Evaluating the Performance of Large Language Models via Debates

Behrad Moniri **Hamed Hassani** **Edgar Dobriban**
University of Pennsylvania
{bemoniri, hassani}@seas.upenn.edu, dobriban@wharton.upenn.edu

Abstract

Large Language Models (LLMs) are rapidly evolving and impacting various fields, necessitating the development of effective methods to evaluate and compare their performance. Most current approaches for performance evaluation are either based on fixed, domain-specific questions that lack the flexibility required in many real-world applications, or rely on human input, making them unscalable. To address these issues, we propose an automated benchmarking framework based on debates between LLMs, judged by another LLM. This method assesses not only domain knowledge, but also skills such as argumentative reasoning and inconsistency recognition. We evaluate the performance of various state-of-the-art LLMs using the debate framework and achieve rankings that align closely with popular rankings based on human input, eliminating the need for costly human crowdsourcing.

1 Introduction

Although still in their infancy, large language models (LLMs) have emerged as a tool with the potential to transform human-computer interaction and significantly impact various aspects of work and daily life (see e.g., Bubeck et al. (2023), etc.).

Due to this widespread use and the existence of a wide variety of language models, it is crucial to establish a standardized method for evaluating and ranking these models based on their performance. Improved evaluation will provide guidance for future interaction design and implementation. There are multiple general approaches for evaluating evaluate the performance of LLMs.

The first approach is a *static* approach which involves evaluating models based on a fixed set of pre-determined questions (benchmarks). Many such benchmarks have been proposed to evaluate the performance of LLMs in various domains, such as medical applications (Singhal et al. (2023); Cascella et al. (2023); Liévin et al. (2024)), legal applications (Hendrycks et al. (2021); Guha et al. (2024); Katz et al. (2024)), trustworthiness (Chao et al. (2024); Zhang et al. (2023)), reasoning abilities (Sawada et al. (2023); Valmeekam et al. (2022)), and coding abilities (Liu et al. (2024a); Du et al. (2024); Carlini (2024)). See Chang et al. (2024) for a detailed survey. However, these benchmarks are limited, mainly because they may become contaminated over time as new language models are introduced with the benchmarks potentially included as their training data (see e.g., Bubeck et al. (2023); Ibrahim et al. (2024)). Thus, rankings derived using these methods may not generalize to other, new tasks and questions, even within the same domain.

The second approach to LLM evaluation is a *human-based* approach in which human evaluators are asked to interact with and compare models by prompting, then ranking their performance based on their responses. An example is *Chatbot Arena*, recently introduced by Chiang et al. (2024), which evaluates LLMs using human feedback collected through crowd-sourcing. Chatbot Arena has attracted significant attention and media coverage (see, e.g., Yang and Cui (2024); Roose (2024)). Human-based approaches can effectively resolve the problem of data contamination in the static approaches. However, their reliance on human input limits their scalability. Designing suitable prompts and reading (often long) responses from various models can be very expensive and time-consuming.

The third approach to model evaluation is a *game-based* approach which bypasses the need for human evaluation by designing a structured *game* in which models compete against each other. The game is crafted with automatically checkable winning criteria, so the winner can be determined automatically without human intervention. The models are then ranked based on their performance in the game. The game should be designed in a way to require the relevant skills expected of language mod-

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 2040–2075

els so that the performance in these games serves as a proxy for the model's overall abilities. As a result, a good choice of the game is very important for the success of game-based approaches.

1.1 Evaluation via Debates

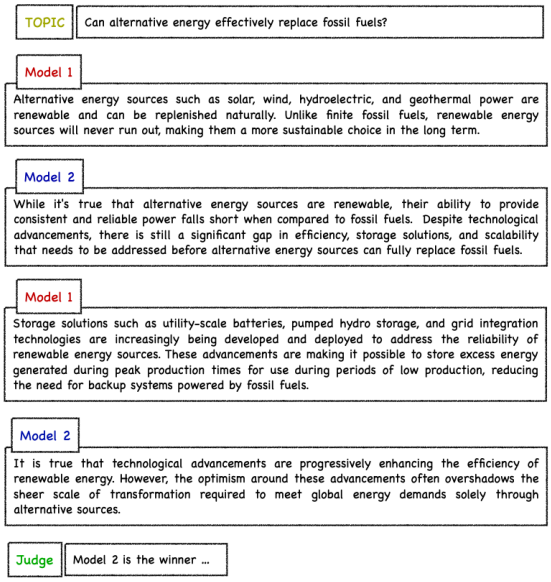
Debate has been a longstanding tradition since antiquity (Dutilh Novaes, 2022, Historical Supplement). It serves as a structured forum for testing conversational and reasoning skills, and has been integral to philosophical discourse, legal proceedings, and civic engagement (Proksch and Slapin (2015); Holbrook (1999); Benoit et al. (2003)). Although success in debates requires mastery of the debate topic and domain knowledge, it also needs skills such as defining the problem, identifying and challenging assumptions, recognizing inconsistencies, and prioritizing the relevance of various details within the overall argument (Roy and Macchiette, 2005; Kennedy, 2009). Such skills are crucial for the effective application of LLMs, which motivates using debate settings in game-based approach to LLM evaluation.

In this paper, we take the game-based approach to model evaluation and design the game to be a structured *debate* between competing language models, judged by a pre-specified language model. In this framework, models debate on a set of predetermined topics. The script of each debate is then given to a pre-determined judge, which is also an LLM, to evaluate and score the arguments presented by each side. Based on these evaluations, the language model with the better overall performance is announced as the winner. Through experiments with state-of-the-art language models, we demonstrate that this fully automated setup can be used to rank these LLMs, producing rankings generally consistent with those of Chatbot Arena, which is one of the most widely used methods of LLM evaluation.

2 Related Works

Collaboration and debates have been used as a framework to enhance the performance of language models. Liang et al. (2023); De Ridder (2024) propose multi-agent debate frameworks in which models express their solutions to a problem to encourage divergent thinking. Xiong et al. (2023) uses a formal debate framework to examine whether LLMs can collaborate to reach a consensus for a shared goal. Chern et al. (2024) proposes multi-round discussions between models to assist human annotators in finding the most capable LLM to be used as an evaluator. Khan et al. (2024) shows that having access to the a debate between two strong LLMs that posses the necessary information to answer the question of the topic will result in more truthful answers from a weak LLM or a human expert that do not have that piece of information. Lee et al. (2024) study the applications of LLMs in debate education. See also Pham et al. (2023); Fu et al. (2023); Madaan et al. (2024); Chao et al. (2024) for other results leveraging the interactions between language models to achieve a given goal.

Figure 1: A snippet of debates. Two language models engage in debates on a list of topics, and a judge model announces the winner for each topic. The language model with the most wins across all topics is declared the overall winner.



3 Debate Framework

Assume that $\mathcal{T}$ is a pre-defined list of debate topics and $\mathcal{LM}$ is a set of large language models that we want to rank based on their performance. The topics are open-ended questions like *"Can alternative energy effectively replace fossil fuels?"* that have two possible sides. We assume that we have *black-box query access* to all the language models. The ranking is based on multiple debates between the language models on different topics.

Each debate is a multi-round interaction on a topic $t \in \mathcal{T}$ between two models, a language model $\mathrm{LM}_1 \in \mathcal{LM}$ that goes first and support one side of the argument, and a language model $\mathrm{LM}_2 \in \mathcal{LM}$ that goes second and supports the other side. In the first round, LM_1 is asked to start the debate by providing the arguments to supports their side based on logic, facts, and evidence. Then, in the

next round LM_2 responds, refutes the arguments raised, and provides new evidence supporting the other side. The debate between the models continues for a predefined number T of rounds. Finally, the model LM_2 is asked to conclude the debate. All prompts used in our experiments can be found in Section A. At the end of each debate, the script of the T rounds is given to a judge LLM, which is asked to consider specific pre-defined factors, to score LM_1 and LM_2, and to announce the winner.

In this framework, to compare two models LM_a and LM_b from the set $\mathcal{L}$, we conduct two debates on each topic $t \in \mathcal{T}$. In the first debate, we set $\mathrm{LM}_1 = \mathrm{LM}_a$, and in the second debate, we set $\mathrm{LM}_1 = \mathrm{LM}_b$. The judge language models evaluate both debates. Based on their assessments, one of the models LM_1 or LM_2 is declared the winner for the topic t, or the result is a draw. We run two debates because the two sides of the argument might not be equally hard to argue for. Also, the judge could be biased and favor the model that goes first (or last). By running the debate with the role of the models flipped, we account for these biases.

This process is repeated for all topics $t \in \mathcal{T}$. The model with the highest number of wins across different topics is declared the overall winner. Finally, after comparing all pairs $(\mathrm{LM}_a, \mathrm{LM}_b)$ of models, the overall ranking of the models in $\mathcal{L}$ is generated.

3.1 Algorithm Details

Since LLMs are typically not directly trained to debate, specifying the rules of the task is crucial. To effectively do this, we provide a detailed system prompt to the models. The systems prompts ask the models to support a side; ask that the arguments and rebuttals should be backed by logic, facts, and evidence; and that the answers should be convincing, factual and concise. The history of all previous rounds of the debate is given to the debating LLMs.

The same is also true for the judge language model. We set a detailed system prompt for the judge so that it considers factors such as clarity of arguments, factuality and use of evidence, rebuttal and counterarguments, logical consistency, persuasiveness, conciseness, and coherence in the evaluation. Further, in the system prompt for the judge, we specify the exact format for the output. The script of the debate is given to the judge model using its prompt and the judge model announces a winner.

3.2 The Choice of the Judge LLM

LLMs have been used as judges for various applications (see e.g., Zheng et al. (2024); Chen et al. (2024); Chao et al. (2024); Kim et al.; Hada et al. (2024b,a); Wu and Aji (2023); Kim et al. (2024)).

Because of the symmetry in the game, the evaluation based on the game is *fair* regardless of the strength or weakness of the judge model. However, when we use weaker language models as a judge, we give preference to language models that are stronger at *convincing the judge that their argument is coherent*, instead of the models that are *actually* giving coherent responses. In other words, models that are able to generate responses that are marked by the judge to be more coherent are scored higher. Note that the same is true for human judges. For example, in Chatbot Arena, users select the response of an LLM not necessarily based on whether it is more “coherent", but based on whether it is more “coherent". Although this might not be the right metric for all applications, we argue that for the conversation-based tasks that chatbots are used for everyday, the persuasiveness is the key ability users seek when choosing the language model. For example, see OpenAI o1 System Card (OpenAI, 2024, Section 4.7.1).

For the task of evaluating debates, Liu et al. (2024b) showed that of GPT-4 outperforms humans and other state-of-the-art LLMs fine-tuned on extensive datasets in debate evaluation. More generally, GPT-4 models have consistently been demonstrated to closely match with human intentions when acting as a judge and have been used as a judge extensively in LLM literature (see e.g., Zheng et al. (2024); Achiam et al. (2023) and references therein). Based on these findings, we choose GPT-4 as the judge model in our experiments. In Section 4.2, conduct some experiments with Llama-3-70b as the judge and demonstrate that the rankings do not change significantly when Llama-3-70b is used as judge. Also, we conducted an experiment where a human judge also asked to determine the winner of debates by reading them. We showed that the winners chosen by GPT-4 is consistent with the winners chosen by the human evaluator.

In this paper, we evaluate the overall conversation ability of language models with general questions and use general-purpose LLMs as judge. However, we note that if all the debate topics are from a specific subject, choosing a Retrieval-Augmented Generation LLM (see e.g., (Lewis

et al., 2020)) with a proper knowledge-base might be better suited for determining the factuality of the claims by different sides of the debate.

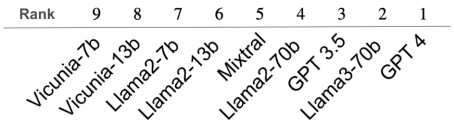
4 Experimental Results

In this section, we consider the list of debate topics from Section B and run debates with four rounds, i.e., $T = 4$. As the debating models, we use Llama-2-7b, Llama-2-13b, Llama-2-70b (Touvron et al., 2023), Llama-3-70b (Meta AI, 2024), Vicuna-7b-v1.5,Vicuna-13b-v1.5 (Chiang et al., 2023), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024), gpt-4-0125-preview, and gpt-3.5-turbo-0125 (Achiam et al., 2023). To access GPT models, we use the OpenAI API. For other models, we use the API from `https://www.together.ai/`.

4.1 Rankings

We run a total of 50 debates on 25 topics (Section B) between each pair of models. In this section we use gpt-4-0125-preview as the judge. On each topic, the model that wins both rounds is considered the winner. Table 1 (Left) shows the number of wins in the debates between various model pairs. For example, Llama-2-7b has won in no topics against Llama-2-70b, whereas Llama-2-70b has won in nine debates against Llama-2-7b. The detailed results in different topics can be found in Section C. See Table 1 and Figure 2

Figure 2: Overall ranking of LLMs with GPT-4 as judge.



From these results, as a sanity check, it is seen that for different families of models (Llama-2, GPT, Vicuna), models with more parameters rank better; i.e., Llama-2-70b ranks better compared to Llama-2-7b. Also, newer generations of each model rank better than their older counterpart; i.e., GPT-4 ranks better compared to GPT-3.5. Similarly, Llama-3-70b ranks better compared to Llama-2-70b.

Comparing this ranking with the rankings available on the Chatbot Arena leaderboard website, accessed on June 14th, 2024, we see that the rankings in Table 1 are generally consistent with Chatbot Arena. Specifically, the normalized Kendall tau distance between these two rankings is **0.0833**. This distance takes values in $[0, 1]$ where 0 means identical rankings and 1 means reversed rankings.

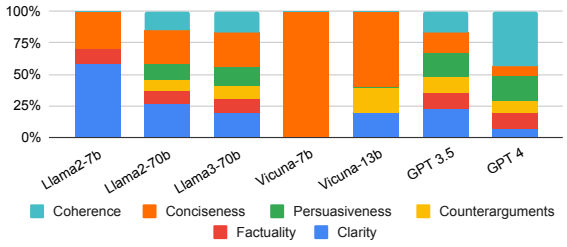
Table 1: The results from Section 4.1 with GPT-4 as judge. Numbers indicate the number of topics in which each model has won against other models

	Llama-2-7b	Llama-2-13b	Llama-2-70b	Llama-3-70b	Vicuna-7b	Vicuna-13b	Mixtral-8x7B	GPT-3.5	GPT-4
Llama-2-7b	1-0	0-6	0-9	0-11	9-0	6-2	1-8	0-5	0-24
Llama-2-13b		0-2	0-7	2-13	14-0	8-0	0-5	0-7	0-22
Llama-2-70b			1-1	0-9	13-0	12-0	2-1	1-2	0-21
Llama-3-70b				0-1	23-0	17-0	6-0	5-0	0-13
Vicuna-7b					2-1	0-3	0-11	0-18	0-24
Vicuna-13b						0-0	0-13	0-13	0-23
Mixtral-8x7B							3-1	0-3	0-24
GPT-3.5								3-2	1-14
GPT-4									2-2

4.2 Other Experiments

Analysis of Content. We prompt the judge (GPT-4) model to state the main reason for its choice among "clarity, factuality, counterarguments, persuasiveness, conciseness, and coherence." Figure 3 illustrates the stated reasons for which different models won the debates against their opponents. This shows that clarity and coherence have been the most decisive factors in the decisions. It also reveals that, for example, the responses by GPT-4 were seen by the judge as being more coherent, whereas the arguments by Vicuna-7b were seen to be more concise.

