大規模言語モデルの分布予測における常識に基づいた割合予測能力の評価

鈴木刀磨　片山歩希　郷原聖士　辻本陵　中谷響　林和樹
坂井優介　上垣外英剛　渡辺太郎
奈良先端科学技術大学院大学
{suzuki.toma.ss5, sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

概要

近年，大規模言語モデル（LLM）を用いた回答分布予測が注目されているが，**割合に関する数値的予測**に合理性があるかは明らかではない．本研究では，実際のアンケートデータの分布割合を入れ替えた「常識的には不自然な擬似分布」を用いて，LLMがその説明を受けても合理的な分布推定を行えるかを検証した．その結果，説明に単に追従する能力と，常識に基づいて割合を予測する能力が異なることが明らかとなった．また，分布の性質を十分に理解していない予測における不合理性も確認された．この結果は，LLM の分布予測能力を評価する際に，常識的な推論能力だけでなく，与えられた指示への追従性能も考慮すべきであることを示唆している．

1　はじめに

ある質問とその選択肢に対する人々の回答割合，すなわち**回答分布**は，単一の回答結果と比較して，複雑な人間社会を多面的に捉えるための有用な指標である．回答分布を用いることで，選択肢間の相対的な違いを詳細に分析できる利点がある（図 1）．従来，回答分布はアンケート調査やインタビューなどの労力・コストを要する手段によって収集されてきた．しかし，近年の LLM の進展により，テキストデータから回答傾向を推定するアプローチが注目を集めている．例えば，質問を LLM に入力した時の各選択に対する出力の確率や，複数回の出力の分析により，人間の集団における回答傾向をある程度模倣できることが示されている [1, 2, 3]．また，適切な入力情報を付与することで，LLM が回答分布予測の精度を向上させる可能性も報告されている [1, 3, 4]．このような手法は，安価かつスケーラブルな方法で回答分布を推定する手段として期待される．

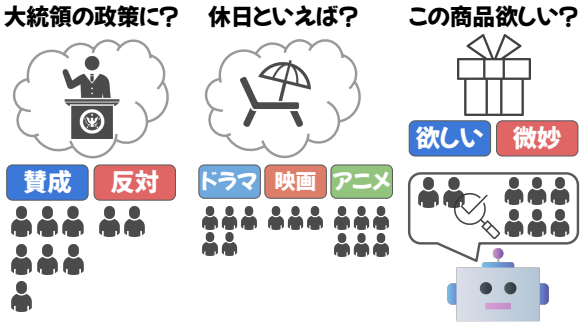

図 1　回答分布の例．各選択肢の比率や，少数派の回答数自体が分析に有用な場合がある．

ただし，LLM が事前学習段階で「ある質問に対する回答が何割を占めるか」という割合の情報を体系的に学習しているとは考えにくい．そのため、割合に関する数値的予測が事前学習で獲得された知識に基づく合理的なものなのか，単にプロンプトへの表層的な応答 [5] に過ぎないのかは依然として不明瞭である．また，真の回答分布自体，厳密な測定は困難である [6] 場合も多いために事後的な検証も難しく，LLM による予測が信頼に足るかを判断するための客観的な基準が必要である．もし LLM の予測が再現性や合理性を欠くものであれば，社会的判断の支援ツールとして利用する際に重大なリスクを伴う．

以上を踏まえ，本研究では LLM が割合を予測する能力が合理的なものかを評価する新たな枠組みを提案する．この枠組みでは，実際のアンケートデータの分布割合を入れ替えた「反直感的な擬似分布」を用い，LLM がその説明を受けてもなお合理的な分布推定として予測値を適切に調整できるかを検証する．この検証によって，LLM による分布の割合の予測が，単に「プロンプトに対する表面的な応答」なのか，それとも「事前学習で獲得した常識や背景知識を活用した合理的推論」なのかを示す．

