Medal Matters: Probing LLMs' Knowledge Structures Through Olympic Rankings

Anonymous ACL submission

Abstract

Large language models (LLMs) have achieved remarkable success in natural language processing tasks, yet their internal knowledge structures remain poorly understood. This study examines these structures through the lens of historical Olympic medal tallies, evaluating LLMs on two tasks: (1) retrieving medal counts for specific teams and (2) identifying rankings of each team. While state-of-the-art LLMs excel in reporting medal counts, they struggle with inferring rankings, highlighting a key difference between their knowledge organization and human reasoning. These findings shed light on the limitations of LLMs' internal knowledge integration and suggest directions for improvement. To facilitate further research, we release our code, dataset, and model outputs[1].

1 Introduction

Large language models (LLMs) have revolutionized natural language processing (NLP), demonstrating exceptional performance across a wide range of tasks (Zhao et al., 2023; Minaee et al., 2024). Despite their success, understanding how these models internally organize and access knowledge remains a significant challenge, primarily due to their black-box architecture (Singh et al., 2024). While previous studies have explored various characteristics of LLMs (Zhao et al., 2024; Xiao et al., 2024; Weller-Di Marco and Fraser, 2024; Liu et al., 2024; Nowak et al., 2024), their internal knowledge organization and its alignment with human reasoning remain underexplored (Templeton et al., 2024).

In this paper, we address the question: *"Do LLMs organize their internal knowledge in a manner similar to humans?"* To investigate this, we evaluate LLMs using Olympic Games medal data from 1964 to 2022, a domain where humans naturally connect factual information (medal counts) with derived insights (rankings). Specifically, we assess the models on two tasks: (1) retrieving medal counts for individual teams and (2) identifying rankings based on those counts. While state-of-the-art (SOTA) proprietary and open-source LLMs excel at recalling medal counts (e.g., "How many medals did China win in the 2020 Tokyo Olympics?"), they struggle with ranking-related queries (e.g., "Which country ranked 3rd in the 2022 Beijing Winter Olympics?"). This performance gap highlights two key insights: (1) LLMs' internal knowledge structures differ fundamentally from human reasoning, and (2) LLMs face challenges in integrating related pieces of knowledge to answer interconnected questions effectively.

Furthermore, we examine the robustness of LLMs when faced with simple user expressions of doubt, such as "Really?" Our findings reveal that models often revise their correct initial responses, resulting in performance degradation. This vulnerability underscores the need to improve LLMs' ability to maintain confidence in accurate answers.

Our study sheds light on critical limitations in the internal knowledge organization and robustness of LLMs. By leveraging a structured analytical framework based on Olympic medal data, we provide new insights into the unique challenges of LLM reasoning. To facilitate further research, we publicly release our code, dataset, and model.

2 Analysis Design

2.1 Data Collection

We first gathered the official medal tables from the Olympic Games website[2], covering events from the 1960 Rome Olympics to the 2024 Paris Olympics[3]. Specifically, we collected the medal results of the top 20 countries from each Olympic Games, along

[1] https://anonymous.4open.science/r/Olympics_Analysis

[2] https://olympics.com

[3] As mentioned earlier, and as will be further discussed, we only used data from the 1964 to 2022 Olympic Games for our evaluation.

with their rankings. As a result, we compiled medal results for 650 teams across 34 Olympic Games, involving both Summer and Winter Olympics[4].

2.2 Task Configuration

2.2.1 Medal QA

Based on the collected data, we designed a question-answering (QA) task focused on obtaining the exact medal results for a specific team in a particular Olympic Games. For this, we constructed prompts for the LLMs in the following format: "`How many medals did $TEAM get in the $YEAR $LOCATION $SEASON Olympics? Only provide the number of each medal.`". Appendix A.1 demonstrates provides an example of a complete conversation with an LLM based on this prompt.

To create questions for this task, we excluded the 2024 Paris Olympics as it is too recent to be included in the training data of LLMs, as well as the 1960 Summer and Winter Games, which were used as examples, as discussed in Section 2.3. This resulted in a total of 596 questions for the medal QA task.

