Reliability of Distribution Predictions by LLMs: Insights from Counterintuitive Pseudo-Distributions

Toma Suzuki Ayuki Katayama Seiji Gobara
Ryo Tsujimoto Hibiki Nakatani Kazuki Hayashi
Yusuke Sakai Hidetaka Kamigaito Taro Watanabe
Nara Institute of Science and Technology (NAIST)
{suzuki.toma.ss5, katayama.ayuki.kc1, gobara.seiji.gt6
tsujimoto.ryo.tq0, nakatani.hibiki.ni4, hayashi.kazuki.hlj4,
sakai.yusuke.sr9, kamigaito.h, taro}@is.naist.jp

Abstract

The proportion of responses to a question and its options, known as the response distribution, enables detailed analysis of human society. Recent studies highlight the use of Large Language Models (LLMs) for predicting response distributions as a cost-effective survey method. However, the reliability of these predictions remains unclear. LLMs often generate answers by blindly following instructions rather than applying rational reasoning based on pretraining-acquired knowledge. This study investigates whether LLMs can rationally estimate distributions when presented with explanations of "artificially generated distributions" that are against commonsense. Specifically, we assess whether LLMs recognize counterintuitive explanations and adjust their predictions or simply follow these inconsistent explanations. Results indicate that smaller or less human-optimized LLMs tend to follow explanations uncritically, while larger or more optimized models are better at resisting counterintuitive explanations by leveraging their pretraining-acquired knowledge. These findings shed light on factors influencing distribution prediction performance in LLMs and are crucial for developing reliable distribution predictions using language models.

1 Introduction

The proportion of responses to a question and its options, known as the response distribution, provides valuable insights into human society beyond individual responses. Response distributions allow detailed analysis of relative differences between options (see Figure 1). Traditionally, they have been collected through labor-intensive and costly methods like surveys and interviews. Recent advances in Large Language Models (LLMs), however, offer new approaches for estimating response tendencies from textual data.

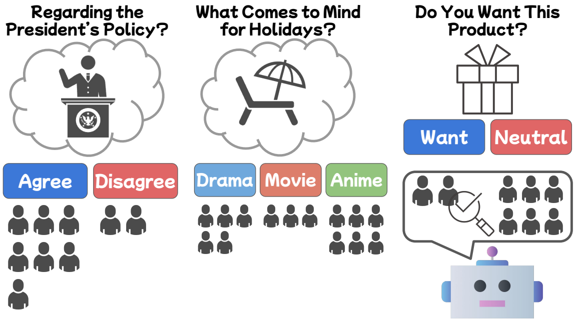
Regarding the President's Policy?
What Comes to Mind for Holidays?
Do You Want This Product?
Agree
Disagree
Drama
Movie
Anime
Want
Neutral

Figure 1: Example of response distribution. Analyzing both the ratios of each choice and the number of minority responses yields valuable insights.

LLMs have demonstrated the ability to partially replicate human collective tendencies by analyzing output probabilities or aggregating multiple outputs (Santurkar et al., 2023; Paruchuri et al., 2024; Hayashi et al., 2025). Providing appropriate input information has further improved the accuracy of these predictions (Durmus et al., 2024; Santurkar et al., 2023; Meister et al., 2024). These methods show promise as cost-effective and scalable alternatives to traditional techniques.

However, LLMs are unlikely to acquire systematic ratio-related knowledge during pretraining, e.g., the expected proportions of responses to a question such as *"What food do you associate with Christmas?"*[1]. This raises concerns about whether their ratio predictions reflect meaningful understanding or mere prompt-following (Kavumba et al., 2022). Additionally, measuring true response distributions is challenging (Baan et al., 2022), complicating validation and emphasizing the need for objective evaluation standards. If LLM predictions lack rationality or reproducibility, their use in social decision-making could pose risks.

In this study, we propose a framework to evaluate the reliability of LLMs' distribution prediction. Specifically, we introduce counterintuitive pseudo-distributions by altering existing survey

[1] This questionnaire is taken from Yahoo! News Polls: https://news.yahoo.co.jp/polls/48833.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 538–550
April 30 - May 1, 2025 ©2025 Association for Computational Linguistics

data and examine whether LLMs adjust their predictions or simply follow inconsistent explanations. Our findings indicate that smaller or less human-optimized models tend to follow inconsistent explanations uncritically, whereas larger or preference-optimized models are better at resisting counterintuitive distributions by leveraging pretraining-acquired knowledge. These results provide insights into factors influencing distribution prediction performance and highlight the variability in trustworthiness across different models, contributing to the development of more reliable distribution predictions using LLMs.

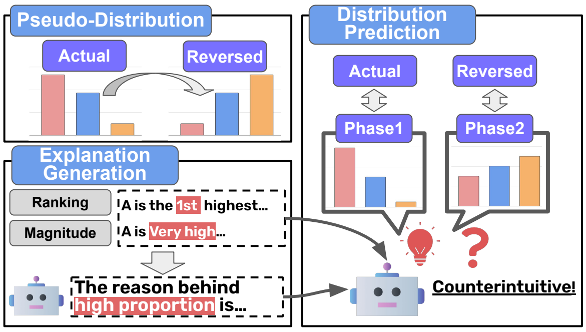

Figure 2: Overview of the proposed method. The actual distribution and a distribution with altered proportions are prepared, and explanations are generated for each. The score difference when estimating the distribution based on these explanations can be interpreted as the extent to which LLMs adjust based on commonsense knowledge.