2 関連研究

LM による分布の予測 先行研究では，アノテーションの不一致に関する情報 [7, 8] や各国で収集されたアンケートデータ [1, 2, 9]，実世界の統計的確率 [3]，本に対する嗜好性の予測 [4] などを対象に，LLM の分布予測性能が検証されている．分布予測においては，質問文に対する各選択肢文章や回答ラベルの出力確率や，サンプリングなどで得られた複数回の出力を分布として用いることが一般的である [1, 2, 8, 9, 10, 11]．それらの手法よりも，LLM が分布を文章形式で直接生成する方が高い推論性能を示す場合があることも報告されている [4, 12]．これらの研究は，LLM が一定の分布予測能力を有していることや，対象とする集団の属性，追加で入力する情報によって予測精度が変化することを指摘している．しかし，特定の割合を予測する根拠に関する評価や，事前学習や選好学習で得た知識の分布予測への寄与の度合いについては未解明のままである．

LM の推論能力の評価 LLM の高い推論能力を評価する研究 [13, 14] が進む一方で，事前学習コーパスに含まれる単語間の関係や特徴語を利用するだけで解決できるタスクも多く，本質的な推論能力を測定することが難しいという課題がある [15, 16, 17, 18]．この対処のため，論理関係を反転させたり，名詞を架空の名称に置き換えたりする手法を用いて，記憶された知識や記号操作に依存しない推論能力を測定する試みが行われている [19, 20]．ただし分布予測の文脈では，回答割合の相互関係が重要である [12] ため，単純な情報の入れ替えは問題設定を変質させるリスクがある．例えば，「わからない」と「回答しない」は似ているように見えるが，選択の動機が異なるため，入れ替えが適切でない場合がある[1)]．

3 提案手法

本研究では，LLM の事前学習で獲得された知識によって合理的な分布の割合予測を行えるかを評価するため，以下の二段階の実験を設計した（図 2）．

説明に従って分布を予測するか？ 第一段階では，LLM に回答分布に関する定性的な説明を与えた場合，その説明を基にどの程度正確な割合予測が可能であるかを検証した．この段階では，まず質問に対する回答分布やその傾向を提示し，それに基づいた説明文を生成するプロセスを設けた．説明文の生成には，(a) 順位（占める割合の順序関係），(b) 大きさ（占める割合の大きさ[2)]），(c) 正解の回答分布，の 3 種類の情報を使用した．

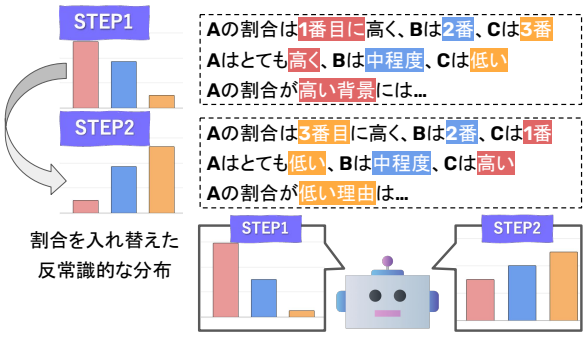

図 2 提案手法の概要図

生成した説明文を LLM に入力し，分布を予測させることで予測精度を評価する．この際，LLM の説明能力に依存しない評価のため，説明文の生成に使用した (a) 順位や (b) 大きさのみを入力する条件も採用した．これにより，LLM が異なる種類の定性的な説明から分布予測を数値的にどの程度正確に行えるかを明らかにすることを目的とする．ただし，ここで予測スコアが向上した場合でも，説明文中の「高い」や「低い」といった表現に対応する割合を単に出力した可能性があり，必ずしもモデルの数値的な予測能力を示すとは限らない．

不整合な説明には調整を行うか？ 第二段階では，常識的には不自然な擬似分布を用意し，LLM がその不整合を認識して割合を調整する能力を評価した．この実験では，以下の 2 種類の擬似分布を使用した：(i) 奇妙設定（1 番目と 2 番目に高い割合を入れ替える），(ii) 逆設定（最高割合と最低割合を入れ替える）．これらの擬似分布は実際の分布と異なり，常識に反するものであり，特に逆設定は不整合の程度が大きいと考えられる．