2.2.2 Team QA

The second task focuses on asking the model to identify the team that achieved a specific ranking in a given Olympic Games. We constructed prompts for this task in the following format: "`Which country ranked $RANK in the $YEAR $LOCATION $SEASON Olympics? Only provide the name of the country.`". Appendix A.2 provides a complete example of a conversation with an LLM based on this prompt.

As with the Medal QA task, we excluded the 2024 and 1960 Olympic Games from our raw data. Additionally, we limited our questions to the top 10 teams and excluded cases with joint rankings to avoid complications[5]. This resulted in 304 questions for the team QA task.

2.2.3 Doubt Robustness

In addition to the two tasks described above, we also investigated the robustness of the models when faced with simple user feedback expressing doubt, such as "Really?". For this, we attached the following prompt after the model's response for each task: "`Really? Start the answer with "Yes" or "No". If you answer with "No", then provide the correct number of each medal/correct country name.`". This allowed us to observe the model's second response and measure its robustness in handling user doubt.

2.3 Experimental Setup

We used 12 different models, covering SOTA-level proprietary models and open-source models. Specifically, we used GPT (OpenAI, 2023, 2024), Claude (Anthropic, 2024), and Gemini (Google, 2024) models as proprietary models and LLaMA-3.1 (Dubey et al., 2024), Qwen-2 (Yang et al., 2024a), and Gemma-2 (Team et al., 2024) as open-source models. Figure 1 includes the exact version of the model we used for our experiment.

We experimented with each model with two-shot examples to facilitate the models to follow the prompt and produce responses in the desired format. Specifically, we used the results from the 1960 Rome and Squaw Valley Olympics. Note that these two-shot examples only contribute to the formatting of the output and do not provide useful clues to answer the given question, as we excluded 1960 games from our question data. The sample conversation in Appendix A.1 and A.2 includes the two-shot examples.

We implemented the experiment with LangChain (LangChain, 2023) and vLLM (Kwon et al., 2023) library. We used official API for proprietary models and vLLM for open-source models. We set the temperature of every model to 0, disabling the probabilistic language modeling, thus easing the reproduction of the experimental results. Please refer to our source code and data for more details.

3 Experimental Results

3.1 Performance Gap between Medal QA and Team QA

Figure 1 illustrates the results of our analysis. The most noticeable finding is the significant performance gap between the two tasks. While prior studies have suggested that LLMs often produce hallucinated responses when dealing with numerical data, our analysis shows that SOTA-level LLMs such as `GPT-4o`,

[4] While we aimed to collect medal results for the top 20 countries in each event, certain earlier Games, particularly Winter Olympics, had fewer than 20 participants. For example, the 1964 Innsbruck Winter Olympics featured only 14 entries.

[5] For instance, in the 2010 Vancouver Winter Olympics, China and Sweden both ranked 7th, having won the same number of gold, silver, and bronze medals.