Figure 3: The percentage of the times each model won against another model for each of the six reasons.



Human as Judge. In this experiment, we asked three human student volunteers to read the contents of the debates between Llama-2-13b and Llama-2-70b and judge the debates. The human evaluators each judged a total of 50 debates and their judgments matched the results of Section C.11 in 81.3% of the time. In particular, each participant agreed with the judgment from Section C.11 in 43, 41, and 38 debates out of 50.

Llama-3 as Judge. Finally, we conduct similar experiments, but with Llama-3-70b as the judge.

For demonstration, we only repeat the debates between Llama-2-13b with all other models. The score of Llama-2-13b against other models with Llama-3-70b as judge is shown in Table 2. The winners in this experiment are identical to the winners announced by GPT-4, in all but one opponent models (Mixtral).

Table 2: Score of Llama-2-13b vs. other models (first numbers for Llama-2-13b), with Llama-3-70b as judge.

Model	Score	Model	Score
Llama-2-7b	7-0	Vicuna-13b	17-0
Llama-2-70b	0-4	Mixtral-8x7B	8-4
Llama-3-70b	0-6	GPT-3.5	0-4
Vicuna-7b	16-2	GPT-4	0-21

5 Conclusion

In this paper, we developed an automated framework to rank the performance of LLMs, based on a multi-round debate between LLMs on different topics, and an evaluation by a judge LLM. We showed that this framework can yield rankings consistent with rankings that rely on human crowdsourcing, while being more scalable.

6 Limitations

List of Topics. The method proposed in this paper requires human input to create the list of debate topics. While this is significantly less time-consuming and less expensive than reading and scoring the debates, it can still pose some scalability issues. We leave the task of automating topic generation for future work.

LLM as Judge. The debate framework introduced in this paper is a game-based approach to model evaluation and uses a judge LLM as a part of the game. Because of the symmetry in the game design, the judge cannot bias the game; however, if a weak LLM is chosen as the judge, it is possible that the judge may not be capable of fully evaluating qualities such as the “factuality" of the arguments. In such cases, the winner is determined by the abilities of different models to *convince* the judge that they are more factual, instead of *actually* giving more factual answers. Although this still demonstrates the abilities of that language model, it might not necessarily be the qualities that we try to evaluate. Note that the same is true for human judges. For example, in Chatbot Arena, users select an LLM not necessarily based on whether it is generated more “factual" responses, but based on whether it can generate more “convincing" outputs.

In our experiments in Section 4.2, it was shown that the decisions are mostly based on factors such as coherence, conciseness, or clarity that are easier for a language model to evaluate. However, despite the evidence (see e.g., Liu et al. (2024b)) that models such as GPT-4 perform well on debate evaluation tasks, in general this limits the applicability of the debate framework to evaluate the ability of models in tasks other than the general conversational ability that was studied here, where factors such as factuality are more important.

Choice of Language. Our method has been primarily evaluated on English, which is a language with relatively limited morphological complexity. This could restrict the applicability of our approach to other languages.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

William L Benoit, Glenn J Hansen, and Rebecca M Verser. 2003. A meta-analysis of the effects of viewing us presidential debates. *Communication monographs*, 70(4):335–350.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Nicholas Carlini. 2024. My benchmark for large language models. Accessed: 06-03-2024.

Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of medical systems*, 47(1):33.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. 2024. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *arXiv preprint arXiv:2401.16788*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. *arXiv preprint arXiv:2403.04132*.

Alexander De Ridder. 2024. Ai-driven debates as a tool for advancing understanding and decision-making.

Xueying Du, Mingwei Liu, Kaixin Wang, Hanlin Wang, Junwei Liu, Yixuan Chen, Jiayi Feng, Chaofeng Sha, Xin Peng, and Yiling Lou. 2024. Evaluating large language models in class-level code generation. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.

Catarina Dutilh Novaes. 2022. Argument and Argumentation. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2022 edition. Metaphysics Research Lab, Stanford University.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024a. Metal: Towards multilingual meta-evaluation. *arXiv preprint arXiv:2404.01667*.

Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024b. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Thomas M Holbrook. 1999. Political learning from presidential debates. *Political Behavior*, 21:67–89.

Lujain Ibrahim, Saffron Huang, Lama Ahmad, and Markus Anderljung. 2024. Beyond static AI evaluations: advancing human interaction evaluations for LLM harms and risks. *arXiv preprint arXiv:2405.10632*.

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*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254.

Ruth R Kennedy. 2009. The power of in-class debates. *Active learning in higher education*, 10(3):225–236.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Unggi Lee, Yeil Jeong, Junbo Koh, Gyuri Byun, Yunseo Lee, Youngsun Hwang, Hyeoncheol Kim, and Cheolil Lim. 2024. Can chatgpt be a debate partner? developing chatgpt-based application "debo" for debate education, findings and limitations. *Educational Technology & Society*, 27(2):pp. 321–346.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking

in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2024a. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems*.

Xinyi Liu, Pinxin Liu, and Hangfeng He. 2024b. An empirical analysis on large language models in debate evaluation. *arXiv preprint arXiv:2406.00050*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*.

Meta AI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

OpenAI. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A Plummer, Zhaoran Wang, and Hongxia Yang. 2023. Let models speak ciphers: Multiagent debate through embeddings. *arXiv preprint arXiv:2310.06272*.

Sven-Oliver Proksch and Jonathan B Slapin. 2015. *The politics of parliamentary debate*. Cambridge University Press.

Kevin Roose. 2024. A.I. has a measurement problem. *New York Times*.

Abhijit Roy and Bart Macchiette. 2005. Debating the issues: A tool for augmenting critical thinking skills of marketing students. *Journal of Marketing Education*, 27(3):264–276.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. Arb: Advanced reasoning benchmark for large language models. *arXiv preprint arXiv:2307.13692*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

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*.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Angela Yang and Jasmine Cui. 2024. ChatGPT still reigns supreme in many AI rankings, but the competition is on. *NBC News*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging LLM-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A System Prompts and Prompts

In this section, we list the system prompts and prompts used in the experiments.

A.1 Debating Models

The first language model LM_1 is asked to start the debate. This language model will have the following system prompt.

System Prompt for LM_1

We are having a debate and the topic is "**TOPIC**". You are representing "**SIDE 1**" and you should zealously support it. This the first round and you are going first. You must bring-up arguments supporting your side backed by logic, facts and evidence. Your answer should also be convincing, factual and concise.

We will prompt LM_1 as follows.

Prompt for LM_1

Please start the debate.

In the second part of the debate, the second language model LM_2 will be given the response of LM_1. The second model is asked to provide a rebuttal to the points raised by LM_1. Also, it is asked to provide new arguments supporting the second side. This language model will have the following system prompt.

System Prompt for LM_2

We are having a debate and the topic is "**TOPIC**". You are representing "**SIDE 2**" and you should zealously support it. The other side has started the debate and you will be given their arguments. You must first provide a rebuttal to the points raised by them, and then provide new arguments supporting your side. All your arguments should be backed by logic, facts and evidence. Your answer should also be convincing, factual and concise.

We prompt LM_2 as follows.

Prompt for LM_2

The other side said: "**Response of** LM_1 **in Part 1**".

In all following steps, we use the following system prompts for the models.

System Prompt

We are having a debate and the topic is "**TOPIC**". You are representing "**SIDE**" and you should zealously support it. You will be given all the arguments so far. In your response, you should support your side and refute the points raised by the other side. Your arguments should be backed by logic, facts and evidence. Your answer should also be convincing, factual and concise.

We prompt the models LM_1 and LM_2 as follows.

Prompt for LM_1

You initially said: "**Response of** LM_1 **in Part 1**". You the other side responded: "**Response of** LM_2 **in Part 2**". Then you said: "**Response of** LM_1 **in Part 3**", ...

Prompt for LM_2

The other side initially said: "**Response of LM_1 in Part 1**". You then responded: "**Response of LM_2 in Part 2**". The other side said: "**Response of LM_1 in Part 3**", ...

A.2 Judge Model

A judge language model gives scores from 1 to 10 to each debate participant based on the following criteria: (1) Clarity of arguments (2) Factuality and use of evidence (3) Rebuttal and counterarguments (4) Logical consistency (5) Persuasiveness and impact (6) Conciseness (7) Coherence. The judges will have the following system prompt.

System Prompt for the Judge

We had a debate and the topic was "**TOPIC**". The two sides in the debate each provided arguments to prove their side and refute the points raised by the opponent. You are a judge for this debate. You should be impartial and as objective as possible. The debate script will be given. You should give a score from 1 to 10 to each side of the debate. In your judgement, you should take into account the following criteria: clarity of arguments, factuality and use of evidence, rebuttal and counterarguments, logical consistency, persuasiveness and impact, conciseness, coherence. Also, you should choose the side who you think is the overall winner. Your answer MUST follow the following format: "side1: [[score of side 1]], side2: [[score of side 2]], winner: [[name of winner]]" where score of side 1 and score of side 2 are numbers from 1 to 10 and name of winner is either "1" or "2".

The judge generates the scores given the whole script of the debate, using the following prompt.

Prompt for the Judge

The script of the debate is as follows: Side 1: "**Response of LM_1 in Part 1**". Side 2: "**Response of LM_2 in Part 2**". Side 1: "**Response of LM_1 in Part 3**". Side 2: "**Response of LM_2 in Part 4**".

B Topic of Debates

The twenty-five debate topics used in the experiments are as follows.

1. Can alternative energy effectively replace fossil fuels?
2. Should K-12 students dissect animals in science classrooms?
3. Is artificial intelligence good for society?
4. Should bottled water be banned?
5. Is a college education worth it?
6. Should the United States keep daylight saving time?
7. Should school dress codes be implemented and enforced?
8. Should the drinking age be lowered from 21 to a younger age?
9. Should the election day be made a national holiday?
10. Should the governments use Large Language Models for advice?
11. Should employers be able to mandate vaccinations?
12. Should fighting be allowed in hockey?
13. Should fur clothing be banned?

14. Should genetically modified organisms (GMOs) be grown?
15. Is golf a sport and are golfers athletes?
16. Is homework beneficial?
17. Is the internet "making us stupid?"
18. Should medical aid in dying be legal?
19. Is obesity a disease?
20. Should the penny stay in circulation?
21. Are the Olympic games an overall benefit for their host countries and cities?
22. Is there really a Santa Claus?
23. Should Halloween be moved permanently to Saturday?
24. Should students have to wear school uniforms?
25. Is social media good for society?

C Experimental Results

In this section, we report the experimental results with GPT-4 as judge. Each table is the result of the debates between two models on all topics. *Home* and *Away* correspond to to runs of the debate on each topic, each time with one LLM going first.

C.1 Llama-2-7b vs Llama-2-7b

	Home			Away		
Topic	Winner	Side 1 Llama-2-7b	Side 2 Llama-2-7b	Winner	Side 1 Llama-2-7b	Side 2 Llama-2-7b
1	Side 1	8	7	Side 1	9	7
2	Side 2	7	8	Side 2	8	9
3	Side 1	8	7	Side 1	8	7
4	Side 1	8	7	Side 1	8	7
5	Side 1	8	7	Side 2	8	9
6	Side 1	8	7	Side 1	8	7
7	Side 2	7	8	Side 2	7	8
8	Side 2	7	8	Side 2	6	8
9	Side 1	8	7	Side 1	8	7
10	Side 1	8	7	Side 1	8	7
11	Side 1	9	7	Side 1	8	7
12	Side 2	7	8	Side 2	7	8
13	Side 1	8	7	Side 1	8	7
14	Side 1	8	7	Side 1	8	7
15	Side 1	8	6	Side 1	8	5
16	Side 2	7	8	Side 2	7	8
17	Side 2	7	8	Side 2	7	9
18	Side 1	8	7	Side 1	8	7
19	Side 1	9	7	Side 1	9	7
20	Side 2	7	8	Side 2	7	8
21	Side 2	7	8	Side 2	8	9
22	Side 2	6	8	Side 2	6	8
23	Side 1	8	7	Side 1	8	7
24	Side 2	7	8	Side 2	7	8
25	Side 1	8	7	Side 1	8	7

C.2 Llama-2-7b vs Llama-2-13b

	Home			Away			
Topic	Winner	Side 1 Llama-2-13b	Side 2 Llama-2-7b	Winner	Side 1 Llama-2-7b	Side 2 Llama-2-13b	Overall
1	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
2	Llama-2-7b	7	8	Llama-2-13b	7	8	Tie
3	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
4	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
5	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
6	Llama-2-13b	8	7	Tie	8	8	Llama-2-13b
7	Llama-2-7b	8	9	Llama-2-13b	7	9	Tie
8	Llama-2-7b	6	8	Llama-2-13b	7	8	Tie
9	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
10	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
11	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
12	Llama-2-7b	7	8	Llama-2-13b	7	8	Tie
13	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
14	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
15	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
16	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
17	Llama-2-7b	6	8	Llama-2-13b	7	8	Tie
18	Llama-2-13b	8	7	Llama-2-7b	8	7	Tie
19	Llama-2-13b	9	7	Llama-2-7b	8	7	Tie
20	Llama-2-7b	6	8	Llama-2-13b	7	8	Tie
21	Llama-2-13b	8	7	Llama-2-13b	8	9	Llama-2-13b
22	Llama-2-7b	6	8	Llama-2-13b	6	8	Tie
23	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
24	Llama-2-13b	8	7	Llama-2-13b	8	9	Llama-2-13b
25	Llama-2-13b	8	7	Llama-2-13b	8	9	Llama-2-13b

C.3 Llama-2-7b vs. Llama-2-70b

	Home			Away			
Topic	Winner	Side 1 Llama-2-7b	Side 2 Llama-2-70b	Winner	Side 1 Llama-2-70b	Side 2 Llama-2-7b	Overall
1	Llama-2-7b	8	7	Llama-2-70b	9	7	Tie
2	Llama-2-70b	7	9	Llama-2-7b	7	8	Tie
3	Llama-2-7b	8	7	Llama-2-70b	8	7	Tie
4	Llama-2-7b	8	7	Llama-2-70b	8	7	Tie
5	Llama-2-70b	8	9	Llama-2-70b	8	7	Llama-2-70b
6	Llama-2-70b	7	8	Llama-2-70b	8	7	Llama-2-70b
7	Llama-2-70b	7	8	Llama-2-70b	8	7	Llama-2-70b
8	Llama-2-70b	6	8	Llama-2-7b	7	8	Tie
9	Llama-2-7b	8	7	Llama-2-70b	8	7	Tie
10	Llama-2-70b	8	8	Llama-2-70b	8	8	Llama-2-70b
11	Llama-2-7b	8	7	Llama-2-70b	8	7	Tie
12	Llama-2-70b	7	9	Llama-2-7b	8	9	Tie
13	Llama-2-70b	7	8	Llama-2-70b	8	7	Llama-2-70b
14	Llama-2-7b	8	7	Llama-2-70b	8	7	Tie
15	Llama-2-7b	8	7	Llama-2-70b	8	7	Tie
16	Llama-2-7b	8	8	Llama-2-70b	8	7	Tie
17	Llama-2-70b	7	8	Llama-2-7b	7	8	Tie
18	Llama-2-7b	8	7	Llama-2-70b	8	7	Tie
19	Llama-2-7b	9	8	Llama-2-70b	8	7	Tie
20	Llama-2-70b	6	8	Llama-2-7b	7	8	Tie
21	Llama-2-70b	8	9	Llama-2-70b	8	7	Llama-2-70b
22	Llama-2-70b	6	8	Llama-2-70b	8	7	Llama-2-70b
23	Llama-2-7b	8	7	Llama-2-70b	8	7	Tie
24	Llama-2-70b	7	8	Llama-2-70b	8	7	Llama-2-70b
25	Llama-2-70b	7	8	Llama-2-70b	9	8	Llama-2-70b

