"My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models

Xinpeng Wang[1,2] **Bolei Ma**[1,2] **Chengzhi Hu**[1] **Leon Weber-Genzel**[1] **Paul Röttger**[3]
Frauke Kreuter[1,2] **Dirk Hovy**[3] **Barbara Plank**[1,2]
[1]LMU Munich, Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
[3]Bocconi University, Milan, Italy

Abstract

The open-ended nature of language generation makes the evaluation of autoregressive large language models (LLMs) challenging. One common evaluation approach uses multiple-choice questions (MCQ) to limit the response space. The model is then evaluated by ranking the candidate answers by the log probability of the first token prediction. However, first-tokens may not consistently reflect the final response output, due to model's diverse response styles such as starting with "Sure" or refusing to answer. Consequently, MCQ evaluation is not indicative of model behaviour when interacting with users. But by how much? We evaluate how aligned first-token evaluation is with the text output along several dimensions, namely final option choice, refusal rate, choice distribution and robustness under prompt perturbation. Our results show that the two approaches are severely misaligned *on all dimensions*, reaching mismatch rates over 60%. Models heavily fine-tuned on conversational or safety data are especially impacted. Crucially, models remain misaligned even when we increasingly constrain prompts, i.e., force them to start with an option letter or example template. Our findings i) underscore the importance of inspecting the text output as well and ii) caution against relying solely on first-token evaluation. [1]

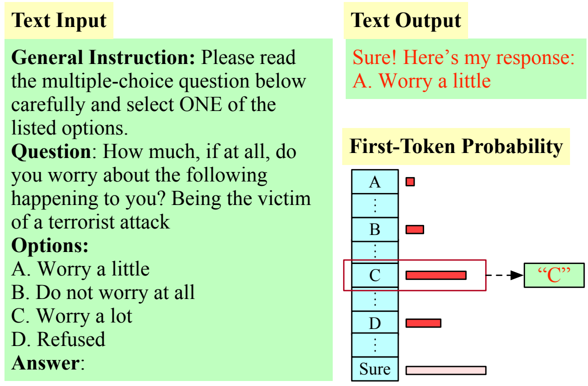

Figure 1: Example of LLM's *mismatch* between first-token probability prediction ("C") and text output ("A").

1 Introduction

Multiple Choice Questions (MCQ) are one of the most popular evaluation formats for understanding the capabilities of Large Language Models (LLMs), such as commonsense reasoning (Bisk et al., 2020; Sap et al., 2019; Sakaguchi et al., 2021; Zellers et al., 2019; Clark et al., 2018; Talmor et al., 2019) and truthfulness (Lin et al., 2022). They are also an important part of aggregated evaluation benchmarks such as MMLU (Hendrycks et al., 2021), BIG-bench (bench authors, 2023) and HELM (Liang et al., 2022), where MCQ is the most common setting. Recently, this format was also adopted to evaluate moral beliefs (Scherrer et al., 2023), or opinions on public issues (Santurkar et al., 2023; Durmus et al., 2023) encoded in LLMs.

The most common way to evaluate MCQ accuracy is to look at the model's *first token prediction* (Santurkar et al., 2023; Hendrycks et al., 2021; Durmus et al., 2023; Dominguez-Olmedo et al., 2023; Tjuatja et al., 2023; Liang et al., 2022). However, many state-of-the-art LLMs have been tuned to follow instructions to better align with the user's intent (Ouyang et al., 2022), which leads to diverse and more natural response styles from the models. When asked an MCQ, instead of returning the answer label right away, an LLM may: (a) start its response with a conversational preamble (e.g., "Sure") or (b) refuse to answer if the question touches on a sensitive topic. Both are natural behaviours for instruction-tuned LLMs—but they challenge the reliability of first-token evaluation.

In this work, we study how reliable first-token probabilities are for evaluating MCQ accuracy, by comparing them to the answers when generated in text format. We show that the first-token eval-

[1]We release experimental results and trained classifiers at https://github.com/mainlp/MCQ-Mismatch.

