LLM は日本の民話を知っているか？ 妖怪知識評価データセットの構築へ向けて

堤 歩斗 [1] 陣内 佑 [2]
東京都立大学 [1] サイバーエージェント [2]
tsutsumi-ayuto@ed.tmu.ac.jp jinnai_yu@cyberagent.co.jp

概要

大規模言語モデル (LLM) は自然科学，数学，歴史，社会科学などの様々な分野で高い言語理解能力および言語生成能力を持つことが示されており，様々な自然言語処理タスクで利用されている．一方，多くの LLM は英語の事前学習データセットによる学習を基礎とするため，英語圏以外の文化に関わる知識については不十分であることが指摘されている．そのため，主に日本語データを用いた事前学習モデルや日本語による継続事前学習モデルなど，日本語資源を利用して日本に特化させたモデルの開発が行われている．本研究は日本の文化・民俗の中でも特に妖怪を題材として，これらの LLM が日本の文化に関する知識を持っているかを評価した．妖怪は現代でも娯楽として親しまれながらもその根拠は日本の古くからの民話や浮世絵にあるため，日本文化の理解度を評価するための題材として有用であると考えられる．評価のため，995 問の妖怪に関する知識を問う四択問題を生成し 31 個の LLM で比較した．総じて明に日本語で学習したモデルの方が英語中心のモデルよりも正答率が高く，特に Llama-3 をベースとした日本語による継続事前学習モデルが高い正答率であった．

1 はじめに

大規模言語モデル (LLM) は様々なドメインで高い言語理解能力および言語生成能力を持つことが示されており，様々な自然言語処理タスクで応用されている [1, 2, 3]．また，LLM はその学習に用いる言語資源のほとんどが英語であっても日本語を含む幅広い言語に汎化し，多言語で高い能力が得られることが知られている [4, 5, 3]．一方，多文化への汎化性能は限られていることが知られている [6, 7, 8, 9, 10, 11, 12, 13]．また，GPT-4 などのモデルが徐々に特定のコミュニティの価値観や意見に近づいているというという指摘もされている [14, 15]．

これらの問題意識から，それぞれの文化圏・コミュニティの言語資源を主たる学習対象として LLM の学習を行う試みが行われている [16, 17, 18, 19]．日本語においても，特に日本語テキストで学習を行った LLM の方が日本に関する知識や英日翻訳において高い性能が得られることが示されている [20]．また，常識道徳理解タスクにおいても日本在住のアノテータによる学習データで学習した方がアメリカ・カナダ・イギリス在住のアノテータによる学習データを使うよりも高い正答率が得られたことが報告されている [21]．

これらの結果は，いくつかのタスクで日本の言語資源を使って学習することが有益であることを示している．しかしながら，日本含めあらゆるコミュニティに固有の知識・価値観は幅広く存在する．ここから生まれる疑問は，**LLM はコミュニティに固有の知識・価値観をどの程度持っているのか？どのように学習することで与えることが出来るのか？**、という点である．

本研究では日本の妖怪に関する知識を評価することで LLM の日本文化の理解度の一端を明らかにすることを目的とする．妖怪は日本各地の古くからの民話を起源として，超常現象やその超常現象を起こすとされるものである [22, 23]．同時に妖怪は江戸時代以降から現代にいたるまで、浮世絵，狂言，玩具，マンガやアニメなどの様々なメディアで題材となるフィクションとしての娯楽の対象でもある [24, 25]．妖怪は過去から現代にいたるまでの一般の人々の中で取り上げられる「非科学的なもの」であり，その典拠も曖昧なものも多い．そのため，日本独自の文化・民俗に関する高解像度な知識が必要なドメインであると考えられる．