C.4 Llama-2-7b vs. Llama-3-70b

	Home			Away			
Topic	Winner	Side 1 Llama-3-70b	Side 2 Llama-2-7b	Winner	Side 1 Llama-2-7b	Side 2 Llama-3-70b	Overall
1	Llama-3-70b	9	8	Llama-2-7b	9	8	Tie
2	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
3	Llama-2-7b	8	9	Llama-3-70b	8	9	Tie
4	Llama-3-70b	8	6	Llama-2-7b	8	7	Tie
5	Llama-3-70b	8	7	Llama-2-7b	8	7	Tie
6	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
7	Llama-3-70b	8	7	Llama-3-70b	7	9	Llama-3-70b
8	Llama-3-70b	8	7	Llama-3-70b	6	8	Llama-3-70b
9	Llama-3-70b	8	7	Llama-2-7b	8	7	Tie
10	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
11	Llama-3-70b	8	7	Llama-2-7b	8	7	Tie
12	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
13	Llama-3-70b	8	6	Llama-2-7b	8	7	Tie
14	Llama-3-70b	8	7	Llama-2-7b	8	7	Tie
15	Llama-3-70b	9	7	Llama-2-7b	8	7	Tie
16	tie	8	8	Llama-3-70b	8	9	Llama-3-70b
17	Llama-2-7b	7	8	Llama-3-70b	7	9	Tie
18	Llama-3-70b	8	7	Llama-2-7b	8	7	Tie
19	Llama-3-70b	9	8	Llama-2-7b	9	7	Tie
20	Llama-2-7b	7	8	Llama-3-70b	6	8	Tie
21	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
22	Llama-2-7b	7	8	Llama-3-70b	6	8	Tie
23	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
24	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
25	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b

C.5 Llama-2-7b vs. Vicuna-7b-v1.5

	Home			Away			
Topic	Winner	Side 1 Llama-2-7b	Side 2 Vicuna-7b-v1.5	Winner	Side 1 Vicuna-7b-v1.5	Side 2 Llama-2-7b	Overall
1	Llama-2-7b	9	7	Vicuna-7b-v1.5	8	7	Tie
2	Vicuna-7b-v1.5	6	8	Llama-2-7b	7	8	Tie
3	Llama-2-7b	8	7	tie	8	8	Llama-2-7b
4	Llama-2-7b	8	7	Vicuna-7b-v1.5	8	7	Tie
5	Llama-2-7b	9	8	Llama-2-7b	8	9	Llama-2-7b
6	Llama-2-7b	8	7	Llama-2-7b	7	8	Llama-2-7b
7	Vicuna-7b-v1.5	7	8	Llama-2-7b	7	8	Tie
8	Vicuna-7b-v1.5	7	8	Llama-2-7b	7	8	Tie
9	Llama-2-7b	8	7	Vicuna-7b-v1.5	8	7	Tie
10	Llama-2-7b	8	7	Llama-2-7b	7	8	Llama-2-7b
11	Llama-2-7b	8	7	tie	8	8	Llama-2-7b
12	Vicuna-7b-v1.5	7	8	Llama-2-7b	6	8	Tie
13	Llama-2-7b	9	7	Llama-2-7b	7	8	Llama-2-7b
14	Llama-2-7b	8	7	Vicuna-7b-v1.5	8	7	Tie
15	Llama-2-7b	8	7	Vicuna-7b-v1.5	8	7	Tie
16	Llama-2-7b	7	6	Llama-2-7b	7	8	Llama-2-7b
17	Vicuna-7b-v1.5	6	8	Llama-2-7b	6	8	Tie
18	Llama-2-7b	9	7	Vicuna-7b-v1.5	8	7	Tie
19	Llama-2-7b	9	7	Llama-2-7b	8	9	Llama-2-7b
20	Llama-2-7b	8	7	Llama-2-7b	6	8	Llama-2-7b
21	Llama-2-7b	8	7	Llama-2-7b	8	9	Llama-2-7b
22	Vicuna-7b-v1.5	6	8	Llama-2-7b	5	7	Tie
23	Vicuna-7b-v1.5	6	8	Llama-2-7b	6	8	Tie
24	Llama-2-7b	8	7	Llama-2-7b	7	8	Llama-2-7b
25	Vicuna-7b-v1.5	7	8	Llama-2-7b	7	8	Tie

C.6 Llama-2-7b vs. Vicuna-13b-v1.5

	Home			Away			
Topic	Winner	Side 1 Llama-2-7b	Side 2 Vicuna-13b-v1.5	Winner	Side 1 Vicuna-13b-v1.5	Side 2 Llama-2-7b	Overall
1	Llama-2-7b	8	7	Vicuna-13b-v1.5	8	7	Tie
2	Vicuna-13b-v1.5	7	8	Llama-2-7b	7	8	Tie
3	Llama-2-7b	8	7	Vicuna-13b-v1.5	8	7	Tie
4	Llama-2-7b	8	7	Llama-2-7b	7	8	Llama-2-7b
5	Llama-2-7b	8	7	Vicuna-13b-v1.5	8	7	Tie
6	Llama-2-7b	8	7	Llama-2-7b	7	8	Llama-2-7b
7	Vicuna-13b-v1.5	8	9	Llama-2-7b	7	8	Tie
8	Vicuna-13b-v1.5	6	8	Llama-2-7b	7	8	Tie
9	Llama-2-7b	8	7	Vicuna-13b-v1.5	7	6	Tie
10	Llama-2-7b	8	7	Llama-2-7b	8	9	Llama-2-7b
11	Llama-2-7b	8	7	Vicuna-13b-v1.5	8	7	Tie
12	Vicuna-13b-v1.5	8	9	Llama-2-7b	6	9	Tie
13	Llama-2-7b	8	7	Llama-2-7b	7	8	Llama-2-7b
14	Llama-2-7b	8	7	Vicuna-13b-v1.5	8	7	Tie
15	Llama-2-7b	8	6	Llama-2-7b	7	8	Llama-2-7b
16	Vicuna-13b-v1.5	8	8	Llama-2-7b	7	8	Tie
17	Vicuna-13b-v1.5	7	8	Llama-2-7b	7	9	Tie
18	Llama-2-7b	8	7	Vicuna-13b-v1.5	8	7	Tie
19	Llama-2-7b	9	7	Llama-2-7b	8	9	Llama-2-7b
20	Vicuna-13b-v1.5	7	8	Llama-2-7b	6	8	Tie
21	Vicuna-13b-v1.5	8	9	Llama-2-7b	7	8	Tie
22	Vicuna-13b-v1.5	7	8	Llama-2-7b	7	8	Tie
23	Vicuna-13b-v1.5	7	8	Vicuna-13b-v1.5	8	7	Vicuna-13b-v1.5
24	Llama-2-7b	8	7	Llama-2-7b	6	8	Llama-2-7b
25	tie	8	8	Vicuna-13b-v1.5	8	7	Vicuna-13b-v1.5

C.7 Llama-2-7b vs. Mixtral-8x7B

	Home			Away			
Topic	Winner	Side 1 Llama-2-7b	Side 2 Mixtral-8x7B	Winner	Side 1 Mixtral-8x7B	Side 2 Llama-2-7b	Overall
1	Llama-2-7b	8	7	Mixtral-8x7B	8	7	Tie
2	Mixtral-8x7B	7	8	Llama-2-7b	7	8	Tie
3	Llama-2-7b	8	7	Mixtral-8x7B	8	7	Tie
4	Llama-2-7b	8	7	Mixtral-8x7B	8	7	Tie
5	Llama-2-7b	8	7	Mixtral-8x7B	8	7	Tie
6	Llama-2-7b	8	7	Llama-2-7b	8	9	Llama-2-7b
7	Mixtral-8x7B	7	8	Llama-2-7b	7	8	Tie
8	Mixtral-8x7B	7	8	Llama-2-7b	7	8	Tie
9	Mixtral-8x7B	8	9	Draw	8	8	Mixtral-8x7B
10	Llama-2-7b	8	7	Mixtral-8x7B	8	7	Tie
11	Llama-2-7b	8	7	Mixtral-8x7B	8	7	Tie
12	Mixtral-8x7B	7	8	Llama-2-7b	8	9	Tie
13	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
14	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
15	Llama-2-7b	8	7	Mixtral-8x7B	8	7	Tie
16	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B
17	Mixtral-8x7B	7	8	Llama-2-7b	7	8	Tie
18	Llama-2-7b	8	7	Mixtral-8x7B	8	7	Tie
19	Llama-2-7b	8	7	Mixtral-8x7B	8	7	Tie
20	Mixtral-8x7B	7	8	Llama-2-7b	7	8	Tie
21	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B
22	Mixtral-8x7B	6	8	Llama-2-7b	7	8	Tie
23	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
24	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
25	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B

C.8 Llama-2-7b vs GPT 3.5

	Home			Away			
Topic	Winner	Side 1 Llama-2-7b	Side 2 GPT-3.5	Winner	Side 1 GPT-3.5	Side 2 Llama-2-7b	Overall
1	Llama-2-7b	8	7	GPT-3.5	9	7	Tie
2	GPT-3.5	7	8	Llama-2-7b	7	8	Tie
3	GPT-3.5	8	9	Llama-2-7b	8	9	Tie
4	Llama-2-7b	8	7	Llama-2-7b	7	8	Llama-2-7b
5	GPT-3.5	8	9	GPT-3.5	8	7	GPT-3.5
6	GPT-3.5	8	9	Llama-2-7b	7	8	Tie
7	GPT-3.5	7	8	Llama-2-7b	7	8	Tie
8	GPT-3.5	6	8	Llama-2-7b	8	7	Tie
9	Llama-2-7b	8	7	GPT-3.5	8	7	Tie
10	GPT-3.5	8	9	GPT-3.5	8	7	GPT-3.5
11	Llama-2-7b	8	7	GPT-3.5	8	7	Tie
12	GPT-3.5	7	9	Llama-2-7b	7	9	Tie
13	Llama-2-7b	8	7	GPT-3.5	8	7	Tie
14	GPT-3.5	8	9	GPT-3.5	8	7	GPT-3.5
15	Llama-2-7b	8	7	GPT-3.5	8	7	Tie
16	GPT-3.5	8	9	Llama-2-7b	8	9	Tie
17	GPT-3.5	7	9	Llama-2-7b	7	8	Tie
18	Llama-2-7b	8	7	GPT-3.5	8	7	Tie
19	Llama-2-7b	8	7	GPT-3.5	8	7	Tie
20	GPT-3.5	7	8	Llama-2-7b	7	8	Tie
21	GPT-3.5	8	9	GPT-3.5	8	7	GPT-3.5
22	GPT-3.5	6	8	Llama-2-7b	7	8	Tie
23	GPT-3.5	7	8	GPT-3.5	8	7	GPT-3.5
24	GPT-3.5	7	8	Llama-2-7b	7	8	Tie
25	GPT-3.5	8	9	Llama-2-7b	8	9	Tie

C.9 Llama-2-7b vs GPT 4

	Home			Away			
Topic	Winner	Side 1 Llama-2-7b	Side 2 GPT-4	Winner	Side 1 GPT-4	Side 2 Llama-2-7b	Overall
1	Llama-2-7b	9	8	GPT-4	9	7	Tie
2	GPT-4	7	9	GPT-4	8	7	GPT-4
3	GPT-4	8	9	GPT-4	8	7	GPT-4
4	GPT-4	8	9	GPT-4	8	7	GPT-4
5	GPT-4	8	9	GPT-4	9	8	GPT-4
6	GPT-4	7	9	GPT-4	8	7	GPT-4
7	GPT-4	7	9	GPT-4	8	7	GPT-4
8	GPT-4	6	8	GPT-4	8	7	GPT-4
9	GPT-4	7	8	GPT-4	9	7	GPT-4
10	GPT-4	7	8	GPT-4	8	7	GPT-4
11	GPT-4	7	8	GPT-4	8	7	GPT-4
12	GPT-4	6	9	GPT-4	8	7	GPT-4
13	GPT-4	7	8	GPT-4	9	7	GPT-4
14	GPT-4	7	9	GPT-4	9	7	GPT-4
15	GPT-4	7	8	GPT-4	9	7	GPT-4
16	GPT-4	7	9	GPT-4	8	7	GPT-4
17	GPT-4	7	9	GPT-4	8	7	GPT-4
18	GPT-4	8	9	GPT-4	9	7	GPT-4
19	GPT-4	8	9	GPT-4	8	7	GPT-4
20	GPT-4	6	8	GPT-4	8	7	GPT-4
21	GPT-4	7	9	GPT-4	8	7	GPT-4
22	GPT-4	7	9	GPT-4	8	7	GPT-4
23	GPT-4	7	8	GPT-4	8	7	GPT-4
24	GPT-4	7	9	GPT-4	8	7	GPT-4
25	GPT-4	8	9	GPT-4	8	7	GPT-4

C.10 Llama-2-13b vs Llama-2-13b

	Home			Away		
Topic	Winner	Side 1 Llama-2-13b	Side 2 Llama-2-13b	Winner	Side 1 Llama-2-13b	Side 2 Llama-2-13b
1	Side 1	8	7	Side 1	8	7
2	Side 2	8	9	Side 2	7	8
3	Side 1	8	7	Side 1	8	7
4	Side 1	8	7	Side 1	8	7
5	Side 1	8	7	Side 1	8	7
6	Side 2	8	9	Side 1	8	7
7	Side 1	9	8	Side 1	8	7
8	Side 2	7	8	Side 2	7	8
9	Side 1	8	7	Side 1	8	7
10	Side 1	8	7	Side 1	8	7
11	Side 2	7	8	Side 1	8	7
12	Side 2	7	8	Side 2	7	8
13	Side 1	8	7	Side 1	8	7
14	Side 1	8	7	Side 1	8	7
15	Side 1	8	7	Side 1	8	7
16	Side 1	8	8	Side 1	8	8
17	Side 2	7	8	Side 2	7	8
18	Side 1	8	7	Side 1	8	7
19	Side 1	9	7	Side 1	9	7
20	Side 2	6	8	Side 2	7	8
21	Side 2	8	9	Side 2	8	9
22	Side 2	6	8	Side 2	6	8
23	Side 1	8	7	Side 1	8	7
24	Side 2	8	9	Side 2	8	9
25	Side 2	8	9	Side 2	8	9