2 Background and Related Work

2.1 Predicting Distributions by LLMs

Previous studies have explored LLMs' distribution prediction performance in contexts like annotation disagreements, survey data across countries, real-world probabilities, and preference predictions (Nie et al., 2020; Santurkar et al., 2023; Ohagi et al., 2024; Paruchuri et al., 2024; Meister et al., 2024). Common approaches involve using output probabilities for response options or aggregating multiple outputs to approximate distributions (Santurkar et al., 2023; Jiang et al., 2024; Zhou et al., 2022). Some studies report better reasoning performance when LLMs directly generate distributions in textual form (Meister et al., 2024; Suzuki et al., 2024).

While these studies confirm that LLMs exhibit some distribution prediction capabilities, the underlying rationale behind specific ratio predictions and the extent to which pretraining or preference learning influences these predictions remain unclear. Moreover, several studies have found that simple uniform distribution baselines, such as assigning equal ratios to all options, can sometimes outperform LLM-based predictions (Meister et al., 2024; Suzuki et al., 2024). This raises concerns about whether LLMs genuinely possess predictive capabilities or merely capture broad tendencies while generating numerically arbitrary estimates. Some studies suggest that LLMs can at least estimate majority opinions, even for questions where a definitive correct answer does not exist (Talmor et al., 2019; Nie et al., 2020; Sakai et al., 2024b).

2.2 Reasoning Abilities in LLMs

Many studies evaluate the reasoning capabilities of LLMs (Wei et al., 2022; Chowdhery et al., 2022), but numerous tasks can be solved by relying on word relationships or salient terms from the pretraining corpus, complicating the assessment of intrinsic reasoning abilities (Manning, 2006; Hosseini et al., 2021; Kung and Peng, 2023; Han et al., 2024). To overcome this, methods like reversing logical relationships or substituting nouns with fictitious names have been proposed to test reasoning independently of memorized knowledge or symbolic manipulation (Wu et al., 2024; Sakai et al., 2024a). However, in distribution prediction, response ratio interrelations are crucial (Suzuki et al., 2024), and simple substitutions risk altering the problem's intent. For example, while *"I don't know"* and *"No response"* appear similar, their motivations differ: *"I don't know"* indicates a lack of understanding, whereas *"No response"* signifies an intentional decision not to answer. Conflating them may result in misinterpreting the distributions.

3 Proposed Method

Our evaluation involves inputting explanations of the actual distribution along with the question, either to support or potentially distract the prediction, in order to better capture the model's true prediction ability. As shown in Figure 2, we design a two-phase experimental framework to evaluate whether LLMs can make rational distribution predictions based on knowledge acquired during pretraining.

Phase 1: Do LLMs Predict Distributions Based on Provided Explanations? This phase investigates whether LLMs can accurately predict response ratios from qualitative explanations. Figure 3 provides an overview of the explanation generation process. First, LLMs generate explanatory

Usage	*Example*
Question	Which team do you think will win the World Series, the Dodgers or the Yankees?
Options	Dodgers, Yankees, Not sure
Actual Distribution	{“Dodgers”: *0.81*, “Yankees”: 0.14, “Not sure”: *0.05*}
Reversed Distribution	{“Dodgers”: *0.05*, “Yankees”: 0.14, “Not sure”: *0.81*}
Ranking	The percentage for “Dodgers” is *the first highest*, “Yankees” is *the second highest*, and “Not sure” is *the third highest*.
Magnitude	The percentage for “Dodgers” is *very high*, the percentage for “Yankees” is *low*, and the percentage for “Not sure” is *low*.
Ranking Explanation	This distribution of responses is shaped by factors such as fan support, past team performance, and recent results. The high level of support for the “Dodgers” is likely due to their popularity, strong performance, or strong backing from local fans. The “Yankees,” being a traditional powerhouse team with a large fan base, receive the second highest level of support. Those who chose “Not sure” likely reflect uncertainty about the outcome of the games or a lack of in-depth knowledge about baseball.

Table 1: An example of a question with its options and original proportions along with an altered set of proportions. Also shown are (a) Ranking and (b) Magnitude information for this question, along with a sample explanation based on (a) Ranking. This explanation was generated by the Qwen 2.5 (Qwen et al., 2024) model with 32B parameters. The original inputs were in Japanese, but are translated into English here.

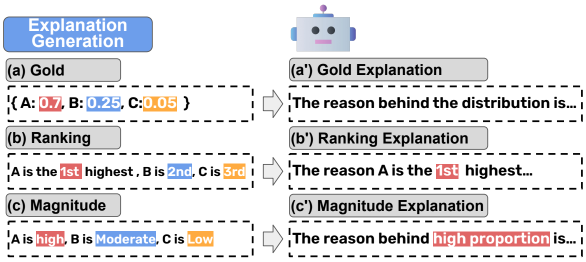
Explanation Generation
(a) Gold
{ A: 0.7, B: 0.25, C:0.05 }
(b) Ranking
A is the 1st highest , B is 2nd, C is 3rd
(c) Magnitude
A is high, B is Moderate, C is Low
(a') Gold Explanation
The reason behind the distribution is...
(b') Ranking Explanation
The reason A is the 1st highest...
(c') Magnitude Explanation
The reason behind high proportion is...