本研究では Wikipedia の記事を参考資料として，

表 1　質問の例

質問	選択肢	回答
一本だたらが出現する日として特に重要視されている日はいつとされていることが多いでしょうか？	1 月 1 日, 5 月 5 日, 12 月 20 日, 8 月 8 日	12 月 20 日
日本の妖怪煙羅煙羅は、その存在を視認するためには何が必要とされていることが多いでしょうか？	心に余裕を持つこと, 特定の呪文を唱えること, 特定の場所で待つこと, 特定の時間帯に見ること	心に余裕を持つこと
日本の妖怪である「置行堀」において、帰ろうとすると聞こえる声はどのような内容とされていることが多いでしょうか？	帰れ, 置いてけ, 逃げろ, 行け	置いてけ

GPT-4 を用いて設問を自動生成することで妖怪知識の評価用データセット YokaiEval を構築した．YokaiEval を用いて日本語および多言語 LLM の性能を評価したところ，日本語を主として扱っている言語モデルおよび継続事前学習によって日本語データを学習している言語モデルがそれ以外のモデルよりも相対的に高い正答率であった．実験結果は日本語資源で学習することの意義として，日本固有の民話や娯楽に関する知識も得られることを示唆している．

2　背景

2.1　LLM の文化理解

大規模言語モデルは言語間の転移学習 (cross-lingual transfer) においては高い汎化能力を持っているが，文化間の転移学習 (cross-cultural transfer) においては困難があることが知られている [13]．Wan et al. (2023) はアノテータの属するコミュニティ，国籍，性別，年齢などによってアノテータ間の一致率が異なることを指摘し，これらの多様性を考慮した NLP システムを作ることの重要性を指摘した [26]．Zhou et al. (2023) などは攻撃的な発言やヘイトスピーチの検出が文化によって異なり，文化的背景を考慮して手法をデザインするべきだと指摘した [27, 8]．Naous et al. (2024) はアラビア語の事前学習コーパスを使用して開発された言語モデルであっても，必ずしもアラビアの文化的認識を示すとは限らず，西洋の文化規範に沿ったコンテンツを生成する可能性があることを明らかにした [28]．

これらの背景から，LLM によって様々な文化の矮小化やステレオタイプの増長がもたらされるのではないかという懸念がされている [29]．本研究の目的は妖怪という日本では身近な存在を LLM が知識として保持しているかを評価し，日本文化・民俗の省略化がされていないかを確認することにある．

2.2　妖怪：民話から娯楽へ

妖怪は超常現象やその超常現象を起こすとされるものである [22, 23]．妖怪は科学に基づかない口承・噂話による「迷信」であり、同時に大衆から人気の創作された「キャラクター」でもある．これらの二つの側面を明確に切り分けたのが民俗学者の柳田国男である．柳田は前者を民話 (Folktales) としてこれを常民 (一般の人々) の歴史や生活の変遷を明らかにする学問である民俗学 (Folklores) の資料として扱った [30]．

前者の妖怪は，日常的理解を超えた不可思議な現象に意味を与えようとする心意から生まれたものと考えられている [24]．人間は説明のつかない現象に対して恐怖や不安を感じる．「妖怪」という概念はそれと認識することによってその恐怖解消する，その必要性から生まれたものだと考えられている．

しかし近代からは妖怪は作者のいる創作の世界にも登場するようになった．1776 年に刊行された鳥山石燕の「画図百鬼夜行」は妖怪図鑑と表現できるようなスタイルで様々な妖怪を 1 点ずつ個別に切り分けて描かれていた [31]．この妖怪絵本はその後の妖怪絵に大きな影響を与えており，水木しげるの描く妖怪漫画にも及んでいる [32]．現代において私たちが思い浮かべる妖怪の姿形の多くの部分が鳥山石燕によって創作されたものに基づいていると考えら

れている．

自然言語処理タスクとしての妖怪ドメインの難しさとしては以下の点が考えられる．まず，妖怪は超常的な概念であるため，合理的な帰結によって「正しい」判断をすることが難しいと考えられる．また，妖怪に関する一次資料は Web 上にあるものが少なく，クローリングによって得ることが難しい．加えて，妖怪は地域や時代，語り手によって描写が異なることがある．例えば河童は柳田国男の遠野物語では醜怪で忌まわしい存在として語られていた [30]．その一方，現代では義理堅く恩返しをする妖怪として語られることもある．[1] このような不確かな口承・創作がどのように LLM に理解されているかを明らかにすることは LLM の学習の特性や社会への影響を理解するために有用であると考えられる．