C.11 Llama-2-13b vs Llama-2-70b

	Home			Away			
Topic	Winner	Side 1 Llama-2-13b	Side 2 Llama-2-70b	Winner	Side 1 Llama-2-70b	Side 2 Llama-2-13b	Overall
1	Llama-2-13b	8	6	Llama-2-70b	8	7	Tie
2	Llama-2-70b	7	8	Llama-2-13b	8	9	Tie
3	Llama-2-70b	8	9	Llama-2-70b	8	7	Llama-2-70b
4	Llama-2-13b	8	7	Llama-2-70b	8	7	Tie
5	Llama-2-13b	8	7	Llama-2-70b	8	7	Tie
6	Llama-2-70b	8	9	Llama-2-70b	8	7	Llama-2-70b
7	Llama-2-70b	7	8	Llama-2-70b	8	7	Llama-2-70b
8	Llama-2-70b	7	8	Llama-2-70b	8	7	Llama-2-70b
9	Llama-2-70b	7	8	Llama-2-70b	8	7	Llama-2-70b
10	Llama-2-13b	8	7	Llama-2-70b	8	7	Tie
11	Llama-2-13b	8	7	Llama-2-70b	8	7	Tie
12	Llama-2-70b	7	8	Llama-2-13b	8	9	Tie
13	Llama-2-13b	8	7	Llama-2-70b	8	7	Tie
14	Llama-2-13b	8	7	Llama-2-70b	8	7	Tie
15	Llama-2-13b	8	6	Llama-2-70b	8	6	Tie
16	Llama-2-70b	8	9	Llama-2-70b	8	7	Llama-2-70b
17	Llama-2-70b	7	8	Llama-2-13b	7	8	Tie
18	Llama-2-13b	8	7	Llama-2-70b	8	7	Tie
19	Llama-2-13b	8	6	Llama-2-70b	8	6	Tie
20	Llama-2-70b	6	8	Llama-2-13b	7	8	Tie
21	Llama-2-70b	8	9	Llama-2-70b	8	7	Llama-2-70b
22	Llama-2-70b	6	8	Llama-2-13b	7	8	Tie
23	Llama-2-13b	8	7	Llama-2-70b	8	7	Tie
24	Llama-2-70b	7	8	Llama-2-13b	8	9	Tie
25	Llama-2-13b	8	7	Llama-2-70b	8	8	Tie

C.12 Llama-2-13b vs. Llama-3-70b

	Home			Away			
Topic	Winner	Side 1 Llama-3-70b	Side 2 Llama-2-13b	Winner	Side 1 Llama-2-13b	Side 2 Llama-3-70b	Overall
1	Llama-3-70b	9	7	Llama-2-13b	8	7	Tie
2	tie	8	8	Llama-3-70b	7	8	Llama-3-70b
3	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
4	Llama-3-70b	8	7	Llama-2-13b	8	7	Tie
5	Llama-3-70b	8	7	Llama-2-13b	8	7	Tie
6	Llama-2-13b	8	9	Llama-2-13b	8	7	Llama-2-13b
7	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
8	Llama-2-13b	7	8	Llama-3-70b	6	8	Tie
9	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
10	Llama-3-70b	8	7	Llama-2-13b	8	7	Llama-2-13b
11	Llama-3-70b	8	6	Llama-3-70b	7	8	Llama-3-70b
12	Llama-3-70b	8	7	Llama-3-70b	7	9	Llama-3-70b
13	Llama-3-70b	8	6	Llama-3-70b	7	8	Llama-3-70b
14	Llama-3-70b	8	7	Llama-2-13b	8	7	Tie
15	Llama-3-70b	9	7	Llama-2-13b	8	7	Tie
16	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
17	Llama-2-13b	7	8	Llama-3-70b	7	9	Tie
18	Llama-3-70b	9	8	Llama-3-70b	7	8	Llama-3-70b
19	Llama-3-70b	9	7	Llama-2-13b	8	7	Tie
20	Llama-2-13b	7	8	Llama-3-70b	6	8	Tie
21	Llama-3-70b	9	8	Llama-3-70b	7	8	Llama-3-70b
22	Llama-2-13b	6	8	Llama-3-70b	6	8	Ties
23	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
24	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
25	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b

C.13 Llama-2-13b vs. Vicuna-7b-v1.5

	Home			Away			
Topic	Winner	Side 1 Llama-2-13b	Side 2 Vicuna-7b-v1.5	Winner	Side 1 Vicuna-7b-v1.5	Side 2 Llama-2-13b	Overall
1	Llama-2-13b	8	7	Vicuna-7b-v1.5	8	7	Tie
2	Vicuna-7b-v1.5	7	8	Llama-2-13b	7	8	Tie
3	Llama-2-13b	9	8	Vicuna-7b-v1.5	8	8	Tie
4	Llama-2-13b	8	7	Vicuna-7b-v1.5	8	7	Tie
5	Llama-2-13b	8	7	Vicuna-7b-v1.5	8	7	Tie
6	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
7	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
8	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
9	Llama-2-13b	8	7	Vicuna-7b-v1.5	8	7	Tie
10	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
11	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
12	Llama-2-13b	8	7	Llama-2-13b	6	8	Llama-2-13b
13	Llama-2-13b	9	7	Llama-2-13b	7	8	Llama-2-13b
14	Llama-2-13b	8	7	Vicuna-7b-v1.5	8	7	Tie
15	Llama-2-13b	8	6	Vicuna-7b-v1.5	8	7	Tie
16	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
17	Llama-2-13b	8	7	Llama-2-13b	6	9	Llama-2-13b
18	Llama-2-13b	8	6	Vicuna-7b-v1.5	8	7	Tie
19	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
20	Vicuna-7b-v1.5	6	8	Llama-2-13b	6	8	Tie
21	Llama-2-13b	8	7	Llama-2-13b	6	8	Llama-2-13b
22	Vicuna-7b-v1.5	6	8	Llama-2-13b	6	8	Tie
23	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
24	Llama-2-13b	8	7	Llama-2-13b	8	9	Llama-2-13b
25	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b

C.14 Llama-2-13b vs. Vicuna-13b-v1.5

	Home			Away			
Topic	Winner	Side 1 Llama-2-13b	Side 2 Vicuna-13b-v1.5	Winner	Side 1 Vicuna-13b-v1.5	Side 2 Llama-2-13b	Overall
1	Llama-2-13b	8	7	Vicuna-13b-v1.5	8	7	Tie
2	Vicuna-13b-v1.5	8	9	Llama-2-13b	7	8	Tie
3	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
4	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
5	Llama-2-13b	8	7	Vicuna-13b-v1.5	8	7	Tie
6	Vicuna-13b-v1.5	7	8	Llama-2-13b	7	8	Tie
7	Vicuna-13b-v1.5	8	9	Llama-2-13b	7	8	Tie
8	Vicuna-13b-v1.5	7	8	Llama-2-13b	6	8	Tie
9	Llama-2-13b	8	7	Vicuna-13b-v1.5	8	7	Tie
10	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
11	Llama-2-13b	8	7	Vicuna-13b-v1.5	8	7	Tie
12	Vicuna-13b-v1.5	8	9	Llama-2-13b	6	8	Tie
13	Llama-2-13b	8	6	Vicuna-13b-v1.5	8	7	Tie
14	Llama-2-13b	8	7	Vicuna-13b-v1.5	8	7	Tie
15	Llama-2-13b	8	6	Vicuna-13b-v1.5	8	7	Tie
16	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
17	Vicuna-13b-v1.5	7	8	Llama-2-13b	6	8	Tie
18	Llama-2-13b	8	7	Vicuna-13b-v1.5	8	7	Tie
19	Llama-2-13b	9	7	Llama-2-13b	7	8	Llama-2-13b
20	Vicuna-13b-v1.5	7	8	Llama-2-13b	6	8	Tie
21	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
22	Vicuna-13b-v1.5	6	8	Llama-2-13b	7	8	Tie
23	Llama-2-13b	8	7	Llama-2-13b	7	8	Llama-2-13b
24	Llama-2-13b	8	7	Vicuna-13b-v1.5	8	7	Tie
25	Llama-2-13b	8	7	Llama-2-13b	8	9	Llama-2-13b

C.15 Llama-2-13b vs Mixtral-8x7B

	Home			Away			
Topic	Winner	Side 1 Llama-2-13b	Side 2 Mixtral-8x7B	Winner	Side 1 Mixtral-8x7B	Side 2 Llama-2-13b	Overall
1	Llama-2-13b	8	7	Mixtral-8x7B	9	7	Tie
2	Mixtral-8x7B	7	8	Llama-2-13b	8	9	Tie
3	Llama-2-13b	8	7	Mixtral-8x7B	8	7	Tie
4	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B
5	Llama-2-13b	8	7	Mixtral-8x7B	8	7	Tie
6	Mixtral-8x7B	8	9	Llama-2-13b	7	8	Tie
7	Mixtral-8x7B	8	9	Llama-2-13b	7	8	Tie
8	Mixtral-8x7B	7	8	Llama-2-13b	6	8	Tie
9	Llama-2-13b	8	7	Mixtral-8x7B	8	7	Tie
10	Llama-2-13b	8	7	Mixtral-8x7B	8	7	Tie
11	Llama-2-13b	8	7	Mixtral-8x7B	8	7	Tie
12	Mixtral-8x7B	6	8	Llama-2-13b	8	9	Tie
13	Llama-2-13b	8	7	Mixtral-8x7B	8	7	Tie
14	Llama-2-13b	8	7	Mixtral-8x7B	8	7	Tie
15	Llama-2-13b	8	6	Mixtral-8x7B	8	7	Tie
16	Llama-2-13b	8	7	Mixtral-8x7B	8	7	Tie
17	Mixtral-8x7B	7	8	Llama-2-13b	7	8	Tie
18	Llama-2-13b	8	7	Mixtral-8x7B	8	7	Tie
19	Llama-2-13b	8	7	Mixtral-8x7B	8	7	Tie
20	Mixtral-8x7B	6	8	Llama-2-13b	6	9	Tie
21	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B
22	Mixtral-8x7B	6	8	Llama-2-13b	6	8	Tie
23	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
24	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
25	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B

C.16 Llama-2-13b vs. GPT 3.5

	Home			Away			
Topic	Winner	Side 1 Llama-2-13b	Side 2 GPT 3.5	Winner	Side 1 GPT 3.5	Side 2 Llama-2-13b	Overall
1	Llama-2-13b	8	7	GPT 3.5	9	7	Tie
2	GPT 3.5	7	9	Llama-2-13b	8	9	Tie
3	Llama-2-13b	8	7	GPT 3.5	8	7	Tie
4	Llama-2-13b	8	7	GPT 3.5	8	7	Tie
5	GPT 3.5	8	9	GPT 3.5	8	7	GPT 3.5
6	GPT 3.5	8	9	Llama-2-13b	7	8	Tie
7	GPT 3.5	8	9	Llama-2-13b	7	8	Tie
8	GPT 3.5	7	9	GPT 3.5	8	7	GPT 3.5
9	Llama-2-13b	8	7	GPT 3.5	8	7	Tie
10	Llama-2-13b	8	7	GPT 3.5	8	7	Tie
11	GPT 3.5	8	9	Llama-2-13b	7	8	Tie
12	GPT 3.5	8	9	Llama-2-13b	7	8	Tie
13	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
14	tie	8	8	GPT 3.5	8	7	GPT 3.5
15	Llama-2-13b	8	7	GPT 3.5	9	7	Tie
16	GPT 3.5	8	9	GPT 3.5	8	7	GPT 3.5
17	GPT 3.5	6	8	Llama-2-13b	8	9	Tie
18	Llama-2-13b	8	7	GPT 3.5	8	7	Tie
19	Llama-2-13b	8	7	GPT 3.5	8	7	Tie
20	GPT 3.5	6	8	Llama-2-13b	6	9	Tie
21	GPT 3.5	8	9	Llama-2-13b	7	8	Tie
22	GPT 3.5	6	8	Llama-2-13b	7	8	Tie
23	GPT 3.5	8	9	GPT 3.5	8	7	GPT 3.5
24	GPT 3.5	7	8	Llama-2-13b	8	9	Tie
25	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5

C.17 Llama-2-13b vs. GPT 4

	Home			Away			
Topic	Winner	Side 1 Llama-2-13b	Side 2 GPT-4	Winner	Side 1 GPT-4	Side 2 Llama-2-13b	Overall
1	GPT-4	8	9	GPT-4	9	6	GPT-4
2	GPT-4	6	9	Llama-2-13b	8	9	Tie
3	GPT-4	8	9	GPT-4	9	8	GPT-4
4	GPT-4	7	8	GPT-4	8	7	GPT-4
5	GPT-4	8	9	GPT-4	9	8	GPT-4
6	GPT-4	7	9	GPT-4	8	7	GPT-4
7	GPT-4	7	9	GPT-4	8	7	GPT-4
8	GPT-4	7	8	GPT-4	8	7	GPT-4
9	GPT-4	8	9	GPT-4	8	7	GPT-4
10	GPT-4	8	9	GPT-4	8	7	GPT-4
11	GPT-4	7	8	GPT-4	9	7	GPT-4
12	GPT-4	7	9	Llama-2-13b	8	9	Tie
13	GPT-4	7	8	GPT-4	8	7	GPT-4
14	GPT-4	7	9	GPT-4	9	8	GPT-4
15	GPT-4	8	9	GPT-4	9	7	GPT-4
16	GPT-4	8	9	GPT-4	8	7	GPT-4
17	GPT-4	6	9	GPT-4	8	7	GPT-4
18	GPT-4	8	9	GPT-4	9	7	GPT-4
19	GPT-4	8	9	GPT-4	8	7	GPT-4
20	GPT-4	6	9	Llama-2-13b	8	9	Tie
21	GPT-4	7	9	GPT-4	8	7	GPT-4
22	GPT-4	5	8	GPT-4	8	7	GPT-4
23	GPT-4	7	8	GPT-4	8	7	GPT-4
24	GPT-4	7	8	GPT-4	8	7	GPT-4
25	GPT-4	8	9	GPT-4	8	7	GPT-4

C.18 Llama-2-70b vs. Llama-2-70b

	Home			Away		
Topic	Winner	Side 1 Llama-2-70b	Side 2 Llama-2-70b	Winner	Side 1 Llama-2-70b	Side 2 Llama-2-70b
1	1	9	7	1	8	7
2	2	7	8	2	7	8
3	2	8	9	2	8	9
4	1	8	7	1	8	7
5	1	8	7	1	8	7
6	2	8	9	2	8	9
7	2	8	9	2	8	9
8	2	8	9	1	8	7
9	1	9	7	1	8	7
10	2	8	9	2	8	9
11	1	8	7	1	8	7
12	2	8	9	2	8	9
13	1	8	7	1	9	7
14	1	8	7	1	8	7
15	1	8	7	1	9	7
16	1	8	7	1	9	8
17	2	7	8	2	8	9
18	1	8	7	1	8	7
19	1	8	7	1	8	7
20	2	6	8	2	7	8
21	1	8	7	2	8	9
22	2	7	8	2	7	8
23	1	8	7	1	8	7
24	1	8	7	1	8	7
25	2	8	9	2	8	9

C.19 Llama-2-70b vs. Llama-3-70b

	Home			Away			
Topic	Winner	Side 1 Llama-3-70b	Side 2 Llama-2-70b	Winner	Side 1 Llama-2-70b	Side 2 Llama-3-70b	Overall
1	Llama-3-70b	9	7	Llama-2-70b	8	7	Tie
2	Llama-2-70b	8	9	Llama-3-70b	7	9	Tie
3	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
4	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
5	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
6	Llama-2-70b	8	9	Llama-3-70b	7	8	Tie
7	Llama-2-70b	7	9	Llama-3-70b	7	8	Tie
8	Llama-2-70b	7	8	Llama-3-70b	7	8	Tie
9	Llama-3-70b	8	7	Llama-2-70b	8	7	Tie
10	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
11	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
12	Llama-2-70b	8	9	Llama-3-70b	7	8	Tie
13	Llama-3-70b	9	7	Llama-3-70b	8	9	Llama-3-70b
14	Llama-3-70b	8	7	Llama-2-70b	8	7	Tie
15	Llama-3-70b	9	6	Llama-2-70b	9	8	Tie
16	Llama-2-70b	8	9	Llama-2-70b	8	7	Tie
17	Llama-2-70b	7	8	Llama-3-70b	7	8	Tie
18	Llama-3-70b	8	7	Llama-2-70b	8	7	Tie
19	Llama-3-70b	8	7	Llama-2-70b	8	7	Tie
20	Llama-2-70b	8	9	Llama-3-70b	7	9	Tie
21	Tie	8	8	Llama-3-70b	8	9	Llama-3-70b
22	Llama-2-70b	7	8	Llama-3-70b	6	8	Tie
23	Llama-3-70b	8	7	Llama-2-70b	8	7	Tie
24	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
25	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b