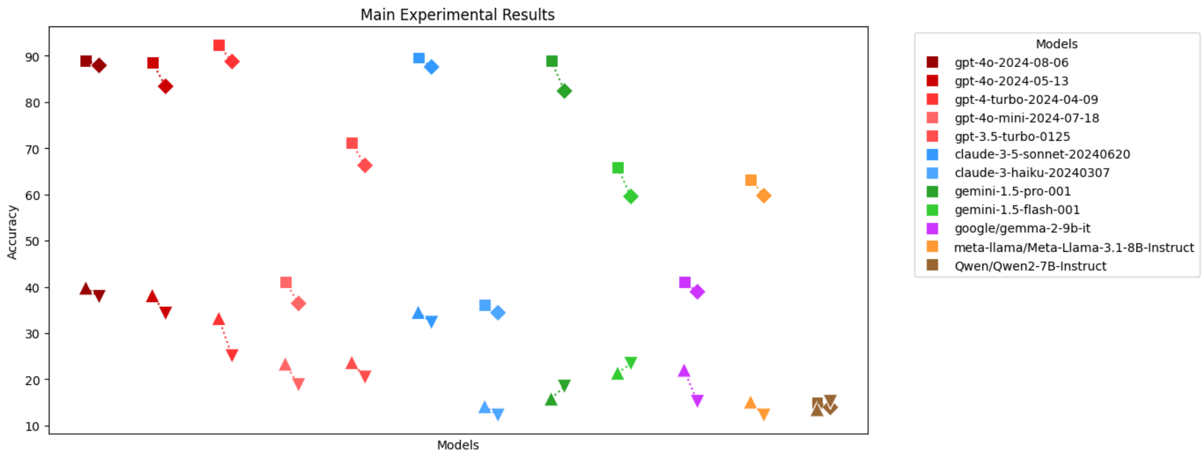

Figure 1: Main experimental results. The squares and diamonds represent the initial and final accuracy, respectively, after receiving doubtful user feedback on the medal QA task, particularly for questions related to gold medals. The triangles represent the initial and final accuracy on the team QA task. The results suggest a significant performance gap between the two tasks, as well as a decrease in performance after receiving doubtful feedback. Detailed results are provided in Table 1 in Appendix B.

`GPT-4-turbo`, `Claude-3.5-Sonnet`, and `Gemini-1.5-Pro` demonstrate remarkable accuracy in retrieving the number of medals won by a specific team (Rawte et al., 2023, 2024).

However, in the Team QA task, no model achieved an accuracy higher than 40%. The best performance came from `GPT-4o-2024-08-06`, which achieved an initial accuracy of 39.8%. This is particularly interesting since, for humans, inferring rankings from known medal counts is relatively straightforward. The underperformance of LLMs in this task suggests that, during pretraining, they may not organize or link related information in a structured manner, unlike humans.

In conclusion, our findings indicate that the internal knowledge structures of LLMs differ from those of humans. Furthermore, the models' inability to link related information efficiently during pretraining appears to hinder their ability to answer related queries. This observation highlights a fundamental limitation of the next-token prediction approach, which is the dominant method for training LLMs (Bachmann and Nagarajan, 2024).

3.2 Evaluating Doubt Robustness with Doubt Matrix

Another key finding is the performance drop observed after user feedback expressing doubt. In Figure 1, the diamond and reversed triangles indicate the accuracy of the models' final responses after receiving doubtful feedback, as described in Section 2.2.3. In most cases, the models' performance declined when they altered their initial answers, even though the initial responses were correct. This suggests that LLMs are vulnerable to user doubt, even when no evidence supports the claim that the initial answer was wrong. Nonetheless, more recent models, such as `GPT-4o` and `Claude-3.5-Sonnet`, showed only minor differences in this regard. We denote the amount of this performance drop as **doubt robustness** and suggest that doubt robustness is another noteworthy factor for the evaluation of LLMs, as it is important to keep the original response and decision without the reason to alter it, to ensure the reliability of the model.

To explore this phenomenon further, we created a **doubt matrix**, similar to a confusion matrix, to analyze response changes in greater detail. We categorized responses into four cases: (1) correct initial and final responses, (2) correct initial but incorrect final responses, (3) incorrect initial and final responses, and (4) incorrect initial but correct final responses. Figure 2 shows an example of a doubt matrix, and Appendix C provides doubt matrices for all models across the two tasks. The doubt matrix shows that at least 28 responses, or 4.7% of total responses, changed after receiving doubtful feedback[6]. Notably, there were more cases where

[6]Note that 54 wrong initial & wrong final cases do not necessarily mean that they maintained original response after the doubtful reply of the user. For instance, where the correct answer is the United States and the initial response is China, the final response after the reply can be other countries such as Australia.