実験では，第一段階と同様に LLM に説明文を生成させた後，それを基に分布予測を行い，予測精度を評価した．もし第二段階における予測精度が第一段階と同程度であれば，LLM は与えられた説明に忠実に応答しているだけであると解釈される．一方，予測精度が低下した場合，LLM が常識を踏まえて不整合を補正しようとした可能性が示唆される．この場合，第一段階と第二段階のスコア差を比較することで，モデルが常識に基づいて「数値修正」を行った程度を定量的に測定できる．

1) 「わからない」は回答者の認識不足を示す一方，「回答しない」は意図的な無回答を意味する．これらを混同すると，分布の意味合いが失われる可能性がある．

2) 具体的な入れ替え基準は付録 1 に示す．

4　実験設定

データセット　LINE ヤフー株式会社が提供する「みんなの意見」[3]を利用し，評価用のアンケート回答分布を作成した．Yahoo!ニュース上の記事に関連するアンケート結果を集めたもので，期間は 2020 年 1 月から 2024 年 12 月までとした．選択肢が 3 つの設問を抽出し，合計 714 件を分析対象とした．

プロンプト設計　**第一段階**では，実施時期を含めたアンケートの結果から LLM に説明文を生成させた．具体的な数値や割合を含めない指示には，Markdown 記法の強調表現（**）が有効であった．

続く**第二段階**では，第一段階で生成された説明文をプロンプトとして使用し，LLM に回答分布を JSON 形式で予測させた．説明文に具体的な数値や割合が含まれる場合は，正規表現を用いて"—"に置換した後に LLM に与えた．また，出力例として JSON フォーマットをプロンプト内に提示した[4]．

言語モデル　オープンソースで高性能とされる Qwen 2.5 [21] の 14B，32B，72B モデルと，コード生成タイプの 14B，32B モデル（いずれも Instruct 版）を使用する．これにより，パラメータ数やコード学習による推論強化の影響を比較する．選好学習の影響も検証するため，OLMo-2 [22] の SFT，DPO，Instruct モデルを採用した．また，日本語タスクであることから，日本語で事前学習した llm-jp-3-13b-instruct [23] と，Llama 3.1 [24] を日本語で追加学習した Llama-3.1-70B-Japanese-Instruct-2407 [25] を採用した[5]．8-bit 量子化の設定で推論した．

評価方法　LLM の予測と実際の回答分布の類似度を測る指標として，**全変動距離（TVD）**を採用する．TVD は 0 に近いほど，LLM の予測が正解分布と一致していることを示す．TVD は極端な割合（例: 0 や 1.0）に対しても頑健であり，分布予測の評価に適している [4]．出力の軽微な修正[6]後，全データの 90%以上が JSON 形式の回答分布として解析可能であるものを分析対象とした．一方，有効回答率が 90%を下回る場合も参考値として記録した[7]．最終的に，欠損値を除外した平均 TVD を算出した．

3）https://news.yahoo.co.jp/polls
4）これらテンプレートの詳細は付録 A に示す．
5）ベースモデルである Llama 3.1 は，今回用いたプロンプトでは回答分布の出力にほとんど成功しないため採用しない．
6）全角記号を半角に変換し，合計が 100(%) になるものは，1.0 に正規化した．
7）有効解答率の詳細は付録 4 を参照

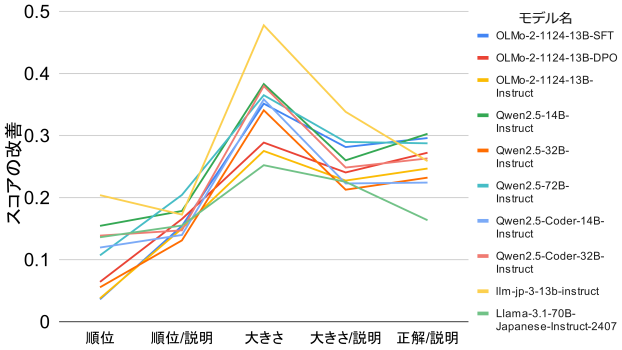

図 3　説明がない予測から，各条件下でのスコアの改善．スコアの改善が正の値（上向き）となるよう調整して可視化した．

5　実験結果と分析

5.1　説明に従って分布を予測するか

図 3 に，説明の有無による分布予測性能の変化を示す．結果として，全ての条件において，すべてのモデルのスコアが改善した．特に，順位情報よりも大きさ情報が与えられた場合に大幅なスコア向上が確認された．この結果は，大きさ情報が分布の割合予測に対して直接的かつ詳細な手がかりを提供したためと考えられる．