C.20 Llama-2-70b vs. Vicuna-7b-v1.5

	Home			Away			
Topic	Winner	Side 1 Llama-2-70b	Side 2 Vicuna-7b-v1.5	Winner	Side 1 Vicuna-7b-v1.5	Side 2 Llama-2-70b	Overall
1	Llama-2-70b	9	7	Vicuna-7b-v1.5	8	7	Tie
2	Vicuna-7b-v1.5	8	9	Llama-2-70b	7	8	Tie
3	Llama-2-70b	9	8	Llama-2-70b	8	9	Llama-2-70b
4	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
5	Llama-2-70b	9	8	Vicuna-7b-v1.5	8	7	Tie
6	Llama-2-70b	8	7	Llama-2-70b	8	9	Llama-2-70b
7	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
8	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
9	Llama-2-70b	8	7	Llama-2-70b	7	9	Llama-2-70b
10	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
11	Llama-2-70b	8	7	Vicuna-7b-v1.5	8	7	Tie
12	Llama-2-70b	8	7	Llama-2-70b	6	8	Llama-2-70b
13	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
14	Llama-2-70b	9	7	Vicuna-7b-v1.5	8	7	Tie
15	Llama-2-70b	8	6	Vicuna-7b-v1.5	8	7	Tie
16	Llama-2-70b	9	8	Llama-2-70b	7	9	Llama-2-70b
17	Vicuna-7b-v1.5	7	8	Llama-2-70b	7	8	Tie
18	Llama-2-70b	9	7	Vicuna-7b-v1.5	8	7	Tie
19	Llama-2-70b	9	7	Vicuna-7b-v1.5	8	7	Tie
20	Vicuna-7b-v1.5	7	8	Llama-2-70b	6	8	Tie
21	Llama-2-70b	8	7	Llama-2-70b	8	9	Llama-2-70b
22	Vicuna-7b-v1.5	7	8	Llama-2-70b	5	8	Tie
23	Llama-2-70b	8	7	Llama-2-70b	6	8	Llama-2-70b
24	Llama-2-70b	8	7	Vicuna-7b-v1.5	8	7	Tie
25	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b

C.21 Llama-2-70b vs. Vicuna-13b-v1.5

	Home			Away			
Topic	Winner	Side 1 Llama-2-70b	Side 2 Vicuna-13b-v1.5	Winner	Side 1 Vicuna-13b-v1.5	Side 2 Llama-2-70b	Overall
1	Llama-2-70b	8	7	Vicuna-13b-v1.5	8	7	Tie
2	Vicuna-13b-v1.5	7	8	Llama-2-70b	7	8	Tie
3	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
4	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
5	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
6	Llama-2-70b	8	7	Llama-2-70b	7	9	Llama-2-70b
7	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
8	Vicuna-13b-v1.5	7	8	Llama-2-70b	6	8	Tie
9	Llama-2-70b	8	7	Vicuna-13b-v1.5	8	7	Tie
10	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
11	Llama-2-70b	8	7	Vicuna-13b-v1.5	8	7	Tie
12	Vicuna-13b-v1.5	8	9	Llama-2-70b	6	9	Tie
13	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
14	Llama-2-70b	8	7	Vicuna-13b-v1.5	8	7	Tie
15	Llama-2-70b	9	7	Vicuna-13b-v1.5	8	7	Tie
16	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
17	Vicuna-13b-v1.5	7	8	Llama-2-70b	7	8	Tie
18	Llama-2-70b	8	7	Vicuna-13b-v1.5	8	7	Tie
19	Llama-2-70b	9	7	Vicuna-13b-v1.5	8	7	Tie
20	Vicuna-13b-v1.5	7	8	Llama-2-70b	6	8	Tie
21	Vicuna-13b-v1.5	8	9	Llama-2-70b	7	8	Tie
22	Llama-2-70b	8	7	Llama-2-70b	6	7	Llama-2-70b
23	Llama-2-70b	8	7	Llama-2-70b	7	8	Llama-2-70b
24	Llama-2-70b	8	7	Llama-2-70b	8	9	Llama-2-70b
25	Vicuna-13b-v1.5	7	8	Llama-2-70b	7	8	Tie

C.22 Llama-2-70b vs. Mixtral-8x7B

	Home			Away			
Topic	Winner	Side 1 Llama-2-70b	Side 2 Mixtral-8x7B	Winner	Side 1 Mixtral-8x7B	Side 2 Llama-2-70b	Overall
1	Llama-2-70b	9	7	Mixtral-8x7B	8	7	Tie
2	Mixtral-8x7B	8	9	Llama-2-70b	7	8	Tie
3	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Tie
4	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Tie
5	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Tie
6	Mixtral-8x7B	8	9	Llama-2-70b	8	9	Tie
7	Llama-2-70b	8	7	Llama-2-70b	8	9	Llama-2-70b
8	Mixtral-8x7B	8	9	Llama-2-70b	7	8	Tie
9	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Tie
10	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Tie
11	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Tie
12	Mixtral-8x7B	8	9	Llama-2-70b	8	9	Tie
13	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
14	tie	8	8	Mixtral-8x7B	8	7	Mixtral-8x7B
15	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Mixtral-8x7B
16	Mixtral-8x7B	8	9	Llama-2-70b	8	9	Tie
17	Mixtral-8x7B	7	8	Llama-2-70b	7	8	Tie
18	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Tie
19	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Tie
20	Mixtral-8x7B	7	8	Llama-2-70b	7	9	Tie
21	tie	8	8	Llama-2-70b	8	9	Llama-2-70b
22	Mixtral-8x7B	6	8	Llama-2-70b	6	8	Tie
23	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Tie
24	Llama-2-70b	8	7	Mixtral-8x7B	8	7	Tie
25	Mixtral-8x7B	8	9	Llama-2-70b	8	9	Tie

C.23 Llama-2-70b vs. GPT 3.5

	Home			Away			
Topic	Winner	Side 1 Llama-2-70b	Side 2 GPT 3.5	Winner	Side 1 GPT 3.5	Side 2 Llama-2-70b	Overall
1	GPT 3.5	7	9	Llama-2-70b	8	9	Tie
2	Llama-2-70b	8	7	GPT 3.5	8	7	Tie
3	Llama-2-70b	9	8	Llama-2-70b	7	8	Llama-2-70b
4	GPT 3.5	7	8	Llama-2-70b	7	8	Tie
5	GPT 3.5	8	9	GPT 3.5	9	8	GPT 3.5
6	Llama-2-70b	9	8	GPT 3.5	8	7	Tie
7	Llama-2-70b	9	8	GPT 3.5	8	7	Tie
8	Llama-2-70b	8	7	GPT 3.5	8	7	Tie
9	GPT 3.5	8	9	Llama-2-70b	7	8	Tie
10	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
11	GPT 3.5	7	8	Llama-2-70b	7	8	Tie
12	Llama-2-70b	8	7	GPT 3.5	8	6	Tie
13	GPT 3.5	7	8	Llama-2-70b	7	8	Tie
14	Llama-2-70b	9	8	GPT 3.5	9	8	Tie
15	GPT 3.5	7	8	Llama-2-70b	7	8	Tie
16	Llama-2-70b	9	8	GPT 3.5	8	7	Tie
17	Llama-2-70b	9	8	GPT 3.5	8	6	Tie
18	GPT 3.5	7	8	Llama-2-70b	7	8	Tie
19	GPT 3.5	7	8	Llama-2-70b	7	8	Tie
20	Llama-2-70b	8	6	GPT 3.5	8	6	Tie
21	Llama-2-70b	9	8	GPT 3.5	8	7	Tie
22	Llama-2-70b	8	7	GPT 3.5	8	6	Tie
23	Llama-2-70b	9	8	GPT 3.5	8	7	Tie
24	Llama-2-70b	9	8	GPT 3.5	8	7	Tie
25	Llama-2-70b	9	8	GPT 3.5	9	8	Tie

C.24 Llama-2-70b vs. GPT 4

	Home			Away			
Topic	Winner	Side 1 Llama-2-70b	Side 2 GPT 4	Winner	Side 1 GPT 4	Side 2 Llama-2-70b	Over
1	Llama-2-70b	8	7	GPT 4	9	7	Tie
2	GPT 4	8	9	Llama-2-70b	8	9	Tie
3	GPT 4	7	8	GPT 4	9	8	GPT 4
4	GPT 4	7	8	GPT 4	8	7	GPT 4
5	GPT 4	8	9	GPT 4	8	7	GPT 4
6	GPT 4	8	9	GPT 4	8	7	GPT 4
7	GPT 4	7	8	GPT 4	8	7	GPT 4
8	GPT 4	7	8	GPT 4	8	7	GPT 4
9	GPT 4	8	9	GPT 4	9	7	GPT 4
10	GPT 4	7	8	GPT 4	8	7	GPT 4
11	GPT 4	7	8	GPT 4	8	7	GPT 4
12	GPT 4	7	9	Llama-2-70b	7	9	Tie
13	GPT 4	7	8	GPT 4	9	7	GPT 4
14	GPT 4	8	9	GPT 4	9	7	GPT 4
15	GPT 4	8	9	GPT 4	8	7	GPT 4
16	GPT 4	8	9	Llama-2-70b	8	9	Tie
17	GPT 4	7	8	tie	8	8	GPT 4
18	GPT 4	8	9	GPT 4	8	7	GPT 4
19	GPT 4	8	9	GPT 4	8	7	GPT 4
20	GPT 4	7	9	GPT 4	8	7	GPT 4
21	GPT 4	7	9	GPT 4	8	7	GPT 4
22	GPT 4	6	9	GPT 4	8	7	GPT 4
23	GPT 4	8	9	GPT 4	8	7	GPT 4
24	GPT 4	7	9	GPT 4	8	7	GPT 4
25	GPT 4	8	9	GPT 4	9	8	GPT 4

C.25 Llama-3-70b vs. Vicuna-7b-v1.5

	Home			Away			
Topic	Winner	Side 1 Llama-3-70b	Side 2 Vicuna-7b-v1.5	Winner	Side 1 Vicuna-7b-v1.5	Side 2 Llama-3-70b	Overall
1	Llama-3-70b	8	7	Llama-3-70b	4	7	Llama-3-70b
2	Llama-3-70b	8	7	Llama-3-70b	7	9	Llama-3-70b
3	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
4	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
5	Llama-3-70b	9	8	Llama-3-70b	8	9	Llama-3-70b
6	Llama-3-70b	8	7	Llama-3-70b	6	8	Llama-3-70b
7	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
8	Llama-3-70b	8	7	Llama-3-70b	6	8	Llama-3-70b
9	Llama-3-70b	9	7	Llama-3-70b	8	9	Llama-3-70b
10	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
11	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
12	Llama-3-70b	9	7	Llama-3-70b	6	9	Llama-3-70b
13	Llama-3-70b	9	7	Llama-3-70b	7	8	Llama-3-70b
14	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
15	Llama-3-70b	8	6	Llama-3-70b	7	8	Llama-3-70b
16	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
17	Vicuna-7b-v1.5	7	8	Llama-3-70b	6	9	Tie
18	Llama-3-70b	9	7	Llama-3-70b	7	8	Llama-3-70b
19	Llama-3-70b	8	7	Vicuna-7b-v1.5	8	7	Tie
20	Llama-3-70b	8	7	Llama-3-70b	6	9	Llama-3-70b
21	Llama-3-70b	8	7	Llama-3-70b	6	8	Llama-3-70b
22	Llama-3-70b	8	7	Llama-3-70b	6	9	Llama-3-70b
23	Llama-3-70b	8	6	Llama-3-70b	6	8	Llama-3-70b
24	Llama-3-70b	8	7	Llama-3-70b	7	9	Llama-3-70b
25	Llama-3-70b	8	7	Llama-3-70b	7	9	Llama-3-70b

C.26 Llama-3-70b vs. Vicuna-13b-v1.5

	Home			Away			
Topic	Winner	Side 1 Llama-3-70b	Side 2 Vicuna-13b-v1.5	Winner	Side 1 Vicuna-13b-v1.5	Side 2 Llama-3-70b	Overall
1	Llama-3-70b	9	7	Llama-3-70b	8	9	Llama-3-70b
2	Vicuna-13b-v1.5	7	8	Llama-3-70b	7	8	Tie
3	Vicuna-13b-v1.5	8	9	Llama-3-70b	7	8	Tie
4	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
5	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
6	Vicuna-13b-v1.5	8	9	Llama-3-70b	7	8	Tie
7	Llama-3-70b	8	7	Llama-3-70b	7	9	Llama-3-70b
8	Llama-3-70b	8	7	Llama-3-70b	6	8	Llama-3-70b
9	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
10	tie	8	8	Llama-3-70b	7	8	Tie
11	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
12	Vicuna-13b-v1.5	8	9	Llama-3-70b	6	9	Tie
13	Llama-3-70b	8	7	Llama-3-70b	7	9	Llama-3-70b
14	Llama-3-70b	9	8	Llama-3-70b	7	8	Llama-3-70b
15	Llama-3-70b	9	7	Llama-3-70b	7	8	Llama-3-70b
16	tie	8	8	Llama-3-70b	7	9	Llama-3-70b
17	Vicuna-13b-v1.5	7	8	Llama-3-70b	6	9	Tie
18	Llama-3-70b	9	7	Llama-3-70b	7	8	Llama-3-70b
19	Llama-3-70b	9	7	Llama-3-70b	8	9	Llama-3-70b
20	Vicuna-13b-v1.5	7	8	Llama-3-70b	6	9	Tie
21	tie	8	8	Llama-3-70b	7	9	Llama-3-70b
22	Vicuna-13b-v1.5	7	8	Llama-3-70b	6	8	Tie
23	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
24	Llama-3-70b	9	8	Llama-3-70b	7	8	Llama-3-70b
25	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b

C.27 Llama-3-70b vs. Mixtral-8x7B

	Home			Away			
Topic	Winner	Side 1 Llama-3-70b	Side 2 Mixtral-8x7B	Winner	Side 1 Mixtral-8x7B	Side 2 Llama-3-70b	Overall
1	Llama-3-70b	9	8	Mixtral-8x7B	8	7	Tie
2	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
3	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
4	Llama-3-70b	8	7	Mixtral-8x7B	8	7	Tie
5	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
6	Mixtral-8x7B	8	9	Llama-3-70b	8	9	Tie
7	Mixtral-8x7B	7	8	Llama-3-70b	7	8	Tie
8	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
9	Mixtral-8x7B	8	9	Llama-3-70b	8	9	Tie
10	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
11	Llama-3-70b	8	7	Mixtral-8x7B	8	7	Tie
12	Mixtral-8x7B	7	9	Llama-3-70b	7	8	Tie
13	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
14	Llama-3-70b	8	7	Mixtral-8x7B	8	7	Tie
15	Llama-3-70b	9	8	Mixtral-8x7B	8	7	Tie
16	Mixtral-8x7B	8	9	Llama-3-70b	8	9	Tie
17	Mixtral-8x7B	8	9	Llama-3-70b	7	8	Tie
18	Llama-3-70b	9	8	Mixtral-8x7B	8	7	Tie
19	Llama-3-70b	8	7	Mixtral-8x7B	8	7	Tie
20	Mixtral-8x7B	7	8	Llama-3-70b	6	8	Tie
21	Mixtral-8x7B	7	8	Llama-3-70b	8	9	Tie
22	Mixtral-8x7B	7	8	Llama-3-70b	6	9	Tie
23	Mixtral-8x7B	7	8	Llama-3-70b	8	9	Tie
24	Mixtral-8x7B	8	9	Llama-3-70b	7	8	Tie
25	Mixtral-8x7B	7	8	Llama-3-70b	7	8	Tie