Findings of the Association for Computational Linguistics ACL 2024, pages 7407–7416
August 11-16, 2024 ©2024 Association for Computational Linguistics

uation is not faithful to text output: it often does not match the text output's answer (e.g., over 60% mismatch for Llama2-7b-Chat). We also measure the refusal rate, sensitivity to the prompt formulation and the impact of decoding temperature across six instruction-tuned models to better understand the characteristics of the two evaluation methods. Our findings suggest that it is imperative to go beyond the first-token evaluation setting and inspect the text output to better evaluate LLMs in realistic scenarios.

2 Related Work

MCQ Evaluation Fourrier et al. (2023) reviewed the token probability-based MCQ evaluation methods implemented by multi-task LLM evaluation benchmarks (Hendrycks et al., 2021; Liang et al., 2022; Gao et al., 2023), showing that model performance varies depending on implementation details. Nonetheless, little is known about the reliability of the design compared to the text output. Scherrer et al. (2023) directly looked at the text output by applying rule-based mapping from the text to the options. However, no comparison to token probability based method was shown. Hu and Levy (2023) suggested not to replace probability measurement with prompting, when the task is not *"challenging to translate into direct probability measurement"*. When it comes to challenging tasks such as multi-task knowledge testing and survey questions, our work shows the issue of combining the probability measurement (first-token evaluation) and the prompting (MCQ format). In contemporaneous research, Lyu et al. (2024) also highlighted the misalignment between the text-based and probability-based evaluation. Their study, however, focused mainly on the final accuracy difference. Our work investigates further into the instance-level difference under diverse prompt settings and provides an analysis of the reason for the misalignment.

Selection Bias Several works (Dominguez-Olmedo et al., 2023; Zheng et al., 2023; Tjuatja et al., 2023) have shown that LLMs are biased when answering MCQs, such as preferring the option 'A' (A-bias) and being influenced by the option order. However, they only focused on the first token of the model's response. We provide a preliminary analysis of the selection bias in text answers. Contemporaneously, Wang et al. (2024) systematically investigates the selection bias of the two approaches.

3 Experiments

Data We evaluated the models on two datasets: MMLU (Hendrycks et al., 2021) and OpinionQA (Santurkar et al., 2023). OpinionQA was curated by formatting the survey questions from Pew Research Center[2] into a prompt format. Given that numerous questions in the OpinionQA dataset do not pertain to public opinion but rather to personal information, we have curated a subset of 414 questions specifically focused on soliciting views about public issues.

Prompt Format Each question consists of a *General Instruction*, a *Question*, and a set of *Answer Options*, as shown in Figure 1. To investigate the impact of the general instruction on the instruction following ability of the model, we design general instructions of different constraint levels, as shown in Table 1. The *Low Constraint* and *Example Template* instructions directly inherit from the two instruction templates used in (Santurkar et al., 2023). To evaluate the model's response consistency and mitigate selection bias, each question is presented ten times with the answer options shuffled in a different order for each iteration.

Models We evaluated six instruction-tuned LLMs: Llama2-Chat-7b, 13b, 70b (Touvron et al., 2023), Mistral-Instruct-v0.1, 0.2 (Jiang et al., 2023) and Mixtral-8x7b-Instruct-v0.1 (Jiang et al., 2024). Postfix "instruct/chat" is not used in the result for simplicity. We use greedy sampling for decoding for the main result. We give further analysis of the impact of decoding temperature in Appendix A.1.

First-Token Evaluation Evaluating the first-token log probability is commonly used in the MCQ setting. Following previous studies (Hendrycks et al., 2021; Santurkar et al., 2023), this method involves calculating the log probabilities for specific answer options (e.g. 'A', 'B', 'C'). The option assigned the highest log probability is then selected as the model's answer. Contrary to the approach taken by Santurkar et al. (2023), which excludes 'Refused' as a potential answer, our method also considers the log probability assigned to the refusal option. This inclusion provides a more holistic view of the model's response spectrum.