Figure 3: Overview of explanation generation. From (a) actual survey, (b) *Ranking* and (c) *Magnitude* are automatically derived. Then, based on (a, b, c), LLMs generate corresponding explanations (a’, b’, c’).

descriptions of the target distributions based on three types of input information: (a) *Gold* (actual numerical distributions), (b) *Ranking* (order relationships, such as “1st,” “2nd,” or “3rd”), and (c) *Magnitude* (relative proportion sizes, such as “Very High,” “High,” or “Low”). The explanations generated from these respective inputs are referred to as (a’) *GoldExp*, (b’) *RankingExp*, and (c’) *MagnitudeExp*. Next, these explanations are provided as input to the LLMs, which then generate reconstructed distributions. Finally, we compare the predicted distributions with the actual ones. An example question and its corresponding explanation used in this phase are shown in Table 1.

Note that evaluating LLMs solely based on the reconstructed distributions from their explanations (a’, b’, c’) may introduce biases unrelated to distribution prediction capability, as the results could be affected by the models’ explanation abilities. To address this, we also measure distribution prediction performance independent of explanation ability by predicting (a) directly from (b, c). Therefore, LLMs predict the distribution from five types of explanations (b, c, a’, b’, c’). Appendix A provides the prompts and detailed descriptions.

Phase 2: Do LLMs Adjust for Counterintuitive Explanations? In the second phase, we evaluated the ability of LLMs to recognize inconsistencies and adjust ratios by introducing pseudo-distributions that are commonsensically implausible. This experiment used the following two types of pseudo-distribution settings: (i) *Swapped*: The proportions of the first and second highest values are swapped. (ii) *Reversed*: The highest and lowest proportions are exchanged. These pseudo-distributions differ from actual distributions and are against commonsense expectations, with *Reversed* setting being considered greater inconsistent.

As in Phase 1, LLMs generate explanations from these pseudo-distributions and predict response distributions. If accuracy remains unchanged, the model is likely following explanations without evaluating plausibility. A decline in accuracy would suggest the model detects inconsistencies and adjusts predictions using commonsense reasoning.

4 Experimental Setup

Dataset We utilized the “Yahoo! News Polls”[2] provided by LY Corporation to create evaluation response distributions. This dataset comprises survey results related to articles published on Yahoo! News, covering the period from January 2020 to December 2024 in Japanese. We extracted questions with three options, resulting in a total of 714 items for analysis. Focusing on Japanese data allows us to reduce the ambiguity in predictions caused by cultural differences compared to conventional English datasets. Furthermore, since this data is based on freely cast votes on the internet, it is considered highly compatible with LLMs, which are primarily pretrained on internet data.

[2] https://news.yahoo.co.jp/polls

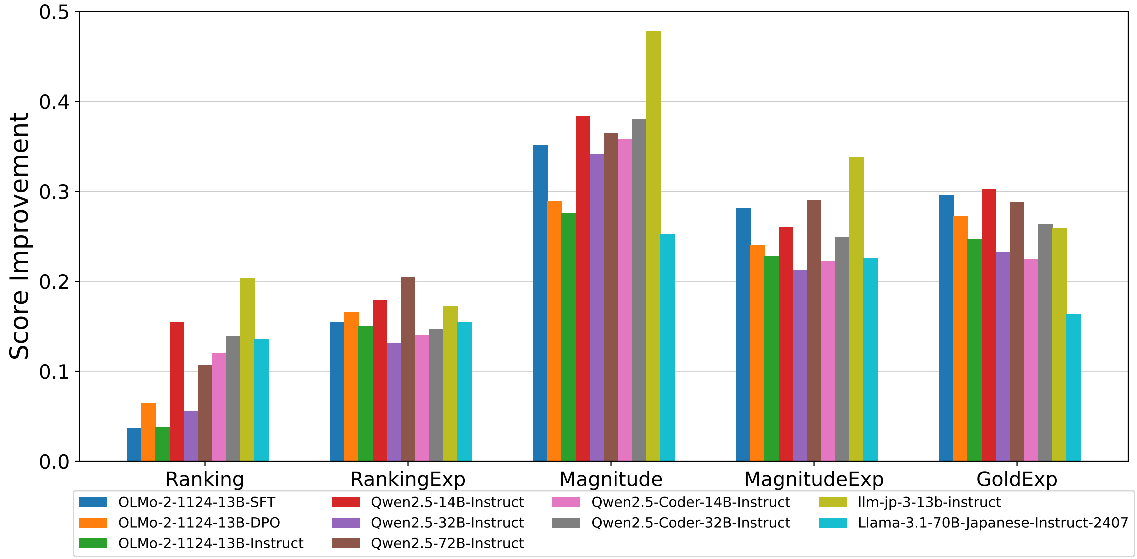

Figure 4: Score improvements across conditions compared to predictions without explanations. Improvements are visualized as positive values (upward).

LLMs We used ten high-performing open-source models, including Qwen 2.5 (Qwen et al., 2024) with 14B, 32B, and 72B parameters, as well as code-generation versions (Hui et al., 2024) with 14B and 32B parameters (all Instruct versions). These models were chosen to examine the effects of parameter size and code-learning on reasoning performance. To evaluate the impact of preference learning, we also included OLMo-2 (OLMo et al., 2024) in its SFT, DPO, and Instruct versions, where human preference alignment is progressively incorporated from supervised finetuning (SFT) to direct preference optimization (DPO) (Rafailov et al., 2023) and further to Reinforcement Learning with Verifiable Rewards (Instruct) (OLMo et al., 2024). Since the evaluation datasets are in Japanese, we employed llm-jp-3-13b-instruct (LLM-jp et al., 2024), which was pretrained in Japanese, and Llama-3.1-70B-Japanese-Instruct-2407 (Ishigami, 2024), a continuously trained Llama 3.1 (Dubey et al., 2024) on Japanese data. We used the 8-bit quantization inferences (Dettmers et al., 2022). We employed greedy decoding in inference.