C.28 Llama-3-70b vs. GPT-3.5

	Home			Away			
Topic	Winner	Side 1 Llama-3-70b	Side 2 GPT-3.5	Winner	Side 1 GPT-3.5	Side 2 Llama-3-70b	Overall
1	Llama-3-70b	8	7	GPT-3.5	8	7	Tie
2	GPT-3.5	8	9	Llama-3-70b	7	8	Tie
3	GPT-3.5	8	9	Llama-3-70b	8	9	Tie
4	Llama-3-70b	8	7	GPT-3.5	8	7	Tie
5	tie	8	8	Llama-3-70b	8	9	Llama-3-70b
6	GPT-3.5	7	8	Llama-3-70b	7	8	Tie
7	GPT-3.5	8	9	Llama-3-70b	7	9	Tie
8	GPT-3.5	7	8	Llama-3-70b	6	8	Tie
9	Llama-3-70b	8	7	GPT-3.5	8	7	Tie
10	GPT-3.5	8	9	Llama-3-70b	8	9	Tie
11	Llama-3-70b	8	7	GPT-3.5	8	7	Tie
12	GPT-3.5	8	9	Llama-3-70b	6	9	Tie
13	Llama-3-70b	8	7	GPT-3.5	8	7	Tie
14	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
15	Llama-3-70b	9	8	GPT-3.5	8	7	Tie
16	Llama-3-70b	9	8	Llama-3-70b	7	8	Llama-3-70b
17	GPT-3.5	8	9	Llama-3-70b	7	9	Tie
18	Llama-3-70b	8	7	Llama-3-70b	8	9	Llama-3-70b
19	Llama-3-70b	8	7	GPT-3.5	8	7	Tie
20	GPT-3.5	7	8	Llama-3-70b	7	9	Tie
21	GPT-3.5	8	9	Llama-3-70b	7	8	Tie
22	GPT-3.5	7	9	Llama-3-70b	6	9	Tie
23	Llama-3-70b	8	7	Llama-3-70b	7	8	Llama-3-70b
24	GPT-3.5	8	9	Llama-3-70b	7	8	Tie
25	GPT-3.5	8	9	Llama-3-70b	8	9	Tie

C.29 Llama-3-70b vs. GPT-4

	Home			Away			
Topic	Winner	Side 1 Llama-3-70b	Side 2 GPT-4	Winner	Side 1 GPT-4	Side 2 Llama-3-70b	Overall
1	Llama-3-70b	8	7	GPT-4	9	7	Tie
2	GPT-4	7	9	Llama-3-70b	7	8	Tie
3	GPT-4	8	9	GPT-4	8	7	GPT-4
4	GPT-4	7	8	GPT-4	8	7	GPT-4
5	GPT-4	8	9	GPT-4	8	7	GPT-4
6	GPT-4	7	8	Llama-3-70b	8	9	Tie
7	GPT-4	8	9	GPT-4	8	7	GPT-4
8	GPT-4	7	8	GPT-4	8	7	GPT-4
9	GPT-4	8	9	GPT-4	8	7	GPT-4
10	GPT-4	8	9	Llama-3-70b	8	9	Tie
11	Llama-3-70b	8	7	GPT-4	8	7	Tie
12	GPT-4	7	8	Llama-3-70b	8	9	Tie
13	Llama-3-70b	8	7	GPT-4	8	7	Tie
14	GPT-4	8	9	GPT-4	9	8	GPT-4
15	GPT-4	8	9	GPT-4	9	7	GPT-4
16	GPT-4	8	9	Llama-3-70b	8	9	Tie
17	GPT-4	8	9	GPT-4	8	7	GPT-4
18	GPT-4	8	9	GPT-4	8	7	GPT-4
19	Llama-3-70b	8	7	GPT-4	9	8	Tie
20	GPT-4	7	9	Llama-3-70b	7	8	Tie
21	GPT-4	7	8	Llama-3-70b	8	9	Tie
22	GPT-4	7	9	GPT-4	8	7	GPT-4
23	GPT-4	7	9	GPT-4	8	7	GPT-4
24	GPT-4	8	9	Llama-3-70b	8	9	Tie
25	GPT-4	8	9	GPT-4	8	7	GPT-4

C.30 Vicuna-7b-v1.5 vs. Vicuna-7b-v1.5

	Home			Away		
Topic	Winner	Side 1 Vicuna-7b-v1.5	Side 2 Vicuna-7b-v1.5	Winner	Side 1 Vicuna-7b-v1.5	Side 2 Vicuna-7b-v1.5
1	1	8	7	1	8	7
2	2	7	8	2	7	8
3	1	8	7	1	8	7
4	1	8	7	1	8	7
5	1	8	7	1	8	7
6	1	8	7	2	7	8
7	1	8	7	2	8	9
8	2	7	8	2	7	8
9	1	8	7	1	8	7
10	1	8	7	1	8	7
11	1	8	7	1	8	7
12	2	6	8	2	7	8
13	2	7	8	2	7	8
14	2	7	8	1	8	7
15	1	8	7	1	8	7
16	1	8	7	1	8	7
17	2	8	9	2	7	8
18	1	8	7	1	8	7
19	1	8	7	1	8	7
20	2	7	8	2	6	8
21	1	8	7	1	8	7
22	2	7	8	2	6	8
23	1	8	7	1	8	7
24	1	8	7	1	8	7
25	2	8	9	2	8	9

C.31 Vicuna-7b-v1.5 vs. Vicuna-13b-v1.5

	Home			Away			
Topic	Winner	Side 1 Vicuna-7b-v1.5	Side 2 Vicuna-13b-v1.5	Winner	Side 1 Vicuna-13b-v1.5	Side 2 Vicuna-7b-v1.5	Overall
1	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	7	Tie
2	Vicuna-13b-v1.5	7	8	Vicuna-7b-v1.5	7	8	Tie
3	Vicuna-13b-v1.5	8	9	Vicuna-13b-v1.5	8	7	Vicuna-13b-v1.5
4	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	7	Tie
5	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	7	Tie
6	Vicuna-13b-v1.5	6	8	Vicuna-7b-v1.5	7	8	Tie
7	Vicuna-13b-v1.5	8	9	Vicuna-7b-v1.5	8	9	Tie
8	Vicuna-13b-v1.5	7	8	Vicuna-7b-v1.5	7	8	Tie
9	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	7	Tie
10	Vicuna-13b-v1.5	8	9	tie	8	8	Vicuna-13b-v1.5
11	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	7	Tie
12	Vicuna-13b-v1.5	6	8	Vicuna-7b-v1.5	6	8	Tie
13	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	7	Tie
14	Vicuna-13b-v1.5	8	9	Vicuna-13b-v1.5	8	7	Vicuna-13b-v1.5
15	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	7	Tie
16	Vicuna-13b-v1.5	7	8	Vicuna-7b-v1.5	8	9	Tie
17	Vicuna-13b-v1.5	7	8	Vicuna-7b-v1.5	7	8	Tie
18	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	6	Tie
19	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	7	Tie
20	Vicuna-13b-v1.5	7	8	Vicuna-7b-v1.5	6	7	Tie
21	Vicuna-13b-v1.5	7	8	Vicuna-7b-v1.5	7	8	Tie
22	Vicuna-13b-v1.5	6	8	Vicuna-7b-v1.5	7	8	Tie
23	Vicuna-13b-v1.5	7	8	Vicuna-7b-v1.5	7	8	Tie
24	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	7	Tie
25	Vicuna-7b-v1.5	8	7	Vicuna-13b-v1.5	8	7	Tie

C.32 Vicuna-7b-v1.5 vs. Mixtral-8x7B

	Home			Away			
Topic	Winner	Side 1 Vicuna-7b-v1.5	Side 2 Mixtral-8x7B	Winner	Side 1 Mixtral-8x7B	Side 2 Vicuna-7b-v1.5	Overall
1	Vicuna-7b-v1.5	8	7	Mixtral-8x7B	9	7	Tie
2	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
3	Vicuna-7b-v1.5	8	7	Mixtral-8x7B	8	7	Tie
4	Vicuna-7b-v1.5	8	7	Mixtral-8x7B	8	7	Tie
5	tie	8	8	Mixtral-8x7B	9	7	Mixtral-8x7B
6	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
7	Mixtral-8x7B	7	8	Vicuna-7b-v1.5	8	9	Tie
8	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
9	Vicuna-7b-v1.5	8	7	Mixtral-8x7B	9	7	Tie
10	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
11	Vicuna-7b-v1.5	8	7	Mixtral-8x7B	8	6	Tie
12	Mixtral-8x7B	6	8	Mixtral-8x7B	8	7	Mixtral-8x7B
13	Vicuna-7b-v1.5	8	7	Mixtral-8x7B	8	6	Tie
14	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B
15	Vicuna-7b-v1.5	8	7	Mixtral-8x7B	8	7	Tie
16	Mixtral-8x7B	7	8	Vicuna-7b-v1.5	8	9	Tie
17	Mixtral-8x7B	7	9	Vicuna-7b-v1.5	8	9	Tie
18	Vicuna-7b-v1.5	8	7	Mixtral-8x7B	8	7	Tie
19	Vicuna-7b-v1.5	8	7	Mixtral-8x7B	8	7	Tie
20	Mixtral-8x7B	7	9	Vicuna-7b-v1.5	7	8	Tie
21	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
22	Mixtral-8x7B	6	8	Vicuna-7b-v1.5	7	8	Tie
23	Vicuna-7b-v1.5	8	7	Mixtral-8x7B	8	7	Tie
24	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
25	Mixtral-8x7B	7	8	Mixtral-8x7B	8	8	Mixtral-8x7B

C.33 Vicuna-7b-v1.5 vs. GPT 3.5

	Home			Away			
Topic	Winner	Side 1 Vicuna-7b-v1.5	Side 2 GPT 3.5	Winner	Side 1 GPT 3.5	Side 2 Vicuna-7b-v1.5	Overall
1	Vicuna-7b-v1.5	8	7	GPT 3.5	9	8	Tie
2	GPT 3.5	7	8	Vicuna-7b-v1.5	7	8	Tie
3	GPT 3.5	8	7	GPT 3.5	8	7	GPT 3.5
4	GPT 3.5	8	7	GPT 3.5	8	7	GPT 3.5
5	GPT 3.5	8	8	GPT 3.5	8	7	GPT 3.5
6	GPT 3.5	7	8	tie	8	8	GPT 3.5
7	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
8	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
9	GPT 3.5	8	7	GPT 3.5	8	7	GPT 3.5
10	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
11	GPT 3.5	8	7	GPT 3.5	8	7	GPT 3.5
12	GPT 3.5	6	8	Vicuna-7b-v1.5	7	8	Tie
13	GPT 3.5	8	7	GPT 3.5	8	7	GPT 3.5
14	GPT 3.5	8	9	GPT 3.5	8	7	GPT 3.5
15	Vicuna-7b-v1.5	8	7	GPT 3.5	9	6	Tie
16	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
17	GPT 3.5	7	9	Vicuna-7b-v1.5	7	8	Tie
18	Vicuna-7b-v1.5	8	7	GPT 3.5	8	7	Tie
19	Vicuna-7b-v1.5	8	7	GPT 3.5	8	7	Tie
20	GPT 3.5	7	9	GPT 3.5	8	7	GPT 3.5
21	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
22	GPT 3.5	6	8	GPT 3.5	8	7	GPT 3.5
23	GPT 3.5	8	7	GPT 3.5	8	7	GPT 3.5
24	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
25	GPT 3.5	7	8	tie	8	8	GPT 3.5

C.34 Vicuna-7b-v1.5 vs. GPT 4

	Home			Away			
Topic	Winner	Side 1 Vicuna-7b-v1.5	Side 2 GPT 4	Winner	Side 1 GPT 4	Side 2 Vicuna-7b-v1.5	Overall
1	GPT 4	8	9	GPT 4	9	7	GPT 4
2	GPT 4	7	9	GPT 4	9	7	GPT 4
3	GPT 4	8	9	GPT 4	9	7	GPT 4
4	GPT 4	7	9	GPT 4	9	8	GPT 4
5	GPT 4	7	9	GPT 4	9	8	GPT 4
6	GPT 4	7	9	GPT 4	9	7	GPT 4
7	GPT 4	7	9	Vicuna-7b-v1.5	8	9	Tie
8	GPT 4	6	8	GPT 4	8	7	GPT 4
9	GPT 4	7	8	GPT 4	9	7	GPT 4
10	GPT 4	7	8	GPT 4	8	7	GPT 4
11	GPT 4	7	9	GPT 4	8	7	GPT 4
12	GPT 4	6	9	GPT 4	8	7	GPT 4
13	GPT 4	7	8	GPT 4	9	7	GPT 4
14	GPT 4	7	8	GPT 4	9	8	GPT 4
15	GPT 4	7	9	GPT 4	9	6	GPT 4
16	GPT 4	7	9	GPT 4	9	8	GPT 4
17	GPT 4	7	9	GPT 4	8	7	GPT 4
18	GPT 4	8	9	GPT 4	9	7	GPT 4
19	GPT 4	8	9	GPT 4	9	8	GPT 4
20	GPT 4	6	9	GPT 4	8	7	GPT 4
21	GPT 4	7	8	GPT 4	8	7	GPT 4
22	GPT 4	6	9	GPT 4	8	7	GPT 4
23	GPT 4	7	9	GPT 4	9	8	GPT 4
24	GPT 4	7	9	GPT 4	9	7	GPT 4
25	GPT 4	7	9	GPT 4	8	7	GPT 4

C.35 Vicuna-13b-v1.5 vs. Vicuna-13b-v1.5

	Home			Away		
Topic	Winner	Side 1 Vicuna-13b-v1.5	Side 2 Vicuna-13b-v1.5	Winner	Side 1 Vicuna-13b-v1.5	Side 2 Vicuna-13b-v1.5
1	1	9	7	1	9	7
2	2	7	8	2	7	8
3	1	8	7	1	8	7
4	1	8	7	1	8	7
5	1	8	7	1	8	7
6	2	8	9	2	7	8
7	2	7	8	2	7	8
8	2	7	8	2	7	8
9	1	8	7	1	8	7
10	1	8	7	1	8	7
11	1	8	7	1	8	7
12	2	7	9	2	6	8
13	1	8	7	1	8	7
14	1	8	7	1	8	7
15	1	8	7	1	8	7
16	1	8	7	1	8	7
17	2	7	8	2	7	8
18	1	8	7	1	8	7
19	1	8	7	1	8	6
20	2	6	8	2	6	8
21	1	8	7	1	8	7
22	2	6	7	2	7	8
23	1	8	7	1	8	7
24	2	7	8	2	7	8
25	2	8	9	1	8	8