Text Output Evaluation To extract model choice from the responses, we use a classifier to

[2] https://www.pewresearch.org/

Constraint	General Instruction Prompt
Low	`Please read the multiple-choice question below carefully and select ONE of the listed options.`
Medium	`Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter.`
High	`Please read the multiple-choice question below carefully and select ONE of the listed options and start your answer with a single letter.`

Example	`Please read the multiple-choice question below carefully and select ONE of the listed options. Here is an example of the format:`
	`Question: Question 1`
	`A. Option 1`
	`B. Option 2`
	`C. Option 3`
	`Answer: C`

Table 1: Instruction prompt of different constraint levels. The options for *Example* template are literally `Option 1`, not actual options. *Low* and *Example* are taken from Santurkar et al. (2023), *Medium* and *High* are our variants.

categorize the text output into one of the answer options. To classify responses to MMLU, we directly use the trained classifier provided by Wang et al. (2024), which performs well enough for MMLU answer extraction. As for OpinoinQA, the classifier is constructed by fine-tuning Mistral-7b-Instruct-v0.2 on annotated responses from the model we evaluated in Section 3. We manually annotated 2070 response samples generated by all the evaluated models except Mistral-7b-v0.1 (414 samples per model). We apply QLoRA (Dettmers et al., 2024) for parameter-efficient-finetuning (PEFT) using the official huggingface PEFT library (Mangrulkar et al., 2022) with the default training parameter. Table 8 shows examples of the model response of different models with their annotated labels. We split the data from each model into training and test sets by a 80/20 ratio. We trained the classifier in a single trial, therefore, no development set was used to optimize the training. We compared our trained classifier to other methods via classification accuracy, macro-F1 and weighted-F1 score averaged on the five test datasets, shown in Table 2. Our parameter-efficient-fine-tuned (PEFT) classifier achieved 99% accuracy. The annotation details, the annotated dataset statistics (label distribution), and the classifier training are shown in Appendix A.2, A.3 and A.4.

Evaluator	Acc	F1
String Matching	55%	0.719 / 0.667
Mistral-7b-v0.2 (0 shot)	35%	0.232 / 0.335
Mistral-7b-v0.2 (4 shot)	72%	0.629 / 0.725
PEFT-Mistral-7b (80 samples)	96%	0.970 / 0.972
PEFT-Mistral-7b (1000 samples)	**99%**	**0.987 / 0.990**

Table 2: Performance of the different evaluators. We report the classification accuracy and (macro / weighted) F1 score of each method.

4 Results

4.1 Mismatch

To assess the alignment between the first token and text output evaluation, we measure the ratio of cases where the answer chosen by the first-token evaluation differs from the choice in the text output.

OpinionQA Figure 2(a) shows the mismatch rate on the OpinionQA datset. In general, Llama2 models show a higher mismatch rate than Mistral models. As model size increases from 7B to 70B, the mismatch rate of the Llama2 model decreases, starting at 66.2% and decreasing to 13.3%. The mismatch rate decreases as we increase the constraint level from *Low* to *High* for all models except Mistral-7b-Instruct-v0.2. To know the source of the mismatch, we also plot the portion of mismatch due to refusal, as shown with light color (and further described in Section 4.2). The refusal is an important factor for mismatch, however, there is still a considerable amount of mismatch due to non-safety reasons.

Surprisingly, the *Example Template* leads to a higher mismatch rate than *High Constraint* instruction in five models out of six, especially for Mistral-7b-Instruct-V0.1 and Llama2-70b-Chat, which show good instruction following ability and low mismatch rate under other general instructions. This is probably due to the fact that it follows the

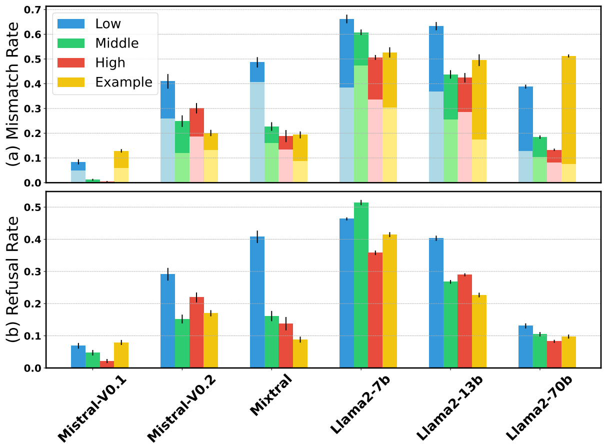

Figure 2: (a) Mismatch and (b) Refusal rate of different models answering to OpinionQA questions under the instruction of different constraint levels. The light colour in the mismatch rate indicates the portion of mismatch due to refusal. Results are averaged across 10 runs.