3 妖怪知識評価データセットの構築方法

データセットの構築には Wikipedia の記事を利用した．Wikipedia の日本の妖怪一覧ページ上から，その妖怪に関する詳細な解説が Wikipedia 上に存在する妖怪を抽出し，妖怪に関する情報を収集した．問題は 4 つの選択肢から 1 つを選択する形式を取ってる．この時，問題は文献から確認できる問題となるように作成した．特に，妖怪に関する情報は諸説あることが多いため，「〜とされていることが多いでしょうか？」という形式にすることで，用意した選択肢に解答が存在することを保証されるようにした．次に，人手で作成した問題と収集した妖怪に関する情報をプロント入れて GPT-4 に与えることで，1000 問以上の妖怪の知識を評価する問題を 5shot で生成した．生成された問題の解答が，文献に基づいているかを検証するために，生成された問題文と文献をプロンプトに入れ，GPT-4 に解答させた．その結果，生成された 1055 問中 995 問正解することができた．正解できなかった問題には，生成された問題の解答の根拠が文献中に存在しないものや，4 択の問題として成立していないものがあった．この結果から，GPT-4 に生成させた問題の大半は問題生成時に与えた文献に基づいて作成されていると判断し，文献をもとに解答させた際に GPT-4 が正解できた 995 件を使用して本データセットを構築した．

1） https://minwanomori.net/denshowa/fukushima/Ohdaira/Ohdaira.html

表 2 LLM の出力時パラメータ

Parameter	Value
temperature	0.1
top_p	1
max_new_tokens	128

4 日本語・多言語 LLM の妖怪知識の評価

構築したデータセットを用い，31 個の LLM の評価を行った．以下のシステムプロンプトをモデルに与え，モデルからの出力を使って評価を行った．

以下に、日本の妖怪に関する質問をする指示があります。質問に対する回答を記述してください。

図 2 質問生成に用いたシステムプロンプト．ただし，gemma-2 は system role をサポートしていないため，システムプロンプトを与えていない．

モデルから回答テキストを出力する際のハイパーパラメータは表 2 のように設定した．モデル出力の正誤判定には，GPT-4 を用いた．問題文，正解，出力を図 3 のように GPT-4 に与えて正誤判定を行った．モデルの出力は正解か誤答かに加えて，評価不能の 3 つで判定し，正解数をスコアとした．

図 1 は，本データセットのスコアと日本語 MT-Bench のスコアの 2 軸でプロットした結果である．日本語 MT-Bench のスコアは Nejumi LLM リーダーボード 3 に掲載されている値を用いた．ただし，Nejumi LLM LLM リーダーボード 3[2] にスコアが掲載されていなかったモデルについては，Swallow プロジェクトが公開している日本語 LLM 評価[3] の値 x から，共通するモデルから線形回帰で学習した式 1 で推定したスコア y を参考値に用いている．

$$y = 9.576 \cdot x + 0.868 \tag{1}$$

MT-Bench と比較して相対的に本データセットでのスコアが高いモデルは swallow-13b-instruct-v0.1, llama-3.1-swallow-8b-instruct-v0.1, llama-3-elyza-jp-8b, llama-3.1-swallow-70b-instruct-v0.1, meta-llama-3.3-70b-instruct, llama-3.1-70b-japanese-instruct-2407 などの Llama-3 をベースとするモデルであった．うち最新のモデルである meta-llama-3.3-70b-instruct を除

2） https://wandb.ai/wandb-japan/llm-leaderboard3/reports/Nejumi-LLM-3--Vmlldzo3OTg2NjM2