C.36 Vicuna-13b-v1.5 vs. Mixtral-8x7B

	Home			Away			
Topic	Winner	Side 1 Vicuna-13b-v1.5	Side 2 Mixtral-8x7B	Winner	Side 1 Mixtral-8x7B	Side 2 Vicuna-13b-v1.5	Overall
1	Vicuna-13b-v1.5	9	8	Mixtral-8x7B	9	8	Tie
2	Mixtral-8x7B	7	9	Vicuna-13b-v1.5	8	9	Tie
3	Mixtral-8x7B	7	8	Mixtral-8x7B	9	8	Mixtral-8x7B
4	Mixtral-8x7B	7	8	Mixtral-8x7B	9	7	Mixtral-8x7B
5	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
6	Mixtral-8x7B	7	8	Vicuna-13b-v1.5	7	8	Tie
7	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
8	Mixtral-8x7B	7	8	Vicuna-13b-v1.5	7	8	Tie
9	Mixtral-8x7B	8	9	Mixtral-8x7B	9	7	Mixtral-8x7B
10	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B
11	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
12	Mixtral-8x7B	7	8	Vicuna-13b-v1.5	7	8	Tie
13	Vicuna-13b-v1.5	8	7	Mixtral-8x7B	8	7	Tie
14	Mixtral-8x7B	8	9	Mixtral-8x7B	9	8	Mixtral-8x7B
15	Mixtral-8x7B	7	8	Mixtral-8x7B	8	5	Mixtral-8x7B
16	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
17	Mixtral-8x7B	6	8	Vicuna-13b-v1.5	7	8	Tie
18	Vicuna-13b-v1.5	8	7	Mixtral-8x7B	8	7	Tie
19	Vicuna-13b-v1.5	8	7	Mixtral-8x7B	9	8	Tie
20	Mixtral-8x7B	6	8	Vicuna-13b-v1.5	7	8	Tie
21	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B
22	Mixtral-8x7B	6	8	Vicuna-13b-v1.5	7	8	Tie
23	Mixtral-8x7B	7	8	Mixtral-8x7B	8	7	Mixtral-8x7B
24	Mixtral-8x7B	8	9	Vicuna-13b-v1.5	8	9	Tie
25	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B

C.37 Vicuna-13b-v1.5 vs. GPT 3.5

	Home			Away			
Topic	Winner	Side 1 Vicuna-13b-v1.5	Side 2 GPT 3.5	Winner	Side 1 GPT 3.5	Side 2 Vicuna-13b-v1.5	Overall
1	Vicuna-13b-v1.5	8	7	GPT 3.5	8	7	Tie
2	GPT 3.5	7	9	Vicuna-13b-v1.5	7	8	Tie
3	GPT 3.5	8	9	Vicuna-13b-v1.5	8	9	Tie
4	Vicuna-13b-v1.5	8	7	GPT 3.5	8	7	Tie
5	GPT 3.5	8	9	GPT 3.5	8	7	GPT 3.5
6	GPT 3.5	7	9	Vicuna-13b-v1.5	7	8	Tie
7	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
8	GPT 3.5	6	9	GPT 3.5	8	7	GPT 3.5
9	GPT 3.5	8	9	GPT 3.5	8	7	GPT 3.5
10	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
11	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
12	GPT 3.5	6	8	Vicuna-13b-v1.5	7	8	Tie
13	GPT 3.5	7	8	GPT 3.5	8	6	GPT 3.5
14	GPT 3.5	8	9	GPT 3.5	9	7	GPT 3.5
15	Vicuna-13b-v1.5	8	7	GPT 3.5	9	7	Tie
16	Vicuna-13b-v1.5	8	7	GPT 3.5	8	7	Tie
17	GPT 3.5	7	9	Vicuna-13b-v1.5	7	8	Tie
18	GPT 3.5	8	9	GPT 3.5	9	8	GPT 3.5
19	Vicuna-13b-v1.5	8	7	GPT 3.5	9	7	Tie
20	GPT 3.5	6	8	Vicuna-13b-v1.5	6	8	Tie
21	GPT 3.5	8	9	GPT 3.5	8	7	GPT 3.5
22	GPT 3.5	7	8	Vicuna-13b-v1.5	7	8	Tie
23	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
24	GPT 3.5	7	8	GPT 3.5	8	7	GPT 3.5
25	GPT 3.5	8	9	tie	8	8	GPT 3.5

C.38 Vicuna-13b-v1.5 vs. GPT 4

	Home			Away			
Topic	Winner	Side 1 Vicuna-13b-v1.5	Side 2 GPT 4	Winner	Side 1 GPT 4	Side 2 Vicuna-13b-v1.5	Overall
1	GPT 4	8	9	GPT 4	9	7	GPT 4
2	GPT 4	6	9	Vicuna-13b-v1.5	8	9	Tie
3	GPT 4	7	8	GPT 4	8	7	GPT 4
4	GPT 4	7	8	GPT 4	8	7	GPT 4
5	GPT 4	7	8	GPT 4	9	8	GPT 4
6	GPT 4	7	9	tie	8	8	GPT 4
7	GPT 4	7	9	Vicuna-13b-v1.5	8	9	Tie
8	GPT 4	6	9	GPT 4	8	7	GPT 4
9	GPT 4	8	9	GPT 4	9	7	GPT 4
10	GPT 4	7	9	GPT 4	8	7	GPT 4
11	GPT 4	7	8	GPT 4	9	7	GPT 4
12	GPT 4	6	9	GPT 4	9	8	GPT 4
13	GPT 4	7	8	GPT 4	9	7	GPT 4
14	GPT 4	7	9	GPT 4	9	7	GPT 4
15	GPT 4	7	8	GPT 4	9	7	GPT 4
16	GPT 4	7	9	GPT 4	9	8	GPT 4
17	GPT 4	6	8	GPT 4	8	7	GPT 4
18	GPT 4	8	9	GPT 4	9	8	GPT 4
19	GPT 4	7	8	GPT 4	9	7	GPT 4
20	GPT 4	6	9	GPT 4	8	7	GPT 4
21	GPT 4	7	9	GPT 4	8	7	GPT 4
22	GPT 4	6	8	GPT 4	8	7	GPT 4
23	GPT 4	7	9	GPT 4	8	7	GPT 4
24	GPT 4	7	9	GPT 4	8	7	GPT 4
25	GPT 4	7	9	tie	8	8	GPT 4

C.39 Mixtral-8x7B vs. Mixtral-8x7B

	Home			Away		
Topic	Winner	Side 1 Mixtral-8x7B	Side 2 Mixtral-8x7B	Winner	Side 1 Mixtral-8x7B	Side 2 Mixtral-8x7B
1	1	8	7	1	8	7
2	2	8	9	2	7	8
3	2	8	9	2	8	9
4	1	8	7	1	8	7
5	1	9	8	1	9	8
6	2	8	9	2	8	9
7	2	8	9	2	7	8
8	1	8	7	1	8	7
9	1	8	7	1	8	7
10	1	8	8	2	8	9
11	1	8	7	1	8	7
12	2	8	9	2	7	8
13	1	8	7	1	8	7
14	1	8	7	1	8	7
15	1	9	7	1	8	7
16	2	8	9	2	8	9
17	2	7	8	1	8	7
18	1	8	7	1	8	7
19	1	8	7	1	8	7
20	2	6	8	2	7	8
21	1	8	7	2	8	9
22	2	7	8	2	7	8
23	1	8	7	1	8	7
24	1	8	7	Tie	8	8
25	2	8	9	2	8	9

C.40 Mixtral-8x7B vs. GPT 3.5

	Home			Away			
Topic	Winner	Side 1 Mixtral-8x7B	Side 2 GPT 3.5	Winner	Side 1 GPT 3.5	Side 2 Mixtral-8x7B	Overall
1	Mixtral-8x7B	8	7	GPT 3.5	8	7	Tie
2	GPT 3.5	7	9	Mixtral-8x7B	7	8	Tie
3	Mixtral-8x7B	8	7	Mixtral-8x7B	8	9	Mixtral-8x7B
4	Mixtral-8x7B	8	7	GPT 3.5	8	7	Tie
5	GPT 3.5	8	9	Mixtral-8x7B	8	9	Tie
6	GPT 3.5	8	9	Mixtral-8x7B	8	9	Tie
7	GPT 3.5	8	9	Mixtral-8x7B	8	9	Tie
8	GPT 3.5	7	8	Mixtral-8x7B	7	8	Tie
9	Mixtral-8x7B	8	7	GPT 3.5	8	7	Tie
10	GPT 3.5	8	9	Mixtral-8x7B	8	9	Tie
11	Mixtral-8x7B	8	7	GPT 3.5	8	7	Tie
12	GPT 3.5	7	9	Mixtral-8x7B	7	8	Tie
13	GPT 3.5	7	8	GPT 3.5	8	7	Mixtral-8x7B
14	GPT 3.5	8	9	GPT 3.5	8	8	Mixtral-8x7B
15	Mixtral-8x7B	8	7	GPT 3.5	8	7	Tie
16	GPT 3.5	8	9	Mixtral-8x7B	8	9	Tie
17	GPT 3.5	7	8	Mixtral-8x7B	7	9	Tie
18	Mixtral-8x7B	8	7	GPT 3.5	8	7	Tie
19	Mixtral-8x7B	8	7	GPT 3.5	8	7	Tie
20	GPT 3.5	7	9	Mixtral-8x7B	7	8	Tie
21	GPT 3.5	7	8	Mixtral-8x7B	8	9	Tie
22	GPT 3.5	7	8	Mixtral-8x7B	7	8	Tie
23	tie	8	8	GPT 3.5	8	7	Mixtral-8x7B
24	GPT 3.5	8	9	Mixtral-8x7B	7	8	Tie
25	GPT 3.5	8	9	Mixtral-8x7B	8	9	Tie

C.41 Mixtral-8x7B vs. GPT 4

	Home			Away			
Topic	Winner	Side 1 Mixtral-8x7B	Side 2 GPT 4	Winner	Side 1 GPT 4	Side 2 Mixtral-8x7B	Overall
1	GPT 4	8	9	GPT 4	9	7	GPT 4
2	GPT 4	6	9	Mixtral-8x7B	8	9	Tie
3	GPT 4	7	8	GPT 4	8	7	GPT 4
4	GPT 4	7	8	GPT 4	8	7	GPT 4
5	GPT 4	7	8	GPT 4	9	8	GPT 4
6	GPT 4	7	9	tie	8	8	GPT 4
7	GPT 4	7	9	Mixtral-8x7B	8	9	Tie
8	GPT 4	6	9	GPT 4	8	7	GPT 4
9	GPT 4	8	9	GPT 4	9	7	GPT 4
10	GPT 4	7	9	GPT 4	8	7	GPT 4
11	GPT 4	7	8	GPT 4	9	7	GPT 4
12	GPT 4	6	9	GPT 4	9	8	GPT 4
13	GPT 4	7	8	GPT 4	9	7	GPT 4
14	GPT 4	7	9	GPT 4	9	7	GPT 4
15	GPT 4	7	8	GPT 4	9	7	GPT 4
16	GPT 4	7	9	GPT 4	9	8	GPT 4
17	GPT 4	6	8	GPT 4	8	7	GPT 4
18	GPT 4	8	9	GPT 4	9	8	GPT 4
19	GPT 4	7	8	GPT 4	9	7	GPT 4
20	GPT 4	6	9	GPT 4	8	7	GPT 4
21	GPT 4	7	9	GPT 4	8	7	GPT 4
22	GPT 4	6	8	GPT 4	8	7	GPT 4
23	GPT 4	7	9	GPT 4	8	7	GPT 4
24	GPT 4	7	9	GPT 4	8	7	GPT 4
25	GPT 4	7	9	tie	8	8	GPT 4

C.42 GPT 3.5 vs. GPT 3.5

	Home			Away		
Topic	Winner	Side 1 GPT 3.5	Side 2 GPT 3.5	Winner	Side 1 GPT 3.5	Side 2 GPT 3.5
1	1	9	8	1	8	7
2	2	7	9	2	8	9
3	2	8	9	2	8	9
4	2	7	8	1	8	7
5	2	8	9	2	8	9
6	2	8	9	2	7	8
7	2	7	8	2	7	8
8	2	7	8	2	7	8
9	1	8	7	2	8	9
10	1	8	7	2	8	9
11	1	8	7	2	8	9
12	2	7	8	2	7	8
13	1	8	7	1	8	7
14	2	8	9	2	8	9
15	1	8	7	1	8	7
16	2	8	9	1	8	7
17	2	7	8	2	7	8
18	1	8	7	1	8	8
19	1	8	7	1	8	7
20	2	7	9	2	7	9
21	2	7	8	2	8	9
22	2	7	8	2	6	8
23	1	8	7	1	8	7
24	2	8	9	2	8	9
25	2	8	9	2	7	8

C.43 GPT 3.5 vs. GPT 4

	Home			Away			
Topic	Winner	Side 1 GPT 3.5	Side 2 GPT 4	Winner	Side 1 GPT 4	Side 2 GPT 3.5	Overall
1	GPT 4	8	9	GPT 4	9	8	GPT 4
2	GPT 4	7	9	GPT 3.5	8	9	Tie
3	GPT 4	8	9	GPT 4	9	8	GPT 4
4	GPT 4	7	8	GPT 4	8	7	GPT 4
5	GPT 4	8	9	GPT 4	8	7	GPT 4
6	GPT 4	7	9	GPT 3.5	8	9	Tie
7	GPT 4	7	9	GPT 3.5	8	9	Tie
8	GPT 4	7	8	GPT 3.5	7	8	Tie
9	GPT 4	8	9	GPT 4	9	8	GPT 4
10	GPT 4	8	9	GPT 4	9	8	GPT 4
11	GPT 4	8	9	GPT 4	9	7	GPT 4
12	GPT 4	6	9	GPT 3.5	8	9	Tie
13	GPT 4	7	8	GPT 4	8	7	GPT 4
14	GPT 4	8	9	GPT 4	9	8	GPT 4
15	GPT 4	8	9	GPT 4	9	8	GPT 4
16	GPT 4	7	9	GPT 4	9	8	GPT 4
17	GPT 4	7	9	GPT 3.5	8	9	Tie
18	GPT 4	8	9	GPT 4	9	8	GPT 4
19	GPT 3.5	8	7	GPT 3.5	9	7	GPT 3.5
20	GPT 4	7	9	GPT 3.5	7	8	Tie
21	GPT 4	7	8	GPT 3.5	8	9	Tie
22	GPT 4	6	8	GPT 3.5	8	9	Tie
23	GPT 4	8	9	GPT 4	8	7	GPT 4
24	GPT 4	7	8	GPT 3.5	8	9	Tie
25	GPT 4	7	8	GPT 4	8	7	GPT 4

C.44 GPT 4 vs. GPT 4

	Home			Away		
Topic	Winner	Side 1 GPT 4	Side 2 GPT 4	Winner	Side 1 GPT 4	Side 2 GPT 4
1	1	8	8	1	8	7
2	2	7	8	2	8	9
3	2	8	9	2	8	9
4	1	8	7	1	8	7
5	2	8	9	1	9	8
6	2	7	8	2	8	9
7	2	8	9	2	7	8
8	2	7	8	2	7	8
9	2	8	9	1	8	7
10	2	8	9	2	8	9
11	1	8	7	1	8	7
12	2	8	9	2	8	9
13	1	8	7	1	8	7
14	1	8	7	2	8	9
15	1	8	7	1	9	8
16	2	8	9	2	8	9
17	2	8	9	2	7	8
18	1	9	8	1	8	7
19	1	8	7	1	9	8
20	2	8	9	2	8	9
21	2	8	9	2	8	9
22	2	8	9	2	8	9
23	1	9	8	2	8	9
24	2	8	9	2	8	9
25	2	8	9	2	8	9

D Other Judges

In this section, we report the results of experiment from Section 4.2 with Llama-3-70b as judge.