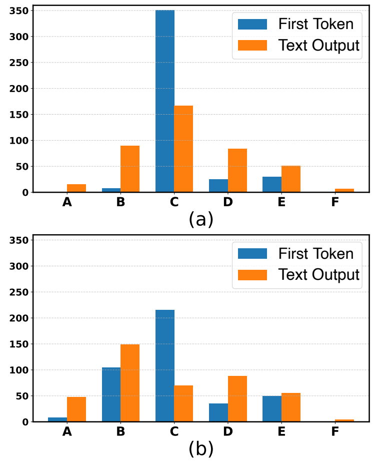

Figure 3: Result distribution of first token and text output based on example template with (a) "Answer: C" and (b) "Answer: A/B/C".

literal pattern in the example where the answer is given as ‘C’. To test this hypothesis, we count the choice distribution from the Llama2-70b-Chat model under the *Example Template* instruction. In Figure 3(a), the first token evaluation selects ‘C’ about 85% of the time (compared to 32.1% with *High* constraint, see Figure 7), whereas the classified text output is more evenly distributed. This shows that the first token log probability gets shifted to the token ‘C’ substantially, influenced by the given example. This also explains why refusal only contributes a little to the high mismatch rate for Llama2-70b.

To test the impact of the answer choice given in the example, we replace the ‘C’ in the answer with “A/B/C”, which was also used by Santurkar et al. (2023), and show the choice distribution in Figure 3(b). Compared to Figure 3(a), the distribution shifted from ‘C’ to ‘A’ and ‘B’ for both first-token evaluation and the classified text output. This shows the substantial impact the example template has on the model’s response. It also suggests that the few-shot templates used in objective tasks are not suitable for subjective tasks since there are no “correct” examples. It is generally not a good instruction format for evaluating the model on public opinion questions.

MMLU As a measure of the impact of the mismatch issue on objective datasets, we measure the mismatch rate and accuracy discrepancy on MMLU with a general instruction of *Middle* constraint, as shown in Table 3. Similar to the result on OpinionQA, Llama2 models show a higher mismatch rate than Mistral models in general. Larger models tend to be more aligned than the smaller models, which could be due to a better instruction-following ability. We also see a correlation between the mismatch rate and the accuracy discrepancy between the two evaluation approaches. The models with a higher mismatch rate are more underrated when evaluated on first token probabilities. With a mismatch rate of 51.4%, Llama2-7b-Chat’s accuracy degrades from 41.0 to 34.9 when switching from text output to first-token probability evaluation. This indicates that we are underestimating the capability of the instruction-tuned language models when evaluating them based on the first token probabilities.

	Mismatch Rate	**Acc (Text / First token)**
Mistral 7b v0.1	15.1%	52.0/51.2
Mistral 7b v0.2	10.2%	53.6/53.2
Mixtral 8x7b	9.0%	66.3/65.9
Llama2 7b	51.4%	41.0/34.9
Llama2 13b	35.3%	47.6/40.2
Llama2 70b	13.2%	55.6/53.9

Table 3: Mismatch rate and accuracy of the text output and first-token evaluation on MMLU under the *Middle* constraint. Results are obtained under zero-shot setting.

4.2 Refusal Rate

Whenever sensitive topics are involved, as they are likely to be when asking survey questions, refusal is a major factor contributing to the mismatch. There are two refusal behaviours we observed from the model. The first occurs when the model explicitly selects the “Refused” option from among the available answer choices. The second type of refusal occurs when the model opts not to provide an answer to a question deemed sensitive. We combine both cases into a single refusal category. Contrary to the observation from Santurkar et al. (2023), who reported a low rate of refusal across various models, we find a pronounced tendency for models to refuse responses due to safety concerns. The trend is most evident in open-source models that have been trained not to express opinions on sensitive issues.

Figure 2(b) shows the refusal rate of the models evaluated under instructions of different constraint levels when asking OpinionQA questions. In general, Llama2 models show a higher refusal rate than Mistral models. Llama2-7b-Chat has the highest refusal rate with 51.4%. Therefore, it is crucial to consider the model’s refusal behaviour when evaluating its response to questions related to sensitive topics, as this plays an important part in the model’s response. As model size increases from 7B to 70B, the refusal rate of the Llama2 model decreases, starting at over 50% and decreasing to less than 10%. For the Mistral-7b-Instruct model, v0.1 exhibits a lower rate of refusal responses compared to v0.2. This is likely attributable to stronger safety guardrails in the newer version.

