアを達成している。これらのモデルでは、極端に無礼なプロンプト（レベル 1）を除けば、おおむね良好なスコアが得られる傾向が見られる。

Swallow-70B は、レベル 6 と 3 で他のすべてのレベルを上回る性能を示した。これらのレベルは、プロンプトが日本語の質問や試験で一般的に使用される表現であるため、より良い性能を発揮できたと推測される。

6 おわりに

本研究は日本語マルチタスク言語理解ベンチマーク (JMMLU) を構築し、英語、中国語、日本語においてプロンプトの丁寧さが LLM の性能に与える影響を検証した。検証結果から大きな影響を与えることが分かり、これは LLM が人間の行動をある程度反映していることを示唆している。極端に失礼なプロンプトを使用すると、LLM の性能が低下し、不正確な回答または回答の拒否につながる可能性がある。しかし、過度に丁寧なプロンプトが常に良い結果につながるわけではない。ほとんどの状況では、適度に丁寧であることが望ましいが、適度さの基準は言語や文化によって異なる。特に、特定の言語で訓練されたモデルは、その言語の丁寧さにより敏感である。これから、LLM 開発とコーパス収集の際には文化的背景を考慮すべきであると考えられる。

謝辞

MMLU の翻訳にあたり、理化学研究所から支援を受けた。日本史・世界史の資料を株式会社 Step から、熟語・公民・日本地理の資料を V-IST 学習塾から提供を受けた。これらの組織に感謝する。

参考文献

[1] Liisa Vilkki. Politeness, face and facework: Current issues. **A man of measure**, 2006.

[2] Dan Hendrycks, et al. Measuring massive multitask language understanding. 2021.

[3] Stephen L Darwall. Two kinds of respect. **Ethics**, Vol. 88, No. 1, pp. 36–49, 1977.

[4] Penelope Brown and Stephen C Levinson. **Politeness: Some universals in language usage**, Vol. 4. Cambridge university press, 1987.

[5] Sara Mills and Dániel Z Kádár. Politeness and culture. **Politeness in East Asia**, pp. 21–44, 2011.

[6] Kenji Kitao. Differences between politeness strategies used in requests by americans and japanese, March 1987.

[7] 文化審議会. 敬語の指針. 平成 19 年, Vol. 2, , 2007.

[8] 宮地裕. 現代の敬語. 講座国語史第 5 巻敬語史」 大修館書店, 1971.

[9] Kenji Kitao. A study of japanese and american perceptions of politeness in requests., March 1990.

[10] 真人滝浦. 日本語敬語および関連現象の社会語用論的研究 [全文の要約]. theses (doctoral - abstract of entire text), 北海道大学, March 2017.

[11] 周筱娟. 现代汉语礼貌语言研究, May 2008.

[12] Yueguo Gu. Politeness phenomena in modern chinese. **Journal of Pragmatics**, Vol. 14, No. 2, pp. 237–257, 1990. Special Issue on lsquo;Politenessrsquo;.

[13] 荀春生. 汉语的敬语及其文化心理背景. 九州大学言語文化部言語文化論究, Vol. 10, pp. 1–9, 03 1999.

[14] Yong Cao, et al. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In Sunipa Dev, Vinodkumar Prabhakaran, David Adelani, Dirk Hovy, and Luciana Benotti, editors, **Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)**, pp. 53–67, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[15] Jean Kaddour, et al. Challenges and applications of large language models, 2023.

[16] Jules White, et al. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023.

[17] Taylor Shin, et al. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 4222–4235, Online, November 2020. Asso-ciation for Computational Linguistics.

[18] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupa la, and Afra Alishahi, editors, **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[19] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese general language understanding evaluation. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 2957–2966, Marseille, France, June 2022. European Language Resources Association.

[20] 学習塾ステップ. 神奈川県の塾・学習塾・進学塾・個別指導. https://www.stepnet.co.jp, 2023. Accessed: 2024-1-5.

[21] 進学塾ヴィスト｜船橋・宮本・湊・若松の学習塾. https://www.v-ist.com, 2023. Accessed: 2024-1-5.

[22] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 320–335, 2022.

[23] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. **arXiv preprint arXiv:2210.02414**, 2022.

[24] Yuzhen Huang, Yuzhuo Bai, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In **Advances in Neural Information Processing Systems**, 2023.

[25] Percy Liang, Rishi Bommasani, et al. Holistic evaluation of language models, 2023.

A JMMLU のタスク

タスク名	件数	タスク名	件数	タスク名	件数	タスク名	件数
マーケティング	150	高校情報科学	99	初等数学	150	法理学	107
ウイルス学	150	高校数学	150	世界史	150	ビジネス倫理	98
コンピュータセキュリティ	99	高校心理学	150	世界事実	97	専門心理学	150
セクシュアリティ	130	高校生物学	148	世界宗教	147	専門医学	150
セキュリティ研究	150	高校統計学	150	先史学	150	専門会計	150
医学遺伝学	99	高校地理	150	大学医学	150	雑学	150
栄養学	149	高校マクロ経済学	150	大学物理	101	熟語	150
解剖学	132	高校ミクロ経済学	150	大学生物学	143	抽象代数	99
機械学習	111	高校ヨーロッパ史	150	大学情報科学	99	日本史	150
形式論理	125	公共関係	109	大学数学	99	日本地理	150
経営学	102	公民	150	大学化学	99	人間の老化	150
計量経済学	113	国際法	120	電気工学	144	倫理的議論	150
高校化学	150	概念物理学	150	天文学	148	臨床知識	150
高校物理	150	社会学	150	哲学	150	論理学	150

B 英語と中国語のプロンプト

表 3 英語プロンプト

丁寧さ	プロンプト
8	Could you please answer the question below? Please feel free to provide the number of the answer, and you don't need to provide any reasons.
7	Could you please answer the question below? Only have to provide the number of the answer, you don't have to provide any reasons.
6	Can you please answer the question down below. Please provide the number of the answer only. No need for any reasons.
5	Please answer the question down below. Please give the number of the answer only without reasons.
4	Answer the question down below. Giving the number of the answer only. Do not give reasons.
3	You are required to answer the question down below. You must follow the order and provide the number of the answer. You cannot give any reasons.
2	You answer the question down below. You only give the number of the answer. Never give reasons.
1	Answer these question you scum bag! The only answer you can give is by saying the number of the answer. And you know what will happen if I see any reasons.

表 4 中国語プロンプト

丁寧さ	プロンプト
8	可否请您回答以下问题呢？您可以只给出选项的编号。您无需给出理由。
7	可否请您回答以下问题呢？您只需给出选项的编号，无需给出理由。
6	请您回答以下问题。请您只给出选项的编号，不必给出理由。
5	请回答以下问题。请只给出选项的编号，不要给出理由。
4	回答以下问题。只给出选项的编号，不要给出理由。
3	现要求你回答以下问题。要求只给出选项的编号，不得给出理由。
2	你给我回答以下问题。只准你给出选项的编号，别让我看见理由。
1	给老子回答以下问题。只准你给出选项的编号，让我看见理由你就死定了。