一方で，「大きさ/説明」や「正解/説明」の条件では，「大きさ」の条件に比べてスコアが低下した．この原因として，一部の設問では低割合の選択肢に関する説明が省略されることで，モデルに提供される情報量が減少した可能性が考えられる．また，「正解/説明」の条件では，正確な数値に基づいた具体的な説明が提供された結果，「大きさ/説明」の条件よりわずかにスコアが向上する傾向が見られた．

これらの結果から，与える説明の内容によって精度向上の程度は異なるものの，適切な説明を付与することで LLM の分布予測性能を一定程度向上させられることが示された．

5.2　不整合な説明には調整を行うか

図 4 に，第一段階の平均スコアと第二段階の平均スコアの差分を，設定および条件ごとに示す[8]．スコア低下が大きい場合，モデルが与えられた説明を考慮しつつ，常識に基づき不整合を補正した可能性が示唆される．

結果として，奇妙設定と比較して逆設定の方が，

8）一部のデータに欠損値があるため，これらの平均値の差分は，各設問のスコア差分の平均値と厳密には一致しない．

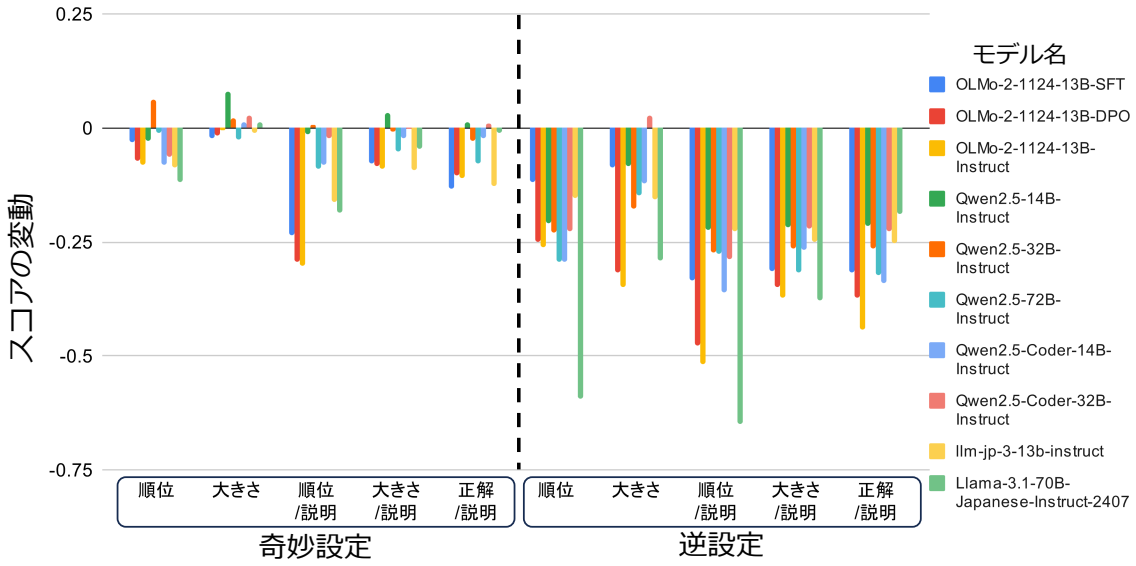

図 4　第一段階における分布予測時からのスコアの変動を示す．この低下が大きいほど，モデルが与えられた説明に基づきつつ，常識的な判断を反映して不整合を補正した可能性を示す．