Evaluation Methods To measure the similarity between the LLM predictions and the gold distributions, we adopted the Total Variation Distance (TVD). TVD is defined as the sum of the absolute differences between the gold (or pseudo-gold, in our experiments) values and the model's predicted values for each option. A lower TVD indicates closer alignment between the LLM predictions and the correct distribution. After minor output adjustments[3], over 90% of the data were analyzable as JSON-formatted response distributions. For cases where the valid response rate fell below 90%, results were recorded as reference values[4]. Finally, the average TVD, excluding missing values, was calculated.

5 Experimental Results

Phase 1: Do LLMs Predict Distributions Based on Explanations? Figure 4 shows the improvement in scores when models were provided with explanations generated based on these attributes, compared to when no explanation was given. All models showed improved scores across all conditions, reinforcing previous findings that providing appropriate contextual information enhances prediction performance.

For *Ranking*, which does not directly provide numerical hints, the condition *RankingExp* where the model supplements relevant background information, led to further score improvements in many models. In contrast, for *Magnitude*, which provides direct numerical hints, the condition *MagnitudeExp*, where an explanation accompanies the magnitude information, resulted in lower scores. This decline is likely due to the omission of explanations for minority options in some questions,

[3] This included converting full-width symbols to half-width and normalizing distributions to 1.0 if their sum equaled 100%.

[4] Details on valid rates are provided in Appendix C.

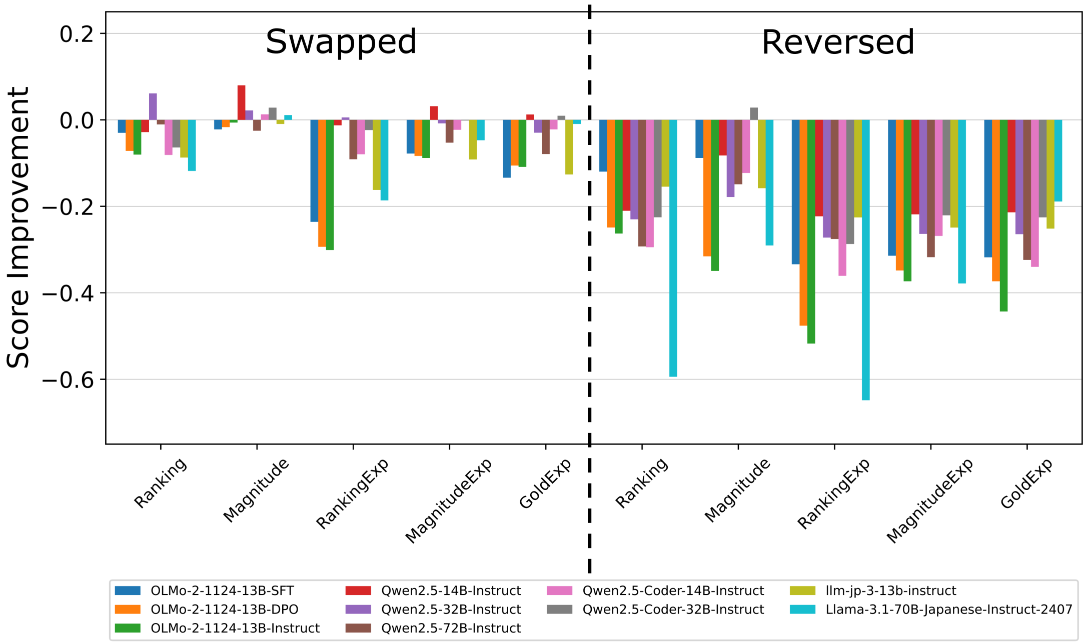

Figure 5: Changes in scores from the first phase of distribution prediction. Larger declines in scores indicate that the model, while considering the provided counterintuitive explanations, made commonsense-based adjustments to correct inconsistencies.

reducing the amount of provided information.

For *GoldExp*, where explanations were generated based on the actual response distributions, exhibited a similar level of improvement to *MagnitudeExp*. This suggests that even approximate magnitude-based explanations can enhance predictive accuracy to a degree comparable to using actual numerical values.

Phase 2: Do LLMs Adjust for Counterintuitive Explanations? Figure 5 shows the differences in average scores between the first and second phases, categorized by settings and conditions.

A large drop in scores was observed under *Reversed* condition, which introduces greater inconsistency compared to *Swapped*. This suggests that many models recognized contradictions between the pseudo-distributions and commonsense expectations. Notably, even in conditions where all models received the same *Ranking* and *Magnitude* information, *Reversed* condition resulted in a greater score decline than *Swapped*. This implies that LLMs leverage pretraining-acquired knowledge to some extent when making predictions.

However, the degree of adjustment varied across models. For example, in OLMo-2, adjustment capabilities improved progressively from SFT to DPO. This trend suggests that DPO, which is designed to align model outputs with human preferences (Rafailov et al., 2023), enhances response distribution prediction performance. Similarly in Qwen 2.5, while smaller models tended to follow counterintuitive explanations, larger models demonstrated more accurate predictions. This pattern was also observed in Japanese-trained models, where Llama-3.1-70B-Japanese showed superior adjustment capabilities. These findings indicate that model size, as well as pretraining and fine-tuning strategies, contribute to improving commonsense-based numerical adjustments.