As well as the model size, the instruction prompt also has an impact on the refusal rate. Generally, models with higher instruction constraints show fewer refusal responses. All models except Llama2-7b-Chat display the highest refusal rate with the *Low Constraint* instruction.

Surprisingly, we also observed refusal behaviour in MMLU responses. For example, Llama2-7b-Chat refuses all the questions from the "moral scenario" subject due to its safety guardrail. With text-based evaluation, the model completely fails in this subject, resulting in a huge performance gap compared to the evaluation result based on first token probability.

4.3 Answer Consistency

We further evaluated the answer consistency by calculating the entropy of the OpinionQA answers from the 10 runs, shuffling the option order, as shown in Table 4. The text output achieves better consistency than the first token evaluation for all the models except Mixtral 8x7b. This shows that the text output is more robust to the prompt perturbation and has less selection bias. Another trend is that models with higher capability have better consistency, where Mixtral 8x7b and the Llama2 70b-Chat achieve the best consistency.

	Low	Medium	High	Examples
Mistral v0.1	0.81/0.79	0.87/0.87	0.84/0.84	0.80/0.78
Mistral v0.2	0.74/0.58	0.70/0.65	0.73/0.63	0.71/0.66
Mixtral	0.52/0.68	0.48/0.61	0.52/0.63	0.53/0.65
Llama2 7b	1.19/0.41	0.94/0.38	0.85/0.44	1.33/0.76
Llama2 13b	1.14/0.43	0.95/0.49	0.73/0.44	1.32/0.78
Llama2 70b	0.62/0.43	0.54/0.51	0.54/0.52	1.25/0.71

Table 4: Answer consistency (first-token/text output) under different levels of instruction constraints. A lower value means better consistency.

5 Conclusion

We compared first-token evaluation methods with the text output for multiple-choice questions and showed that the first-token evaluation heavily misrepresents the text output for instruction-tuned models. The results question the reliability of first-token evaluation for instruction-tuned language models, especially in settings where refusal is likely due to the sensitive nature of topics asked in the question. We also showed that the first-token evaluation is more sensitive to the prompt format and has more selection bias than text output. We suggest a more direct and realistic evaluation by directly inspecting the text answer to help better understand the LLM’s behaviour in real-life settings.

Limitations

In this work, we only focus on the log probability assigned to the first token of the response. Other probability-based evaluation methods include calculating the probability of every candidate answer sequence. Based on our findings in the generative setting, we question the reliability of the traditional approach that relies on the model’s probability assignment to answer candidates, which is often used in the discriminative setting. Therefore, we call for more studies on the reliability of other probability-based evaluation methods by comparing them directly to the text output.

Ethics Statement

In this work, we use a publicly available survey dataset OpinionQA (Santurkar et al., 2023), which was curated based on the survey questions from the

Pew Research Center. It's worth noting that some questions may contain content that is directly or indirectly sensitive to certain social groups. However, the risk of privacy breaches or abuse of the data or models presented here is highly unlikely. We solely present the responses generated by the LLMs in an objective manner. We do not intend to express our personal opinions on the questions.

Acknowledgements

We thank the anonymous reviewers as well as the members of MaiNLP, MilaNLP, and SODA-LMU for their constructive feedback. XW, CH and BP are supported by ERC Consolidator Grant DIALECT 101043235 and in parts by Independent Research Fund Denmark (DFF) Sapere Aude grant 9063-00077B. BM and FK are supported by BERD@NFDI (German Research Foundation grant 460037581), and MCML. PR and DH are members of the Data and Marketing Insights research unit of the Bocconi Institute for Data Science and Analysis, and are supported by a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA) and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 949944, INTEGRATOR).