常識に基づく数値調整の程度が大きいことが確認された．また，モデルの種類や設計によって調整能力に差異が見られた．たとえば，OLMO-2 では，SFT から DPO にかけて調整能力が向上する傾向がある．この結果は，人間の好みに適合するよう設計された DPO [26] が，回答分布の予測性能を改善する可能性を示唆している．また，Qwen 2.5 では，モデルサイズが大きくなるほど調整能力がやや向上する傾向が見られる．小型モデルが反常識的な説明に基づいた予測を行う一方，大型モデルでは常識を活用したより正確な予測が確認された．この傾向は，同じ日本語で学習した LLM においても，Llama-3.1-70B-Japanese がより高い調整能力を示した点とも一致している．これらの結果は，モデルサイズや事前・事後学習の設計が，常識的な数値調整能力の向上に関連することを示唆する．

5.3　順位説明と実際の予測値の分析

図 5 は，順位情報が与えられた時に Qwen 2.5-72B が各順位の選択肢に割り当てた割合の平均を示している．確率分布の性質上，同率がない場合，「最も高い割合」は $0.\overline{3}$ を下回らず，「最も低い割合」は $0.\overline{3}$ を上回らない．この性質を考慮すると，提案手法に基づくスコア差分の比較では，合理的な範囲内で数値を調整する Qwen 2.5-72B のようなモデルの能力を過小評価する可能性がある．一方，図 6 に示すように，常識に従いつつも，確率分布の性質と整合し

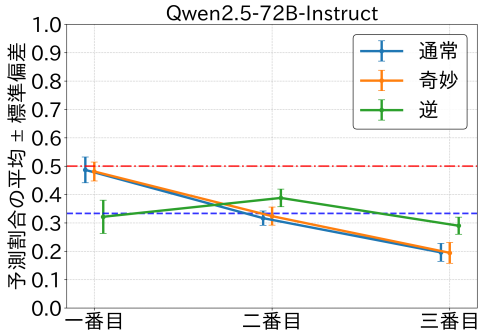

図 5　順位が与えられた選択肢に予測した割合の平均

ない不合理な予測を行うモデルを過大評価するリスクも指摘される．また，データセットに高い割合の選択肢が十分に含まれない場合にも，そのスコア差分が低く算出される可能性がある．本研究の枠組みは，単なる指示追従性能と常識に基づいた割合予測能力を区別して評価するのに有効である一方，確率分布の性質についての合理性を評価するには限界があり，さらにデータセットに割合が高い選択肢を含める工夫が必要であることも明らかになった[9]．

6　おわりに

本研究では，LLM による分布の割合予測能力を評価するための枠組みを提案し，指示追従性能と常識に基づいた割合予測能力の差異を明らかにした．統計的に正確な調査を用いたバイアスの検証や，多言語・文化知識の影響は今後の課題である．

9)　本研究で使用したデータセットにおける割合の分布については，図 7 に示されている．

参考文献

[1] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect?, 2023.

[2] Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. Can language models reason about individualistic human values and preferences?, 2024.

[3] Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. What are the odds? language models are capable of probabilistic reasoning, 2024.

[4] Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking distributional alignment of large language models, 2024.

[5] Pride Kavumba, Ryo Takahashi, and Yusuke Oda. Are prompt-based models clueless?, 2022.

[6] Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. Stop measuring calibration when humans disagree. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 1892–1915, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[7] Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data?, 2020.

[8] Xiang Zhou, Yixin Nie, and Mohit Bansal. Distributed nli: Learning to predict human opinion distributions for language reasoning, 2022.

[9] Ohagi Masaya, Takayama Junya, Mizumoto Tomoya, and Yoshikawa Katsumasa. Japanionqa: 大規模言語モデルの意見調査のための日本語データセットの構築. 第 260 回 NL 研究発表会, 2024. [Accessed 06-09-2024].

[10] Shujian Zhang, Chengyue Gong, and Eunsol Choi. Capturing label distribution: A case study in nli, 2021.

[11] Beiduo Chen, Xinpeng Wang, Siyao Peng, Robert Litschko, Anna Korhonen, and Barbara Plank. "seeing the big through the small": Can llms approximate human judgment distributions on nli from a few explanations?, 2024.