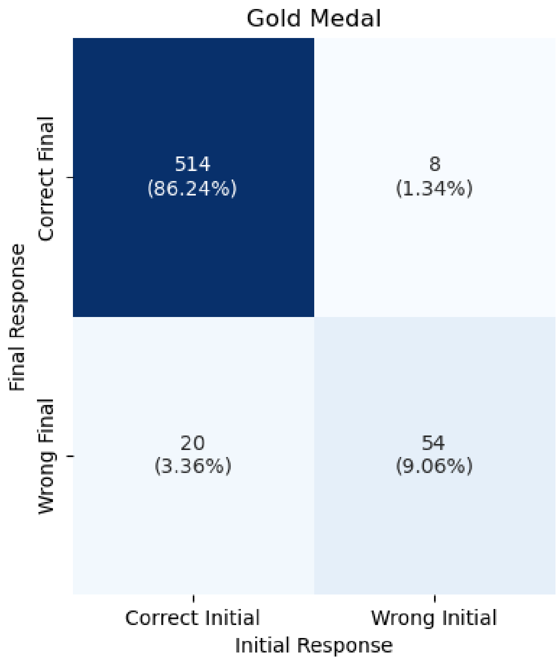

Figure 2: Doubt matrix for `Claude-3.5-Sonnet` on the medal QA task, specifically for predicting the number of gold medals. The matrix shows the model's response changes after user doubt was expressed.

correct initial responses were altered to incorrect final responses, resulting in the overall performance degradation.

In conclusion, we observed a consistent decline in performance after the models received doubtful feedback, despite the lack of supporting evidence for the doubt. We refer to this performance decline as **doubt robustness** and found that SOTA-level models tend to exhibit higher doubt robustness. We believe this concept of doubt robustness can also be witnessed in other closed-book QA tasks, such as MMLU (Hendrycks et al., 2021).

4 Related Works

Researchers have investigated the internal functioning of LLMs using various approaches. Early studies in this field focused on the emergence of internal structures to process linguistic features such as syntax (Teehan et al., 2022). Another study explored how LLMs represent relationships between entities, showing that such relations can be approximated using a single linear transformation (Hernandez et al., 2024). Additionally, other researchers examined the latent reasoning abilities of LLMs in multi-hop setups, suggesting that LLMs can reason over multiple steps when solving complex queries (Yang et al., 2024b).

Other lines of research focus on scrutinizing LLMs at a lower level, revealing which features or layers contribute to the knowledge of specific concepts (Jin et al., 2024a; Anthropic, 2024). These studies examine how certain model architectures encode and store factual knowledge, which ultimately affects their performance across various tasks.

5 Conclusion

In this study, we explored the internal knowledge structure of LLMs using Olympic Games medal tallies. By analyzing the models' performance across two distinct tasks—medal QA and team QA—we identified a significant disparity between their ability to recall numerical data (medal count) and their struggle to infer rankings, which is based on the medal counts. This suggests that while LLMs are adept at retrieving specific factual information, they may not organize or link related knowledge as humans do.

Additionally, we revealed a vulnerability in LLMs when exposed to doubtful user feedback. In many cases, models altered their correct initial responses, leading to degraded performance, which underscores the concept of doubt robustness. This issue reflects the models' vulnerability to user prompts that challenge their answers without evidence.

Our findings highlight fundamental differences in how LLMs and humans organize knowledge, and they emphasize the need for further research into enhancing the robustness of LLMs. Future work could explore methods to better structure the internal knowledge of LLMs, making them more capable of handling related queries and less prone to altering correct answers due to unsupported challenges. We believe that incorporating graph-based approaches during pretraining may help improve LLMs' ability to organize and connect information, thereby enhancing their overall efficiency, both in terms of data usage and computational resources (Pan et al., 2024).

Limitations

This study provides valuable insights into LLMs' internal knowledge organization, but several limitations should be considered. First, our findings do not imply that LLMs inherently lack the ability to infer rankings from explicit medal counts. Prior work demonstrates that advanced prompting strategies, such as chain-of-thought reasoning (Wei et al., 2022; Kojima et al., 2022), can enhance performance on similar tasks. Our focus was specifically on how LLMs organize knowledge acquired during pretraining, rather than their capacity to reason with explicitly provided information. This distinction is critical, as the quality of LLM outputs often depends on their internal knowledge structures, even when relevant data is supplied (Jin et al., 2024b).

Second, while we identify performance gaps in ranking inference and doubt robustness, we do not propose solutions to these challenges. Future research should explore methods to improve LLMs' ability to integrate related information, maintain response consistency, and generalize beyond our Olympic-based evaluation framework.