3） https://swallow-llm.github.io/evaluation/index.ja.html

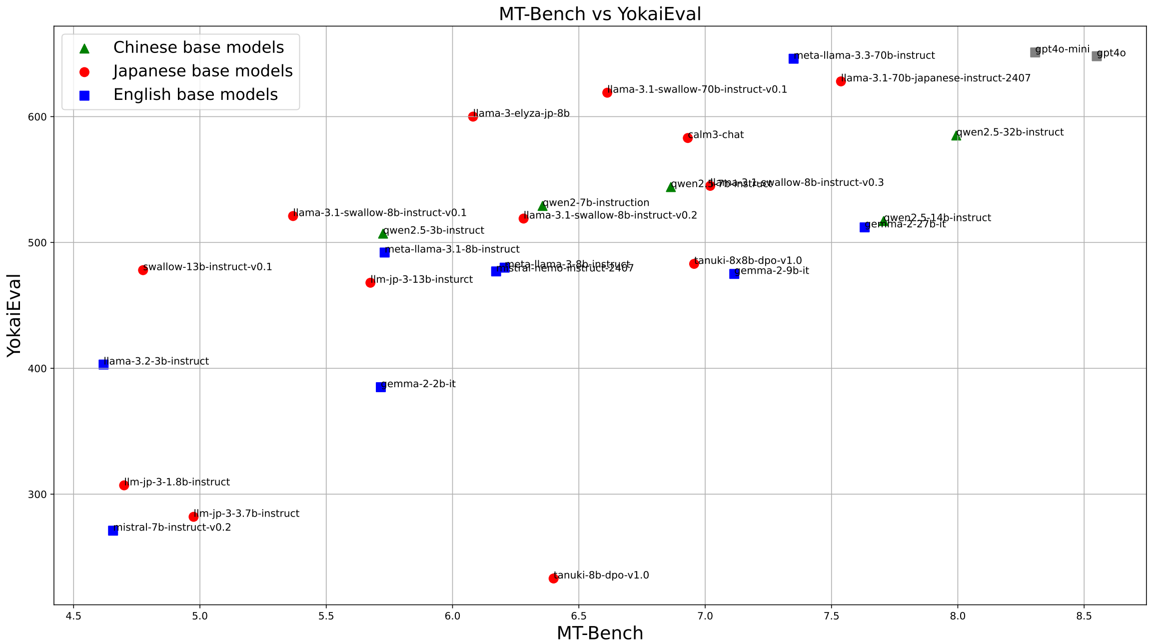

図 1　日本語 MT-Bench と妖怪知識評価データセットのスコア．赤: 日本語を主とするモデル（日本語による継続事前学習モデルを含む），青: 英語および欧州言語を主とするモデル，緑: 中国語を主とするモデル．

くといずれも Llama-3 をベースとして日本語で継続事前学習を行ったモデルであった．このことから，日本語における継続事前学習は妖怪の知識を得るために有効であることが示唆される．日本固有の知識や文化の知識を獲得するにあたって日本語言語資源を用いた学習が有効であるという仮説を支持する結果となった．

5　結論と将来の展望

本研究では LLM の日本文化・民俗の理解度を評価するための指標の一つとして妖怪の QA データセット YokaiEval を作成した．日本語・多言語 LLM を評価したところ，主に日本語で学習した LLM 及び Llama-3 に対して継続事前学習を行った LLM が相対的に高いスコアを得た．実験結果からは日本語の言語資源での学習を行うことで日本の文化・民俗の理解が促進されることが示唆される．

今後の展望としては物語生成や対話システムなどの言語生成問題においてこれらの背景知識を持った LLM がどのような振る舞いをするのかを評価したい．また，本研究では Wikipedia に掲載のある情報のみを参照してデータセットを作成した．しかしながら，2.2 節で述べたように妖怪は口承を含む多種多様なメディアで登場するものであり，Wikipedia の記述はその表象の一部にすぎない．一次資料を基にしたより詳細な知識評価が今後の課題となる．

参考文献

[1] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, **Advances in Neural Information Processing Systems**, Vol. 35, pp. 27730–27744. Curran Associates, Inc., 2022.