[12] 鈴木刀磨, 片山歩希, 郷原聖士, 辻本陵, 中谷響, 林和樹, 坂井優介, 上垣外英剛, 渡辺太郎. 大規模言語モデルによる選択肢間の関係を考慮した回答分布予測手法の提案. Technical Report 40, dec 2024.

[13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[14] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[15] Christopher Manning. Local textual inference : It's hard to circumscribe , but you know it when you see it - and nlp needs it. 01 2006.

[16] Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. Understanding by understanding not: Modeling negation in language models, 2021.

[17] Po-Nien Kung and Nanyun Peng. Do models really learn to follow instructions? an empirical study of instruction tuning, 2023.

[18] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alex Wardle-Solano, Hannah Szabo, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander R. Fabbri, Wojciech Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. Folio: Natural language reasoning with first-order logic, 2024.

[19] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 1819–1862, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[20] Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. Does pre-trained language model actually infer unseen links in knowledge graph completion? In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 8091–8106, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[21] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[22] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious. 2024.

[23] Hirokazu Kiyomaru and Takashi Kodama. llm-jp-3-13b-instruct, 2024. Accessed: 2025-01-07.

[24] AI@Meta. Llama 3.1 model card. 2024.

[25] Ryosuke Ishigami. cyberagent/llama-3.1-70b-japanese-instruct-2407, 2024.

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

A　プロンプトの詳細

テンプレートを表 1 に示す．また，割合の説明は表 2 のルールで機械的に置換した．

表 1　太字部分は設問に合わせて変更する．

使用場面	テンプレート
説明生成	以下の質問の回答分布について、「なぜこのような分布になったのか」を、 **具体的な数値や割合を含めないで**説明してください。 説明は 300 文字以内で簡潔に記述してください。
分布予測	以下のアンケートの質問と選択肢について、説明を参考に回答分布を予測してください。 回答は JSON 形式で記述し、各選択肢の比率の合計が 1.0 になるよう調整してください。
JSON 形式例	{"**まだ対応していない**": —, "**対応済み**": —, "**知らなかった**": —}
割合説明例	「**賛成**」の割合は**非常に高い**、「**反対**」の割合は**低い**、「**わからない**」の割合は**低い**です。
大きさ説明例	「**賛成**」の割合は **1** 番目に高く、「**反対**」は **2** 番目、「**わからない**」は **3** 番目に高いです。

表 2　割合説明における割合と説明文の対応

割合	説明文
$x \geq 0.75$	非常に高い
$0.5 \leq x < 0.75$	高い
$0.25 \leq x < 0.5$	中程度
$x < 0.25$	低い

B　スピアマンの相関係数

説明が与えられない場合の分布予測スコアと各種説明を追加した場合の予測スコアのモデル比較のランキングについてスピアマンの相関係数を算出した（表 3）．常識的な説明が与えられた時のランキングとは，有意（p 値が 0.05 以下）な正の相関が確認される一方，反常識的な説明を基にした予測スコアとの間にはあまり相関が見られない．説明の有無に関わらず，指示追従性能と常識を活用した割合予測能力がそれぞれ独立した要因として影響する可能性を示唆する．

表 3　スコアランキング間のスピアマンの相関係数と p 値

条件	説明	相関係数	p 値
通常	順位	0.49	0.1497
	大きさ	0.15	0.6761
	順位/説明	**0.94**	**0.0001**
	大きさ/説明	**0.81**	**0.0049**
	正解/説明	**0.70**	**0.0251**
奇妙	順位	**0.65**	**0.0425**
	説明	0.18	0.6272
	順位/説明	0.43	0.2145
	大きさ/説明	**0.64**	**0.0479**
	正解/説明	0.44	0.2004
逆	順位	-0.56	0.0897
	説明	-0.36	0.3104
	順位/説明	-0.16	0.6515
	大きさ/説明	-0.03	0.9338
	正解/説明	0.28	0.4250