6 Analysis

6.1 Naturalness as Causal Modeling

We re-tokenized the generated text and calculated its perplexity as a continuation of the input prompt[5]. A higher perplexity value indicates that the output is less natural for the model, allowing for a quantitative evaluation of deviations from pretraining expectations. The results are shown in Table 2.

For OLMo-2, there is little change in perplexity between *Actual* and *Reversed* conditions. In contrast, models larger than Qwen 2.5-14B exhibit increased perplexity in the *Ranking* setting when shifting from *Actual* to *Reversed*. This suggests that

[5] Due to tokenizer effects, the token sequence during re-tokenization may not always match the original generated sequence.

Model	Ranking			Magnitude		
	Actual	Swapped	Reversed	Actual	Swapped	Reversed
OLMo-2-1124-13B-SFT	1.14 ± 0.06	1.12 ± 0.05	1.10 ± 0.05	1.13 ± 0.06	1.12 ± 0.06	1.14 ± 0.07
OLMo-2-1124-13B-DPO	1.28 ± 0.07	1.26 ± 0.07	1.25 ± 0.07	1.28 ± 0.07	1.28 ± 0.07	1.28 ± 0.07
OLMo-2-1124-13B-Instruct	1.04 ± 0.04	1.04 ± 0.03	1.04 ± 0.03	1.05 ± 0.04	1.05 ± 0.04	1.06 ± 0.05
Qwen2.5-14B-Instruct	1.47 ± 0.10	1.46 ± 0.09	1.46 ± 0.11	1.41 ± 0.11	1.37 ± 0.11	1.42 ± 0.10
Qwen2.5-32B-Instruct	1.12 ± 0.12	1.12 ± 0.10	1.32 ± 0.12	1.10 ± 0.09	1.12 ± 0.11	1.17 ± 0.14
Qwen2.5-72B-Instruct	1.07 ± 0.06	1.08 ± 0.06	1.13 ± 0.10	1.05 ± 0.04	1.06 ± 0.05	1.07 ± 0.06
Qwen2.5-Coder-14B-Instruct	1.08 ± 0.02	1.10 ± 0.03	1.09 ± 0.03	1.08 ± 0.03	1.07 ± 0.03	1.08 ± 0.03
Qwen2.5-Coder-32B-Instruct	1.32 ± 0.11	1.31 ± 0.09	1.35 ± 0.09	1.27 ± 0.13	1.26 ± 0.12	1.25 ± 0.13
llm-jp-3-13b-instruct	4.70 ± 1.19	4.69 ± 1.10	4.81 ± 1.11	4.96 ± 1.21	4.77 ± 1.11	4.90 ± 1.20
Llama-3.1-70B-Japanese-Instruct-2407	1.13 ± 0.09	1.16 ± 0.10	1.16 ± 0.10	1.11 ± 0.08	1.09 ± 0.08	1.12 ± 0.09

Table 2: Perplexities for cases with ranking or magnitude information under various settings. A higher perplexity value indicates that the output is less natural for the model.

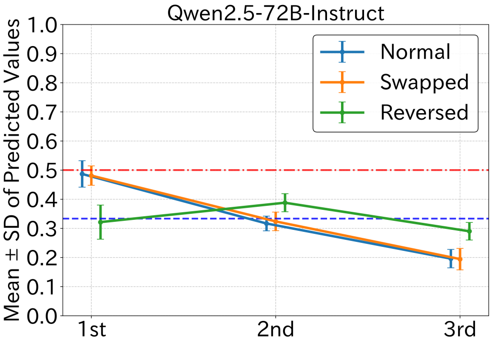

Figure 6: Average proportions predicted for ranked options when ranking information is provided.

while OLMo-2 and similarly sized models, such as Qwen 2.5-14B, do not necessarily treat counterintuitive predictions as unnatural at the internal representation level, larger models are more capable of doing so. Additionally, even for large models, the high standard deviation suggests that model behavior varies largely across different questions. These results suggest the usefulness of leveraging log probabilities of response options or sampling-based methods for distribution prediction, particularly when employing large-scale models.

6.2 Ranking Explanations and Predictions

Figure 6 shows the average proportions assigned by Qwen 2.5-72B to each ranked option when ranking information is available. Given probability distribution properties, the highest proportion does not fall below the dotted blue line indicating 0.33, and the lowest does not exceed 0.33 in the absence of ties. Consequently, models like Qwen 2.5-72B, which adjust values within a rational range, may be underestimated. In contrast, some cases highlight the risk of overestimating models that appear aligned with commonsense reasoning while violating probability constraints, as shown in Figure 7 in Appendix D.

Moreover, if the dataset lacks high-proportion options, score differences may be artificially low, leading to inaccurate model assessments. While our framework effectively distinguishes between instruction-following and commonsense-based predictions, it has limitations in evaluating probability rationality. Ensuring a balanced dataset mitigates these issues. Notably, previous studies have focused on refining distance metrics but overlooked dataset composition, highlighting the need for improvements in evaluation reliability.

7 Conclusion

This study examined whether language models predict response distributions based on rational reasoning with commonsense knowledge or merely follow instructions. By altering survey data ratios, we analyzed model predictions under inconsistent conditions. The experimental results showed that in the highly inconsistent *Reversed* condition, larger models and those fine-tuned with preference learning tended to correct inconsistencies using commonsense knowledge. Smaller models either showed little change or adapted to the inconsistencies. These findings evaluate aspects of prediction capability that conventional studies cannot measure and offer insights into selecting reliable models for distribution prediction. The proposed method is adaptable across languages and dataset types. Therefore, future work should include experiments in multilingual settings, including English, to investigate the influence of cultural factors. Additionally, measuring and mitigating biases using statistically reliable data, such as government-conducted surveys, is an important direction for future research.