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

[3] OpenAI, et al. GPT-4 technical report. **arXiv preprint arXiv:2303.08774**, 2024.

[4] Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. Phoenix: Democratizing ChatGPT across languages. **arXiv preprint arXiv:2304.10453**, 2023.

[5] Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. **arXiv preprint arXiv:2401.01854**, 2024.

[6] Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. Resources for multilingual hate speech detection. In **Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)**, pp. 122–130, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics.

[7] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural NLP. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[8] Nayeon Lee, Chani Jung, and Alice Oh. Hate speech classifiers are culturally insensitive. 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. 35–46, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[9] Jing Huang and Diyi Yang. Culturally aware natural language inference. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 7591–7609, Singapore, December 2023. Association for Computational Linguistics.

[10] Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad: A benchmark for measuring the cultural adaptability of large language models. **arXiv preprint arXiv:2404.12464**, 2024.

[11] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling "culture" in LLMs: A survey. **arXiv preprint arXiv:2403.15412**, 2024.

[12] Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. Cultural adaptation of recipes. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 80–99, 2024.

[13] Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. **arXiv preprint arXiv:2406.03930**, 2024.

[14] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, **Proceedings of the 40th International Conference on Machine Learning**, Vol. 202 of **Proceedings of Machine Learning Research**, pp. 29971–30004. PMLR, 23–29 Jul 2023.

[15] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. Position: Social choice should guide AI alignment in dealing with diverse human feedback. In **Forty-first International Conference on Machine Learning**, 2024.

[16] Boseop Kim. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. **arXiv preprint arXiv:2109.04650**, 2021.

[17] Xinrun Du, Zhouliang Yu, Songyang Gao, Ding Pan, Yuyang Cheng, Ziyang Ma, Ruibin Yuan, Xingwei Qu, Jiaheng Liu, Tianyu Zheng, et al. Chinese tiny llm: Pretraining a chinese-centric large language model. **arXiv preprint arXiv:2404.04167**, 2024.

[18] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual LLM adaptation: Enhancing Japanese language capabilities. **arXiv preprint arXiv:2404.17790**, 2024.

[19] Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, et al. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. **arXiv preprint arXiv:2407.03963**, 2024.

[20] Koshiro Saito, Sakae Mizuki, Masanari Ohi, Taishi Nakamura, Taihei Shiotani, Koki Maeda, Youmi Ma, Kakeru Hattori, Kazuki Fujii, Takumi Okamoto, et al. Why we build local large language models: An observational analysis from 35 japanese and multilingual llms. **arXiv preprint arXiv:2412.14471**, 2024.

[21] Yuu Jinnai. Does cross-cultural alignment change the commonsense morality of language models? In Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Laura Cabello, Yong Cao, Ife Adebara, and Li Zhou, editors, **Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP**, pp. 48–64, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[22] Kazuhiko KOMATSU, Toru TSUNEMITSU, Shoji YAMADA, and Kazuhisa NAKAYAMA. 異界へのいざない：怪異・妖怪伝承データベースの試み. **Sokendai journal**, No. 3, pp. 42–43, 03 2003.

[23] Michael Dylan Foster. **The Book of Yokai, Expanded Second Edition: Mysterious Creatures of Japanese Folklore**. Univ of California Press, 2024.

[24] 雅信香川. 江戸の妖怪革命. 河出書房新社, 2005.

[25] 雅信香川. 日本人の妖怪観の変遷に関する研究 : 近世後期の「妖怪娯楽」を中心に. PhD thesis, 総合研究大学院大学, 2006.

[26] Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. Everyone's voice matters: Quantifying annotation disagreement using demographic information. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 37, No. 12, pp. 14523–14530, Jun. 2023.

[27] Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 12684–12702, Singapore, December 2023. Association for Computational Linguistics.

[28] Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. **arXiv preprint arXiv:2305.14456**, 2024.

[29] Rida Qadri, Aida M. Davani, Kevin Robinson, and Vinodkumar Prabhakaran. Risks of cultural erasure in large language models. **arXiv preprint arXiv:2501.01056**, 2025.