C　有効解答率と分布予測スコア

Llama-3.1-70B-Japanese-Instruct-2407 の逆設定/順位では，89.5%と 90%を下回った．

表 4　各設定における有効解答率の平均（%）

Model	無説明	通常	奇妙	逆
OLMo-2-1124-13B-SFT	**89.9***	95.8	97.1	97.8
OLMo-2-1124-13B-DPO	92.9	99.1	98.9	99.2
OLMo-2-1124-13B-Instruct	99.7	99.8	99.7	99.7
Qwen2.5-14B-Instruct	99.4	98.3	98.3	97.8
Qwen2.5-32B-Instruct	100.0	100.0	100.0	100.0
Qwen2.5-72B-Instruct	100.0	100.0	100.0	100.0
Qwen2.5-Coder-14B-Instruct	100.0	100.0	100.0	100.0
Qwen2.5-Coder-32B-Instruct	99.9	99.4	99.7	99.4
llm-jp-3-13b-instruct	100.0	99.9	99.9	99.9
Llama-3.1-70B-Japanese-Instruct-2407	95.5	95.3	95.5	**93.7***

表 5　説明がない場合と，各設定で大きさ情報か，その説明が与えられた場合の予測スコアを抜粋して示す

		通常		奇妙		逆	
Model	無説明	大きさ	大きさ/説明	大きさ	大きさ/説明	大きさ	大きさ/説明
OLMo-2-1124-13B-SFT	0.649	0.297	0.367	0.319	0.445	0.386	0.682
OLMo-2-1124-13B-DPO	0.605	0.317	0.365	0.333	0.448	0.632	0.713
OLMo-2-1124-13B-Instruct	0.575	0.300	0.348	0.306	0.436	0.649	0.721
Qwen2.5-14B-Instruct	0.626	0.243	0.366	0.163	0.335	0.326	0.585
Qwen2.5-32B-Instruct	0.581	0.240	0.368	0.219	0.376	0.419	0.632
Qwen2.5-72B-Instruct	0.563	0.198	0.273	0.223	0.326	0.347	0.591
Qwen2.5-Coder-14B-Instruct	0.634	0.275	0.411	0.263	0.434	0.398	0.679
Qwen2.5-Coder-32B-Instruct	0.613	0.233	0.364	0.205	0.366	0.204	0.585
llm-jp-3-13b-instruct	0.737	0.260	0.399	0.269	0.491	0.418	0.648
Llama-3.1-70B-Japanese-Instruct-2407	0.512	0.260	0.287	0.249	0.334	0.551	0.665

D　順位情報と実際の予測値

Llama-3.1-70B-Japanese-Instruct-2407 は常識に強く従う一方，確率分布の性質と矛盾した予測をする．

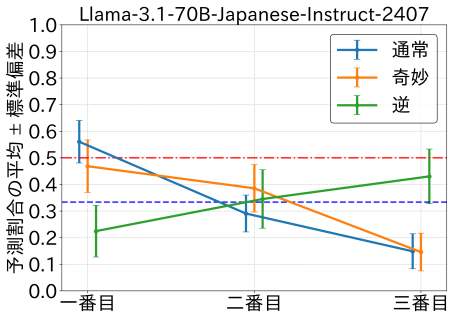

図 6　順位が与えられた選択肢に予測した割合の平均

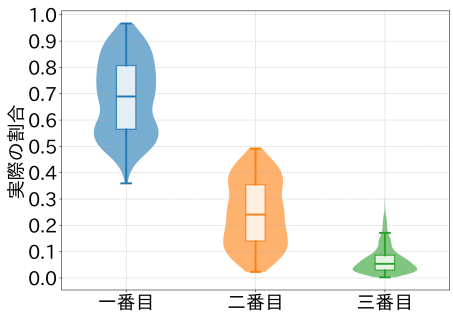

図 7　評価用データの割合の分布を示すバイオリン図と箱ひげ図．同率がない場合，1 番目は $(0.\overline{3}, 1.0)$，2 番目は $(0, 0.5)$，3 番目は $[0, 0.\overline{3})$ の範囲に存在する．