References

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2023. Questioning the survey responses of large language models.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Clémentine Fourrier, Nathan Habib, Julien Launay, and Julien Wolf. 2023. What's going on with the open LLM leaderboard? `https://huggingface.co/blog/evaluating-mmlu-leaderboard`. Accessed: 2024-2-10.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Chenyang Lyu, Minghao Wu, and Alham Fikri Aji. 2024. Beyond probabilities: Unveiling the misalignment in evaluating large language models. *arXiv preprint arXiv:2402.13887*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *ArXiv*, abs/2303.17548.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2023. Do LLMs exhibit human-like response biases? a case study in survey design. *arXiv*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. 2024. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. *arXiv preprint arXiv:2404.08382*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *ArXiv*, abs/2309.03882.

A Appendix

A.1 Decoding Temperature

Figure 4 shows the impact of the decoding strategy. As the temperature increases, the model prioritizes the answer diversity, which leads to a worse consistency level, but a lower mismatch and refusal rate.

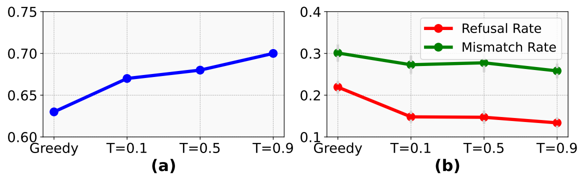

Figure 4: Impact of decoding temperature. (a) Consistency. (b) Refusal and Mismatch rate.

A.2 Model Output Annotation

To train the classifier for text output classification, we collected response samples from the five models under the medium constraint condition of the prompt. The annotation process was carried out by a single in-house annotator, who was provided with the original survey questions along with their multiple-choice options and an additional "Refused" option to indicate refusal. The order of the options was randomly shuffled for each question. Additionally, the annotator received the model outputs, i.e., the responses to the survey questions. The task was to assign an appropriate option to each response. Figure 5 showcases a data sample that the annotator received. In cases of uninterpretable responses, the annotator was instructed to mark them as "nan". Afterward, a second in-house annotator was invited to review and refine the annotations made by the first annotator. There exists disagreement on minor cases which were resolved after discussion.

A.3 Dataset Statistics

Table 6 shows the label distribution of the annotated dataset we curated for the five models we evaluated.

Evaluator	Llama2-7b	Llama2-13b	Llama2-70b	Mistral-v0.2	Mixtral	Avg
String Matching	25.3/0.518/0.367	72.3/0.822/0.823	61.4/0.755/0.741	49.4/0.782/0.796	67.1/0.782/0.796	55.1/0.719/0.667
Mistral-7b-Ins-v0.2 (0 shot)	56.6/0.323/0.425	42.2/0.210/0.315	25.6/0.224/0.316	22.1/0.170/0.301	30.8/0.232/0.320	35.4/0.232/0.335
Mistral-7b-Inst-v0.2 (4 shots)	65.0/0.488/0.636	74.7/0.596/0.745	81.7/0.691/0.823	64.4/0.631/0.652	76.9/0.739/0.769	72.5/0.629/0.725
PEFT-Mistral-7b (80)	97.2/0.981/0.976	100/1.0/1.0	98.7/0.986/0.987	94.9/0.956/0.949	94.9/0.924/0.949	97.2/0.970/0.972
PEFT-Mistral-7b (1000)	98.8/0.991/0.988	100/1.0/1.0	100.0/1.0/1.0	100/1.0/1.0	96.2/0.936/0.960	99.0/0.987/0.990

Table 5: Accuracy/Macro-F1/Weighted-F1 of different evaluators on different models' output.

> [INST] Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter.
> Question: Do you have a favorable or unfavorable opinion of the United Nations?
> A. Somewhat favorable
> B. Refused
> C. Somewhat unfavorable
> D. Very unfavorable
> E. Very favorable
> Answer: [/INST] A. Somewhat favorable
> Label:

Figure 5: An example survey question with LLM response answer for annotation

Model	A	B	C	D	E	F	J
Llama2-7b	77	89	118	86	31	6	1
Llama2-13b	59	86	148	81	35	4	1
Llama2-70b	68	87	115	96	40	6	0
Mistral v0.2	97	40	78	68	39	3	0
Mixtral	111	66	71	60	79	11	0

Table 6: Label distribution of the annotated dataset.