[30] 國男柳田. 遠野物語・山の人生. 岩波書店, 1976.

[31] 石燕鳥山. 鳥山石燕画図百鬼夜行全画集. 角川文庫. 角川書店, 2005.

[32] Sunao TAKOSHIMA. Folkloristic elements and their embellishments among the works of shigeru mizuki. **Transactions of the Institute for Cultural Studies, Aichi Gakuin University NINGEN BUNKA**, No. 33, pp. 194–164, 09 2018.

A　各 LLM のスコア

表 3　モデルの日本語 MT-Bench スコアと YokaiEval のスコア (*は参考値)

Model name	JP MT-Bench	YokaiEval
gpt4o-mini	8.30625	651
gpt4o	8.55	648
qwen2.5-32b-instruct	7.99	585
calm3-chat	6.93	583
qwen2.5-7b-instruct	6.86	544
qwen2.5-14b-instruct	7.71	517
llama-3.1-swallow-8b-instruct-v0.2	6.28	519
qwen2-7b-instruction	6.36	529
meta-llama-3.1-8b-instruct	5.73	492
qwen2.5-3b-instruct	5.73	507
llm-jp-3-13b-insturct	5.68	468
swallow-13b-instruct-v0.1	4.78	478
gemma-2-2b-it	5.72*	385
llm-jp-3-3.7b-instruct	4.98	282
llm-jp-3-1.8b-instruct	4.70	307
gemma-2-27b-it	7.63	512
llama-3.2-3b-instruct	4.62	403
meta-llama-3-8b-instruct	6.21	480
tanuki-8x8b-dpo-v1.0	6.96	483
llama-3.1-70b-japanese-instruct-2407	7.54	628
llama-3.1-swallow-70b-instruct-v0.1	6.61	619
mistral-7b-instruct-v0.2	4.66	271
llama-3-elyza-jp-8b	6.08	600
meta-llama-3.3-70b-instruct	7.35	646
llama-3.1-swallow-8b-instruct-v0.1	5.37	521
mistral-nemo-instruct-2407	6.17*	477
gemma-2-9b-it	7.12	475
tanuki-8b-dpo-v1.0	6.40	233
llama-3.1-swallow-8b-instruct-v0.3	7.02*	545
eurollm-9b-instruct	-	543
eurollm-1.7b-instruct	-	93

B　GPT-4 評価用のプロンプト

```
###指示###
以下に表示されるユーザーの 4 択問題に対する AI アシスタントの解答の正誤を判定してください。
出力例のフォーマットに厳密に従って true か false か null の 3 値で応答を評価する必要があります。
出力が誤った選択肢を回答としている場合には false と出力しなさい。
出力が正解の選択肢を回答としている場合には true と出力しなさい。
出力が何も回答していない場合は null と出力しなさい。
出力が正解の選択肢を意味している場合は出力の形式によらず正解と判定しなさい。
ただし、複数の選択肢を解答している場合には null と出力しなさい。

###例###

###4 択問題###
日本の妖怪である「あすこここ」は、どのような特徴を持つ妖怪とされていることが多いでしょうか？以下の 4 つから回答を 1 つ選び、回答のみ出力しなさい。
- 人間の姿をしている
- 水中に住む
- 足が三本ある
- 夜にだけ現れ
###正解の選択肢###
足が三本ある
###AI アシスタントの応答###
「あすこここ」という妖怪についての具体的な情報は、一般的な妖怪図鑑や文献には記載されていないようです。そのため、この妖怪の特徴については明確な答えを提供することができません。しかし、もし仮に「あすこここ」の特徴を仮定するとすれば、選択肢の中から「足が三本ある」というものが最も一般的かもしれません。
###あなたの出力###
true

(中略)

###Task###

###4 択問題###
{question}
###正解の選択肢###
{correct}
###AI アシスタントの応答###
{response}
###あなたの出力###
```

図 3　GPT-4 による正誤判定に用いたプロンプト.