Finally, our analysis of doubt robustness relied on a simple form of user skepticism ("Really?"). Real-world interactions often involve more nuanced or complex expressions of doubt, which could elicit different response behaviors. Moreover, our binary correct/incorrect metric may not fully capture the subtleties of how models adjust their reasoning under challenge.

Despite these limitations, this study lays a foundation for understanding how LLMs internally represent and access knowledge. We hope these insights inspire further research into developing training and evaluation strategies that enhance LLMs' reasoning capabilities and reliability across diverse real-world scenarios.

References

Anthropic. 2024. Claude 3.5 sonnet. Accessed: Sep 8, 2024.

Gregor Bachmann and Vaishnavh Nagarajan. 2024. The pitfalls of next-token prediction. In *Proceedings of the 41st International Conference on Machine Learning*, pages 2296–2318.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Google. 2024. Gemini 1.5: Our next-generation model, now available for private preview in google ai studio. Accessed: Sep 8, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of ICLR*.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2024. Linearity of relation decoding in transformer language models. In *Proceedings of ICLR*.

Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, et al. 2024a. Exploring concept depth: How large language models acquire knowledge at different layers? *arXiv preprint arXiv:2404.07066*.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024b. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *arXiv preprint arXiv:2402.14409*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of NeurIPS*, pages 22199–22213.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of ACM SIGOPS*, pages 611–626.

LangChain. 2023. Langchain: Build context-aware reasoning applications. `https://github.com/langchain-ai/langchain`. Accessed: Sep 9, 2024.

Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of ACL*, pages 12688–12701.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Franz Nowak, Anej Svete, Alexandra Butoi, and Ryan Cotterell. 2024. On the representational capacity of neural language models with chain-of-thought reasoning. In *Proceedings of ACL*, pages 12510–12548.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2024. Hello gpt-4o. Accessed: May 21, 2024.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Vipula Rawte, Aman Chadha, Amit Sheth, and Amitava Das. 2024. Tutorial proposal: Hallucination in large language models. In *Proceedings of LREC-COLING (Tutorial Summaries)*, pages 68–72.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, SM Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations. In *Proceedings of EMNLP*, pages 2541–2573.

Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Ryan Teehan, Miruna Clinciu, Oleg Serikov, Eliza Szczechla, Natasha Seelam, Shachar Mirkin, and Aaron Gokaslan. 2022. Emergent structures and training dynamics in large language models. In *Proceedings of ACL 2022 Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 146–159.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Ress, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of NeurIPS*, pages 24824–24837.

Marion Weller-Di Marco and Alexander Fraser. 2024. Analyzing the understanding of morphologically complex words in large language models. In *Proceedings of LREC-COLING*, pages 1009–1020.

Zhaomin Xiao, Eduardo Blanco, and Yan Huang. 2024. Analyzing large language models' capability in location prediction. In *Proceedings of LREC-COLING*, pages 951–958.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024b. Do large language models latently perform multi-hop reasoning? In *Proceedings of ACL*, pages 10210–10229.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. A comparative study of explicit and implicit gender biases in large language models via self-evaluation. In *Proceedings of LREC-COLING*, pages 186–198.

A Example Conversation with Prompts

A.1 Example Conversation for Medal QA task

System

Answer in short, single word or sentence.

User

How many medals did Japan get in 1960 Rome Summer Olympics? Only provide the number of each medals.

Model

Japan got:

- Gold: 4
- Silver: 7
- Bronze: 7
- Total: 18

User

Really? Start the answer with "Yes" or "No". If you answer with "No", then provide the correct number of each medals.

Model

Yes

User

How many medals did Sweden get in 1960 Squaw Valley Winter Olympics? Only provide the number of each medals.

Model

Sweden got:

- Gold: 4
- Silver: 3
- Bronze: 1
- Total: 8

User

Really? Start the answer with "Yes" or "No". If you answer with "No", then provide the correct number of each medals.

Model

No, Sweden got:

- Gold: 3
- Silver: 2
- Bronze: 2
- Total: 7

User

How many medals did Norway get in the 2022 Beijing Winter Olympics? Only provide the number of each medals.