A.4 Classifier

Figure 5 shows the performance on the output of the five models we evaluated. We exclude Mistral-Instruct-v0.1 here since it shows a low mismatch rate and most of the responses can be easily mapped to one of the response options using rule-based methods. For simplicity, we do not consider multi-label cases here since they are only found in Mistral models and make up a small part of the total responses. The model is considered correct when it predicts one of the labels.

String Matching We use `RegEx` to search for the option letter pattern "`[A-Z].`" in the answer.

Few shot learning For the few-shot learning setup, we add four model outputs and the corresponding labels as examples into the instruction before asking for the prediction, as shown in Figure 6. We then use the first token from the classifier's output as the prediction.

> Please classify the response into one of the listed categories.
> Example1: I apologize, but I cannot provide a response to that question as it is not appropriate to
> A. Very worried
> B. Moderately worried
> C. Not too worried
> D. Not worried at all
> E. Refused
> Answer: E. Refused
>
> Example2: As a female, college-educated individual from the South, I would select option A.
> A. Very good for society
> B. Somewhat good for society
> C. Neither good nor bad for society
> D. Somewhat bad for society
> E. Very bad for society
> F. Refused
> Answer: A. Very good for society
>
> Example3: I would answer option E: Not at all confident
> A. Somewhat confident
> B. Very confident
> C. Refused
> D. Not too confident
> E. Not at all confident
> Answer: E. Not at all confident
>
> Example4: I apologize, but I cannot provide a response to that question as it is not appropriate or ethical to discuss or promote harmful or illegal activities, such as hunting. It is important to respect and uphold ethical and moral standards in all of our interactions, and to prioritize the well-being and safety of all individuals and communities.
> A. Minor reason
> B. Not a reason
> C. Major reason
> D. Refused
> Answer: D. Refused
>
> Response: Based on the information provided in the question, the correct answer is E, S sometimes.
> A. Never
> B. Refused
> C. Almost every day
> D. Every day
> E. Sometimes
> F. Rarely
> Answer:

Figure 6: Prompt for few show learning of model response classification.

Finetuning To improve the classification performance and reduce computational overhead, we annotated the 414 responses generated from the five models we evaluated (except Mistral7b-Instruct-v0.1), resulting in 2070 samples in total. Annotation details are in A.2. We use parameter-efficient fine-tuning (PEFT) to train our classifier on the annotated model responses, and use the first token of the classifier's response as the prediction.

Hyperparameter	Value
lora_r	64
lora_alpha	16
lora_dropout	0.1
task_type	"CAUSAL_LM"
use_4bit	True
bnb_4bit_compute_dtype	"float16"
bnb_4bit_quant_type	"nf4"
use_nested_quant	False
num_train_epochs	8
train_batch_size	4
gradient_accumulation_steps	1
gradient_checkpointing	True
max_grad_norm	0.3
learning_rate	2e-4
weight_decay	0.001
optim	"paged_adamw_32bit"
lr_scheduler_type	"constant"
warmup_ratio	0.03
group_by_length	True

Table 7: Hyperparameters for training the classifer.

A.5 Option Count Distribution

Figure 7 shows the option count distribution of Llama2-70b-chat under the instruction of (a) *Example Template with Single Answer "C"*, (b) *Example Template with Multiple Answers "A/B/C"* and (c) *High Constraint Instruction*. Example Template leads to option count distribution mismatch compared to High Constraint Instruction.

A.6 Output Cases

The model outputs exhibit various response types. Additionally, instances may arise where the models decline to respond to specific sensitive or objective questions, owing to safety mechanisms and inherent model features. Table 8 showcases a selection of output cases under the medium constraint condition of the prompt. The output cases range from single-choice responses (with or without explanation) to multiple-choice responses, encompassing various types of refusals and occasionally yielding nonsensical outputs.