Limitations

This study relies on internet-based survey data, which could contain biases. However, as internet data is widely used for pretraining language models and aligns with their commonsense knowledge, it serves as a meaningful baseline for evaluating pseudo-distribution consistency with commonsense reasoning. Ensuring statistical accuracy for practical applications remains a challenge, and model predictions may vary over time. While this study does not explicitly address temporal changes, Yahoo! News Polls is publicly accessible, allowing future research to refine statistical accuracy and analyze time-dependent trends. However, limited variations in the prompt templates used in our experiments could affect the experimental outcomes (Sakai et al., 2024c). Investigating such variability in outputs is also left for future work. In addition, we do not take into account factors of confidence during prediction when evaluating performance such as Ozaki et al. (2024). This perspective may yield more insights into our findings.

Ethical Considerations

Rather than reinforcing biases, this study aims to identify and examine them. By analyzing how biases manifest in model predictions, we contribute to a deeper understanding of their impact and support the development of fairer, more robust evaluation methods. Finally, Yahoo! News Polls, which was used in this study, is licensed for research use, so there are no license issues.

References

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao

Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. 2024. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Kazuki Hayashi, Kazuma Onishi, Toma Suzuki, Yusuke Ide, Seiji Gobara, Shigeki Saito, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025. IRR: Image review ranking framework for evaluating vision-language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9939–9956, Abu Dhabi, UAE. Association for Computational Linguistics.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312, Online. Association for Computational Linguistics.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. 2024. Qwen2.5-coder technical report. *Preprint*, arXiv:2409.12186.

Ryosuke Ishigami. 2024. cyberagent/llama-3.1-70b-japanese-instruct-2407.

Liwei Jiang, Sydney Levine, and Yejin Choi. 2024. Can language models reason about individualistic human values and preferences? In *Pluralistic Alignment Workshop at NeurIPS 2024*.

Pride Kavumba, Ryo Takahashi, and Yusuke Oda. 2022. Are prompt-based models clueless? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2333–2352, Dublin, Ireland. Association for Computational Linguistics.

Po-Nien Kung and Nanyun Peng. 2023. Do models really learn to follow instructions? an empirical study of instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1317–1328, Toronto, Canada. Association for Computational Linguistics.

LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, Hiroshi Kataoka, Satoru Katsumata, Daisuke Kawahara, Seiya Kawano, Atsushi Keyaki, Keisuke Kiryu, Hirokazu Kiyomaru, Takashi Kodama, Takahiro Kubo, Yohei Kuga, Ryoma Kumon, Shuhei Kurita, Sadao Kurohashi, Conglong Li, Taiki Maekawa, Hiroshi Matsuda, Yusuke Miyao, Kentaro Mizuki, Sakae Mizuki, Yugo Murawaki, Akim Mousterou, Ryo Nakamura, Taishi Nakamura, Kouta Nakayama, Tomoka Nakazato, Takuro Niitsuma, Jiro Nishitoba, Yusuke Oda, Hayato Ogawa, Takumi Okamoto, Naoaki Okazaki, Yohei Oseki, Shintaro Ozaki, Koki Ryu, Rafal Rzepka, Keisuke Sakaguchi, Shota Sasaki, Satoshi Sekine, Kohei Suda, Saku Sugawara, Issa Sugiura, Hiroaki Sugiyama, Hisami Suzuki, Jun Suzuki, Toyotaro Suzumura, Kensuke Tachibana, Yu Takagi, Kyosuke Takami, Koichi Takeda, Masashi Takeshita, Masahiro Tanaka, Kenjiro Taura, Arseny Tolmachev, Nobuhiro Ueda, Zhen Wan, Shuntaro Yada, Sakiko Yahata, Yuya Yamamoto, Yusuke Yamauchi, Hitomi Yanaka, Rio Yokota, and Koichiro Yoshino. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *Preprint*, arXiv:2407.03963.

Christopher D. Manning. 2006. Local textual inference: It's hard to circumscribe, but you know it when you see it—and nlp needs it. pages 1–12.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. 2024. Benchmarking distributional alignment of large language models. *Preprint*, arXiv:2411.05403.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Masaya Ohagi, Junya Takayama, Tomoya Mizumoto, and Katsumasa Yoshikawa. 2024. Proposal of a response distribution prediction method considering relationships between choices using large language models [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]. In *Technical Report of the 260th Special Interest Group on Natural Language Processing*, volume 16, pages 1–13, Kanazawa, Japan. Information Processing Society of Japan. (In Japanese).

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. 2024. 2 olmo 2 furious.

Shintaro Ozaki, Yuta Kato, Siyuan Feng, Masayo Tomita, Kazuki Hayashi, Ryoma Obara, Masafumi Oyamada, Katsuhiko Hayashi, Hidetaka Kamigaito, and Taro Watanabe. 2024. Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain. *Preprint*, arXiv:2412.20309.

Akshay Paruchuri, Jake Garrison, Shun Liao, John B Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. 2024. What are the odds? language models are capable of probabilistic reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11712–11733, Miami, Florida, USA. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2024a. Does pre-trained language model actually infer unseen links in knowledge graph completion? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8091–8106, Mexico City, Mexico. Association for Computational Linguistics.

Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024b. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14182–14214, Bangkok, Thailand. Association for Computational Linguistics.

Yusuke Sakai, Adam Nohejl, Jiangnan Hang, Hidetaka Kamigaito, and Taro Watanabe. 2024c. Toward the evaluation of large language models considering score variance across instruction templates. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 499–529, Miami, Florida, US. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Toma Suzuki, Ayuki Katayama, Seiji Gobara, Ryo Tsujimoto, Hibiki Nakatani, Kazuki Hayashi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. Proposal of a response distribution prediction method considering relationships between choices using large language models [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]. In *Technical Report of the 262nd Special Interest Group on Natural Language Processing*, volume 40, pages 1–14, Nagoya, Japan. Information Processing Society of Japan. (In Japanese).

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

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

Xiang Zhou, Yixin Nie, and Mohit Bansal. 2022. Distributed NLI: Learning to predict human opinion distributions for language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 972–987, Dublin, Ireland. Association for Computational Linguistics.

A Prompt Details

The templates used in our experiments are shown in Table 3. Additionally, explanations of proportions were mechanically replaced based on the rules provided in Table 4.

In the first phase, LLMs were instructed to generate explanatory descriptions of survey results, excluding specific numerical values or ratios, while considering the survey periods. In the second phase, these generated descriptions were used as prompts, and LLMs were tasked with predicting response distributions in JSON format (Meister et al., 2024; Suzuki et al., 2024). If the explanations contained numerical values or ratios, they were replaced with "—" using regular expressions before being provided to the model. To ensure correct output formatting, JSON-format examples were also included in the prompt.

Usage Scenario	*Template*
Explanation Generation (Translated)	Please explain why the response distribution for the following question turned out this way, without including any specific numbers or percentages. Keep your explanation concise and within 300 characters. Survey period: October 21, 2024 – October 31, 2024 Question: Which team do you think will win the World Series, the Dodgers or the Yankees? Options: "Dodgers", "Yankees", "Not sure" Response Distribution: The percentage for "Dodgers" is the first highest, "Yankees" is the second highest, and "Not sure" is the third highest. Explanation:
Explanation Generation	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。 実施期間: 2024-10-21～2024-10-31 質問: ドジャースとヤンキース、どちらがワールドシリーズを制覇すると思いますか？ 選択肢: "ドジャース", "ヤンキース", "わからない" 回答分布: 「ドジャース」の割合は1番目に高く、「ヤンキース」は2番目、「わからない」は3番目に高いです。 説明:
Distribution Prediction (Translated)	Please predict the response distribution for the following question and options, based on the explanation provided. Your answer should be in JSON format, and the sum of the proportions for all choices must equal 1.0. Survey period: October 21, 2024 – October 31, 2024 Question: Which team do you think will win the World Series, the Dodgers or the Yankees? Options: "Dodgers", "Yankees", "Not sure" Explanation: This distribution of responses is shaped by factors such as fan support, past team performance, and recent results. The high level of support for the "Dodgers" is likely due to their popularity, strong performance, or strong backing from local fans. The "Yankees," being a traditional powerhouse team with a large fan base, receive the second highest level of support. Those who chose "Not sure" likely reflect uncertainty about the outcome of the games or a lack of in-depth knowledge about baseball. Example output format: {"Dodgers": –, "Yankees": –, "Not sure": –} Response distribution:
Explanation Generation	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。 実施期間: 2024-10-21～2024-10-31 質問: ドジャースとヤンキース、どちらがワールドシリーズを制覇すると思いますか？ 選択肢: "ドジャース", "ヤンキース", "わからない" 説明: この回答分布は、ファンの支持やチームの過去のパフォーマンス、最近の成績などの要因によって形成されています。 「ドジャース」への支持が高いのは、彼らの人気や優れたパフォーマンス、あるいは地元ファンからの強い支持があるからでしょう。 「ヤンキース」も伝統のある強豪チームであり、多くのファンや支持者がいるため、2番目の支持を得ています。 「わからない」を選んだ人々は、試合の結果に対する不確実性や、野球の専門知識が不足していることを示しています。 回答分布の出力例: {"ドジャース": –, "ヤンキース": –, "わからない": –} 回答分布:

Table 3: Details of the prompts used in the experiment. All inputs were provided in Japanese. For reference, English translations of the prompts are also included.

Ratio Range	*Descriptive Category*
$x \geq 0.75$	Very High (非常に高い)
$0.5 \leq x < 0.75$	High (高い)
$0.25 \leq x < 0.5$	Moderate (中程度)
$x < 0.25$	Low (低い)

Table 4: Correspondence between ratio ranges and descriptive categories (Japanese are provided in parentheses).

B Spearman's Rank Correlation Coefficient

We calculated Spearman's rank correlation coefficients for model rankings based on distribution prediction scores without explanations and those with various types of added explanations (See Table 5). The rankings with commonsense explanations showed significant positive correlations, whereas there was little correlation with the rankings based on predictions from counterintuitive explanations. This

Condition	Explanation	Correlation Coefficient	p-value
Actual	Ranking	0.49	0.1497
	Magnitude	0.15	0.6761
	Ranking/Explanation	*0.94*	*0.0001*
	Magnitude/Explanation	*0.81*	*0.0049*
	Gold/Explanation	*0.70*	*0.0251*
Swapped	Ranking	*0.65*	*0.0425*
	Explanation	0.18	0.6272
	Ranking/Explanation	0.43	0.2145
	Magnitude/Explanation	*0.64*	*0.0479*
	Gold/Explanation	0.44	0.2004
Reversed	Ranking	-0.56	0.0897
	Explanation	-0.36	0.3104
	Ranking/Explanation	-0.16	0.6515
	Magnitude/Explanation	-0.03	0.9338
	Gold/Explanation	0.28	0.4250

Table 5: Spearman's Rank Correlation Coefficients and p-values Between Score Rankings.

suggests that instruction-following performance and commonsense-based ratio prediction capabilities may independently influence model performance.