D.1 Llama2-13b vs. Llama2-7b

	Home			Away			
Topic	Winner	Side 1 Llama2-7b	Side 2 Llama2-13b	Winner	Side 1 Llama2-13b	Side 2 Llama2-7b	Overall
1	Llama2-7b	8	7	Llama2-13b	8	7	Tie
2	Llama2-13b	8	9	Llama2-7b	8	9	Tie
3	Llama2-7b	8	7	Llama2-13b	8	7	Tie
4	Llama2-7b	8	6	Llama2-13b	8	6	Tie
5	Llama2-7b	8	7	Llama2-13b	8	7	Tie
6	Llama2-7b	8	7	Llama2-13b	8	7	Tie
7	Llama2-13b	8	9	Llama2-7b	8	9	Tie
8	Llama2-13b	8	9	Llama2-7b	8	9	Tie
9	Llama2-7b	8	7	Llama2-13b	8	7	Tie
10	Llama2-13b	8	9	Llama2-13b	8	7	Llama2-13b
11	Llama2-7b	8	7	Llama2-13b	8	6	Tie
12	Llama2-13b	8	9	Llama2-13b	8	7	Llama2-13b
13	Llama2-7b	8	7	Llama2-13b	8	6	Tie
14	Llama2-7b	8	7	Llama2-13b	8	7	Tie
15	Llama2-7b	8	6	Llama2-13b	8	6	Tie
16	Llama2-7b	8	7	Llama2-13b	8	7	Tie
17	Llama2-13b	8	9	Llama2-13b	8	7	Llama2-13b
18	Llama2-7b	8	7	Llama2-13b	8	7	Tie
19	Llama2-7b	8	7	Llama2-13b	8	6	Tie
20	Llama2-13b	7	8	Llama2-7b	8	9	Tie
21	Llama2-13b	8	9	Llama2-13b	8	7	Llama2-13b
22	Llama2-13b	7	8	Llama2-7b	6	8	Llama2-13b
23	Llama2-13b	8	9	Llama2-13b	8	7	Llama2-13b
24	Llama2-13b	8	9	Llama2-13b	8	7	Llama2-13b
25	Llama2-7b	8	7	Llama2-13b	8	7	Tie

D.2 Llama2-13b vs. Llama2-70b

	Home			Away			
Topic	Winner	Side 1 Llama2-13b	Side 2 Llama2-70b	Winner	Side 1 Llama2-70b	Side 2 Llama2-13b	Overall
1	Llama2-13b	8	7	Llama2-70b	8	7	Tie
2	Llama2-70b	7	8	Llama2-70b	8	7	Llama2-70b
3	Llama2-70b	8	9	Llama2-13b	8	9	Tie
4	Llama2-13b	8	7	Llama2-70b	8	7	Tie
5	Llama2-13b	8	7	Llama2-70b	8	7	Tie
6	Llama2-70b	8	9	Llama2-70b	8	7	Llama2-70b
7	Llama2-13b	8	7	Llama2-70b	8	7	Tie
8	Llama2-13b	8	7	Llama2-70b	8	7	Tie
9	Llama2-13b	8	7	Llama2-70b	8	7	Tie
10	Llama2-70b	8	9	Llama2-13b	8	9	Tie
11	Llama2-13b	8	7	Llama2-70b	8	7	Tie
12	Llama2-70b	8	9	Llama2-13b	8	9	Tie
13	Llama2-13b	8	7	Llama2-70b	8	7	Tie
14	Llama2-13b	8	7	Llama2-70b	8	7	Tie
15	Llama2-13b	8	6	Llama2-70b	8	6	Tie
16	Llama2-70b	7	8	Llama2-70b	8	7	Llama2-70b
17	Llama2-13b	8	6	Llama2-70b	8	7	Tie
18	Llama2-13b	8	7	Llama2-70b	8	7	Tie
19	Llama2-13b	8	6	Llama2-70b	8	6	Tie
20	Llama2-70b	7	8	Llama2-13b	8	9	Tie
21	Llama2-13b	8	7	Llama2-70b	8	7	Tie
22	Llama2-70b	6	8	Llama2-13b	7	8	Tie
23	Llama2-13b	8	7	Llama2-70b	8	7	Tie
24	Llama2-70b	8	9	Llama2-13b	8	9	Tie
25	Llama2-70b	8	9	Llama2-70b	8	7	Llama2-70b

D.3 Llama2-13b vs. Llama3-70b

	Home			Away			
Topic	Winner	Side 1 Llama3-70b	Side 2 Llama2-13b	Winner	Side 1 Llama2-13b	Side 2 Llama3-70b	Overall
1	Llama3-70b	8	7	Llama2-13b	8	7	Tie
2	Llama2-13b	8	9	Llama3-70b	8	9	Tie
3	Llama3-70b	8	7	Llama2-13b	8	7	Tie
4	Llama3-70b	8	6	Llama2-13b	8	7	Tie
5	Llama3-70b	8	7	Llama2-13b	8	7	Tie
6	Llama2-13b	8	9	Llama2-13b	8	7	Llama2-13b
7	Llama2-13b	8	9	Llama3-70b	8	9	Tie
8	Llama2-13b	8	9	Llama3-70b	8	9	Tie
9	Llama3-70b	8	7	Llama2-13b	8	7	Tie
10	Llama2-13b	8	9	Llama3-70b	7	8	Tie
11	Llama3-70b	8	7	Llama2-13b	8	7	Tie
12	Llama2-13b	8	9	Llama3-70b	8	9	Tie
13	Llama3-70b	8	7	Llama2-13b	8	7	Tie
14	Llama3-70b	8	7	Llama3-70b	8	9	Llama3-70b
15	Llama3-70b	8	7	Llama2-13b	8	7	Tie
16	Llama2-13b	8	9	Llama2-13b	8	7	Llama2-13b
17	Llama2-13b	8	9	Llama3-70b	7	8	Tie
18	Llama3-70b	8	7	Llama2-13b	8	7	Tie
19	Llama3-70b	8	7	Llama2-13b	8	6	Tie
20	Llama3-70b	8	7	Llama3-70b	8	9	Llama3-70b
21	Llama3-70b	8	7	Llama3-70b	8	9	Llama3-70b
22	Llama2-13b	8	9	Llama3-70b	7	8	Tie
23	Llama3-70b	8	7	Llama3-70b	8	9	Llama3-70b
24	Llama3-70b	8	7	Llama2-13b	8	7	Tie
25	Llama2-13b	8	9	Llama3-70b	8	9	Tie

D.4 Llama2-13b vs. Vicuna-7b-v1.5

	Home			Away			
Topic	Winner	Side 1 Llama2-13b	Side 2 Vicuna-7b-v1.5	Winner	Side 1 Vicuna-7b-v1.5	Side 2 Llama2-13b	Overall
1	Llama2-13b	8	7	Vicuna-7b-v1.5	8	7	Tie
2	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
3	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
4	Llama2-13b	8	6	Vicuna-7b-v1.5	8	7	Tie
5	Llama2-13b	8	7	Vicuna-7b-v1.5	8	7	Tie
6	Llama2-13b	8	7	Llama2-13b	7	9	Llama2-13b
7	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
8	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
9	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
10	Llama2-13b	8	7	Llama2-13b	7	8	Llama2-13b
11	Llama2-13b	8	6	Vicuna-7b-v1.5	8	7	Tie
12	Llama2-13b	8	7	Llama2-13b	7	9	Llama2-13b
13	Llama2-13b	8	6	Vicuna-7b-v1.5	8	7	Tie
14	Llama2-13b	8	6	Llama2-13b	8	9	Llama2-13b
15	Llama2-13b	8	6	Vicuna-7b-v1.5	8	7	Tie
16	Llama2-13b	8	6	Llama2-13b	8	9	Llama2-13b
17	Llama2-13b	8	6	Llama2-13b	7	9	Llama2-13b
18	Llama2-13b	8	6	Llama2-13b	8	9	Llama2-13b
19	Llama2-13b	8	6	Vicuna-7b-v1.5	8	7	Tie
20	Vicuna-7b-v1.5	8	9	Vicuna-7b-v1.5	7	8	Vicuna-7b-v1.5
21	Llama2-13b	8	7	Llama2-13b	7	9	Llama2-13b
22	Vicuna-7b-v1.5	7	8	Vicuna-7b-v1.5	6	8	Vicuna-7b-v1.5
23	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
24	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
25	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b

D.5 Llama2-13b vs. Vicuna-13b-v1.5

	Home			Away			
Topic	Winner	Side 1 Llama2-13b	Side 2 Vicuna-13b-v1.5	Winner	Side 1 Vicuna-13b-v1.5	Side 2 Llama2-13b	Overall
1	Llama2-13b	8	7	Vicuna-13b-v1.5	8	7	Tie
2	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
3	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
4	Llama2-13b	8	6	Llama2-13b	8	9	Llama2-13b
5	Llama2-13b	8	7	Vicuna-13b-v1.5	8	7	Tie
6	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
7	Vicuna-13b-v1.5	8	9	Llama2-13b	8	9	Tie
8	Llama2-13b	8	7	Llama2-13b	7	9	Llama2-13b
9	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
10	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
11	Llama2-13b	8	6	Llama2-13b	8	9	Llama2-13b
12	Vicuna-13b-v1.5	8	9	Llama2-13b	7	9	Tie
13	Llama2-13b	8	6	Llama2-13b	8	9	Llama2-13b
14	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
15	Llama2-13b	8	6	Vicuna-13b-v1.5	8	7	Tie
16	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
17	Llama2-13b	8	6	Llama2-13b	7	8	Llama2-13b
18	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
19	Llama2-13b	8	6	Vicuna-13b-v1.5	8	7	Tie
20	Vicuna-13b-v1.5	8	9	Llama2-13b	7	8	Tie
21	Llama2-13b	8	7	Llama2-13b	7	8	Llama2-13b
22	Vicuna-13b-v1.5	7	8	Llama2-13b	8	9	Tie
23	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
24	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
25	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b

D.6 Llama2-13b vs. Mixtral-8x7B

	Home			Away			
Topic	Winner	Side 1 Llama2-13b	Side 2 Mixtral-8x7B	Winner	Side 1 Mixtral-8x7B	Side 2 Llama2-13b	Overall
1	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B
2	Mixtral-8x7B	8	9	Llama2-13b	8	9	Tie
3	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
4	Llama2-13b	8	7	Mixtral-8x7B	8	7	Tie
5	Llama2-13b	8	7	Mixtral-8x7B	8	7	Tie
6	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
7	Mixtral-8x7B	8	9	Llama2-13b	8	9	Tie
8	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
9	Llama2-13b	8	7	Mixtral-8x7B	8	7	Tie
10	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
11	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B
12	Mixtral-8x7B	8	9	Llama2-13b	8	9	Tie
13	Llama2-13b	8	7	Mixtral-8x7B	8	7	Tie
14	Mixtral-8x7B	8	9	Mixtral-8x7B	8	7	Mixtral-8x7B
15	Llama2-13b	8	6	Mixtral-8x7B	8	6	Tie
16	Llama2-13b	9	9	Llama2-13b	8	9	Llama2-13b
17	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
18	Llama2-13b	8	7	Mixtral-8x7B	8	7	Tie
19	Llama2-13b	8	7	Mixtral-8x7B	8	7	Tie
20	Mixtral-8x7B	8	9	Llama2-13b	7	8	Tie
21	Mixtral-8x7B	8	9	Llama2-13b	8	9	Tie
22	Mixtral-8x7B	6	8	Llama2-13b	7	8	Tie
23	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
24	Llama2-13b	8	7	Llama2-13b	8	9	Llama2-13b
25	Mixtral-8x7B	6	8	Mixtral-8x7B	9	7	Mixtral-8x7B

D.7 Llama2-13b vs. GPT-3.5

	Home			Away			
Topic	Winner	Side 1 Llama2-13b	Side 2 GPT-3.5	Winner	Side 1 GPT-3.5	Side 2 Llama2-13b	Overall
1	GPT-3.5	8	9	GPT-3.5	8	7	GPT-3.5
2	GPT-3.5	8	9	Llama2-13b	8	9	Tie
3	GPT-3.5	8	9	Llama2-13b	8	9	Tie
4	Llama2-13b	8	7	GPT-3.5	8	7	Tie
5	Llama2-13b	8	7	GPT-3.5	8	7	Tie
6	GPT-3.5	8	9	Llama2-13b	8	9	Tie
7	GPT-3.5	8	9	Llama2-13b	7	8	Tie
8	GPT-3.5	8	9	Llama2-13b	8	9	Tie
9	Llama2-13b	8	7	GPT-3.5	9	8	Tie
10	GPT-3.5	8	9	Llama2-13b	8	9	Tie
11	GPT-3.5	8	9	Llama2-13b	8	9	Tie
12	GPT-3.5	8	9	Llama2-13b	7	9	Tie
13	Llama2-13b	8	7	GPT-3.5	8	7	Tie
14	GPT-3.5	8	9	Llama2-13b	8	9	Tie
15	Llama2-13b	8	7	GPT-3.5	8	6	Tie
16	GPT-3.5	8	9	Llama2-13b	8	9	Tie
17	GPT-3.5	8	9	GPT-3.5	8	7	GPT-3.5
18	GPT-3.5	8	9	Llama2-13b	8	9	Tie
19	Llama2-13b	8	7	GPT-3.5	8	7	Tie
20	GPT-3.5	7	9	Llama2-13b	7	8	Tie
21	GPT-3.5	8	9	Llama2-13b	8	9	Tie
22	GPT-3.5	6	8	Llama2-13b	7	9	Tie
23	GPT-3.5	8	9	GPT-3.5	9	8	GPT-3.5
24	GPT-3.5	8	9	GPT-3.5	9	8	GPT-3.5
25	GPT-3.5	8	9	Llama2-13b	7	8	Tie

D.8 Llama2-13b vs. GPT-4

	Home			Away			
Topic	Winner	Side 1 Llama2-13b	Side 2 GPT-4	Winner	Side 1 GPT-4	Side 2 Llama2-13b	Overall
1	GPT-4	8	9	GPT-4	9	7	GPT-4
2	GPT-4	8	9	Llama2-13b	8	9	Tie
3	GPT-4	8	9	GPT-4	8	7	GPT-4
4	GPT-4	8	9	GPT-4	8	6	GPT-4
5	GPT-4	8	9	GPT-4	8.5	7.5	GPT-4
6	GPT-4	7	9	GPT-4	8	7	GPT-4
7	GPT-4	8	9	Llama2-13b	8	9	Tie
8	GPT-4	8	9	GPT-4	8	7	GPT-4
9	GPT-4	8	9	GPT-4	8	6	GPT-4
10	GPT-4	8	9	GPT-4	8	7	GPT-4
11	GPT-4	8	9	GPT-4	8.5	7.5	GPT-4
12	GPT-4	7	9	Llama2-13b	8	9	Tie
13	GPT-4	8	9	GPT-4	8	6	GPT-4
14	GPT-4	8	9	GPT-4	8	7	GPT-4
15	GPT-4	8	9	GPT-4	9	7	GPT-4
16	GPT-4	8	9	GPT-4	8	7	GPT-4
17	GPT-4	7	9	GPT-4	8	6	GPT-4
18	GPT-4	8	9	GPT-4	8	7	GPT-4
19	GPT-4	8	9	GPT-4	8.5	7.5	GPT-4
20	GPT-4	7	9	Llama2-13b	8	9	Tie
21	GPT-4	7	9	GPT-4	8	7	GPT-4
22	GPT-4	6	8	GPT-4	8	7	GPT-4
23	GPT-4	8	9	GPT-4	8	7	GPT-4
24	GPT-4	8	9	GPT-4	8	7	GPT-4
25	GPT-4	8	9	GPT-4	8	7	GPT-4