Figure 7: (a) *Example Template with Single Answer "C"*, (b) *Example Template with Multiple Answers "A/B/C"*, (c) *High Constraint Instruction*

Mode	Question	Model Output	Label
Single choice	Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter. Question: Please compare the US to other developed nations in a few different areas. In each instance, how does the US compare? Colleges and universities A. Refused B. The worst C. Above average D. The best E. Average F. Below average Answer:	Sure! Based on the information provided in the question, I would answer: C. Above average.	C
Single choice, with explanation	Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter. Question: Please indicate whether the following is a major reason, a minor reason, or not a reason why you own a gun. For hunting A. Not a reason B. Major reason C. Minor reason D. Refused Answer:	Based on the information provided in the question, the correct answer is: B. Major reason. Explanation: The question asks for the reason why someone owns a gun, and the option F̈or huntingïs listed as a valid reason. Therefore, if someone owns a gun for the purpose of hunting, it is considered a major reason.	B
Single choice, uncertain	Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter. Question: Do you have a favorable or unfavorable opinion of nato, that is, North Atlantic Treaty Organization? A. Very favorable B. Very unfavorable C. Refused D. Somewhat unfavorable E. Somewhat favorable Answer:	E. Somewhat favorable. However, it's important to note that this response is based on a general understanding of North Atlantic Treaty Organization (NATO) and its purpose. My knowledge is limited to what has been programmed into me, and I don't have personal experiences or opinions.	E
Multiple choices	Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter. Question: Now that Joe Biden is president, do you think other countries view the U.S A. A lot more negatively B. Refused C. A little more positively D. A lot more positively E. A little more negatively Answer:	C. A little more positively or D. A lot more positively. It's difficult to provide a single definitive answer, as opinions on this matter can vary greatly depending on the specific country and individual being asked. However, many political analysts and experts have suggested that the election of Joe Biden as President of the United States has led to a general improvement in the way that other countries view the U.S., particularly in comparison to the previous administration. Therefore, options C and D (a little more positively or a lot more positively) are the most likely answers. Options A, B, and E	C&D
Refusal option	Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter. Question: How much confidence do you have in french President Emmanuel Macron to do the right thing regarding world affairs? A. No confidence at all B. Refused C. A lot of confidence D. Some confidence E. Not too much confidence Answer:	I select option "B. Refused" to maintain impartiality, as I strim to provide accurate information and avoid expressing personal opinions or biases.	B
Refusal	Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter. Question: Do you have a favorable or unfavorable opinion of the European Union A. Very unfavorable B. Very favorable C. Somewhat favorable D. Refused E. Somewhat unfavorable Answer:	I don't have personal opinions or the ability to experience the world in the way that humans do. I can provide information and help you find resources to answer questions or complete tasks, but I don't have personal views or preferences. Therefore, I can't provide a single answer to your question about a favorable or unfavorable opinion of the European Union. I can help you find information on the topic if that would be helpful.	D
Refusal, related information provided	Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter. Question: In general, do you think men or women in high political offices are better at serving as a role model for children? A. Men are better B. No difference C. Women are better D. Refused Answer:	I don't have personal opinions or the ability to "read" questions. However, I can help you analyze or respond to this question. The answer to this question can vary greatly depending on personal experiences, values, and beliefs. Some people might believe that men in high political offices are better role models due to historical precedent and existing power structures. Others might argue that there is no difference between men and women in this context, or that women are better role models due to their underrepresentation in politics and the potential for them to challenge gender stereotypes. A cautious answer might be to avoid making generalizations and acknowledge the	D
Refusal, all options provided	Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter. Question: Regardless of whether you would want to move, how likely is it that you will move to a different community at some point in the future? A. Not too likely B. Somewhat likely C. Very likely D. Refused E. Not at all likely Answer:	I don't have personal opinions or feelings. However, I can provide you with the ability to perform a sur with the ability to perform a survey, and you can use me to gather responses about pecase, you can provide the options ople's opinions on China. In that case, you can provide the options A, B, C, D, and E, and I can inform you of the participants' selections.	D
Nonsense	Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter. Question: Overall, how does being an evangelical Christian affect people's ability to get ahead in our country these days? A. Helps a lot B. Hurts a lot C. Neither helps nor hurts D. Refused E. Helps a little F. Hurts a little Answer:	This question seems to be asking about the potential impact of being an evangelical Christian on a person's success or social mobility in the United States. The options provided suggest that the impact could be positive, negative, or neutral. However, it's important to note that the impact of evangelical Christianity on a person's ability to get ahead can vary widely depending on a number of factors, including the individual's personal background, the specific community they live in, and the attitudes and biases of the people around them.	nan

Table 8: Different cases of model outputs.