C Valid Response Rate

Table 6 shows the average Valid Response Rate across all conditions for each setting. The Valid Response Rate represents the proportion of model outputs that could be parsed as response distributions in JSON format. Asterisks (*) indicate cases where the Valid Response Rate did not exceed the threshold of 90%. Under the reversed setting/ranking condition for Llama-3.1-70B-Japanese-Instruct-2407, the valid response rate fell to 89.5%, below the 90% threshold.

Model	*No Explanation*	*Actual*	*Swapped*	*Reversed*
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**

Table 6: Average Valid Response Rate (%) Across Settings

D Ranking Information and Actual Predicted Values

As in Section 6.2, we plotted the average values assigned to each option when ranking information was provided for all models in Figure 7. Note that Llama-3.1-70B-Japanese-Instruct-2407 strongly adheres to commonsense reasoning but produces predictions that conflict with the properties of probability distributions.

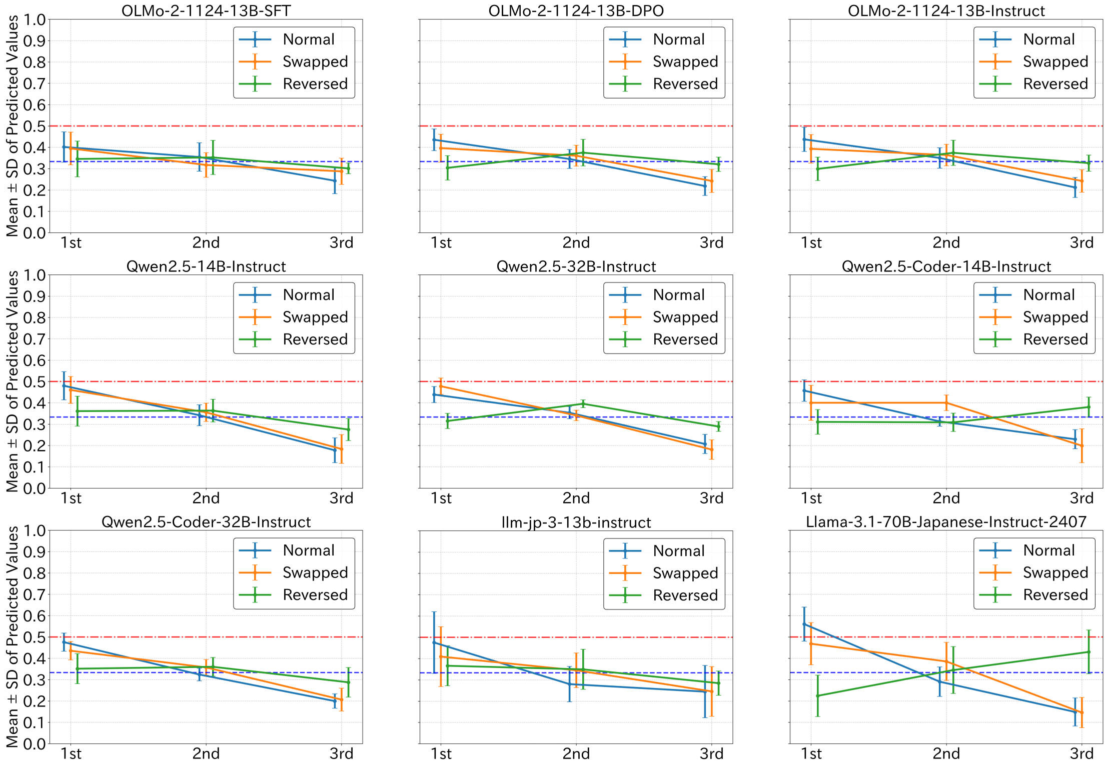

Figure 7: Average proportions predicted for ranked options when ranking information is provided, for all models in our experiment.

E The distribution of proportions in the dataset

Figure 8 shows violin and box plots illustrating the distribution of proportions in the evaluation dataset.

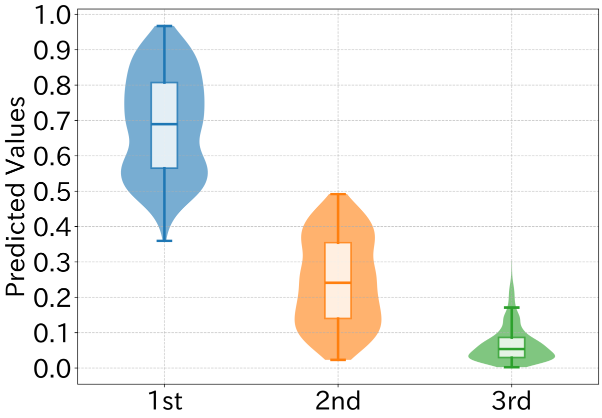

Figure 8: Violin and box plots showing the distribution of proportions in the evaluation dataset. In the absence of ties, the first rank falls within the range $(0.33.., 1.0)$, the second rank within (0, 0.5), and the third rank within [0, $0.33..)$.