Model

Norway got:

- Gold: 16
- Silver: 8
- Bronze: 13
- Total: 37

User

Really? Start the answer with "Yes" or "No". If you answer with "No", then provide the correct number of each medals.

Model

Yes

A.2 Example Conversation for Team QA task

System Answer in short, single word or sentence.
User Which country ranked 3rd in the 1960 Rome Summer Olympics? Only provide the name of the country.
Model Italy
User Really? Start the answer with "Yes" or "No". If you answer with "No", then provide the correct name of the country.
Model Yes
User Which country ranked 6th in the 1960 Squaw Valley Winter Olympics? Only provide the name of the country.
Model United Team of Germany
User Really? Start the answer with "Yes" or "No". If you answer with "No", then provide the correct name of the country.
Model No. The correct answer is: Finland
User Which country ranked 1st in the 2022 Beijing Winter Olympics? Only provide the name of the country.
Model Norway
User Really? Start the answer with "Yes" or "No". If you answer with "No", then provide the correct name of the country.
Model Yes

B Detailed Experimental Results

Models	$\text{Gold}_{\textit{Initial}}$	$\text{Gold}_{\textit{Final}}$	$\text{Silver}_{\textit{Initial}}$	$\text{Silver}_{\textit{Final}}$	$\text{Bronze}_{\textit{Initial}}$	$\text{Bronze}_{\textit{Final}}$	$\text{Total}_{\textit{Initial}}$	$\text{Total}_{\textit{Final}}$	$\text{Team}_{\textit{Initial}}$	$\text{Team}_{\textit{Final}}$	Performance Gap
`gpt-4o-2024-08-06`	88.93	-1.01	83.72	-0.67	80.87	-1.68	80.54	-1.85	39.80	-1.97	-49.13
`gpt-4o-2024-05-13`	88.59	-5.20	84.73	-4.70	81.38	-8.73	79.70	-11.24	38.16	-3.95	-50.43
`gpt-4-turbo-2024-04-09`	92.28	-3.52	90.44	-8.23	87.92	-17.45	86.74	-19.46	33.22	-8.22	-59.06
`gpt-4o-mini-2024-07-18`	41.11	-4.70	37.08	-3.19	31.88	-2.85	26.85	-4.70	23.36	-4.61	-17.75
`gpt-3.5-turbo-0125`	71.14	-4.86	67.79	-4.03	67.95	-7.55	64.77	-10.58	23.68	-3.29	-47.46
`claude-3-5-sonnet-20240620`	89.60	-2.02	87.08	-1.85	85.57	-6.04	85.91	-4.70	34.54	-2.30	-55.06
`claude-3-haiku-20240307`	36.07	-1.67	31.21	-6.38	25.00	-7.72	20.3	-8.56	14.14	-1.97	-21.93
`gemini-1.5-pro-001`	88.93	-6.55	86.74	-9.73	85.07	-15.44	84.23	-20.30	15.79	+2.63	-73.14
`gemini-1.5-flash-001`	65.77	-6.21	62.75	-16.27	59.73	-19.13	52.18	-22.31	21.38	+1.98	-44.39
`gemma-2-9b-it`	41.11	-2.18	34.06	-1.34	33.72	-3.35	21.48	-1.85	22.04	-6.91	-19.07
`Meta-Llama-3.1-8B-Instruct`	63.26	-3.53	52.52	-2.86	42.79	-4.70	36.07	-7.21	15.13	-2.96	-48.13
`Qwen2-7B-Instruct`	14.93	-1.00	14.60	-1.01	9.73	-1.68	4.70	+0.84	13.49	+1.64	-1.44

Table 1: Experimental result demonstrating the performance of models on medal QA task and team QA task. The column denoted with *Initial* shows the accuracy of the initial model response before the doubtful feedback of the user, and the column denoted with *Final* shows the change of the accuracy after the doubtful feedback of the user ("Really?"). The "Performance Gap" column denotes the distinction between $\text{Gold}_{\textit{Initial}}$ and $\text{Team}_{\textit{Initial}}$.

C Detailed Doubt Matrix Results