DebateBench: A Challenging Long Context Reasoning Benchmark For Large Language Models

Utkarsh Tiwari* **Aryan Seth*** **Adi Mukherjee**
Kaavya Mer **Kavish** **Dhruv Kumar**
Birla Institute of Technology and Science, Pilani
f20212221, f20220052@pilani.bits-pilani.ac.in

Abstract

We introduce DebateBench, a novel dataset consisting of an extensive collection of transcripts and metadata from some of the world's most prestigious competitive debates. The dataset consists of British Parliamentary debates from prestigious debating tournaments on diverse topics, annotated with detailed speech-level scores and house rankings sourced from official adjudication data. We curate 256 speeches across 32 debates with each debate being over 1 hour long with each input being an average of 32,000 tokens. Designed to capture long-context, large-scale reasoning tasks, DebateBench provides a benchmark for evaluating modern large language models (LLMs) on their ability to engage in argumentation, deliberation, and alignment with human experts. To do well on DebateBench, the LLMs must perform in-context learning to understand the rules and evaluation criteria of the debates, then analyze 8 seven minute long speeches and reason about the arguments presented by all speakers to give the final results. Our preliminary evaluation using GPT o1, GPT-4o, and Claude haiku, shows that LLMs struggle to perform well on DebateBench, highlighting the need to develop more sophisticated techniques for improving their performance.

1 Introduction

The reasoning capabilities of Large Language Models (LLMs) have been extensively evaluated across a variety of domains, including STEM problem-solving (Cobbe et al., 2021; Clark et al., 2018; Arora et al., 2023; Hendrycks et al., 2021b; Lu et al., 2022; Bubeck et al., 2023), language understanding (Hendrycks et al., 2021a), and code generation (Chen et al., 2021; Austin et al., 2021). However, there remains a notable gap in the availability of sufficiently diverse and challenging natural language datasets that rigorously benchmark reasoning over long-contexts. Additionally, existing benchmarks for long-context reasoning suffer from two main problems: (1) lack of "argument-only" debates, and (2) non-comprehensive scoring metrics. Argument-only debates differ from factual debates in that the judging panel knows no facts beforehand and needs to be convinced of their correctness. Existing metrics in these datasets are also win/loss based, rather than provided by expert annotators. A comprehensive overview of existing benchmarks is provided in Section 2.

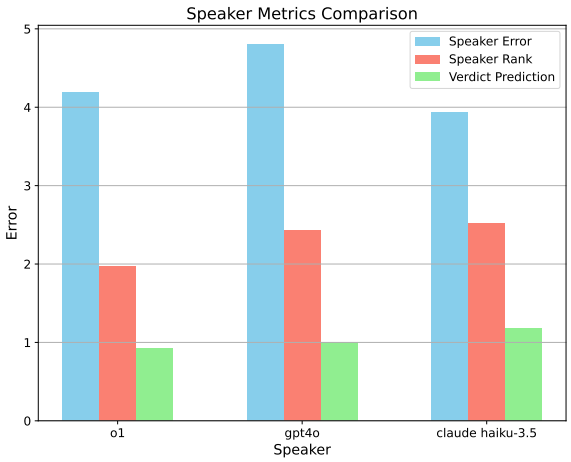

Figure 1: Performance of models on DebateBench, the y-axis represents the mean absolute error (MAE) of the three tasks. More details in 3.2

In this paper, we propose DebateBench, a novel dataset consisting of an extensive collection of transcripts and metadata of British Parliamentary debates from prestigious debating tournaments on diverse topics. It is an annotated dataset containing detailed speech-level scores and house rankings sourced from official adjudication data. We curate 256 speeches across 32 debates with each debate being over 1 hour long with each input being an average of 32,000 tokens. DebateBench serves as a challenging benchmark for evaluating modern large language models (LLMs) on their ability to engage in argumentation, deliberation, and alignment with human experts.

DebateBench includes three primary evaluation tasks: (1) **Speech Scoring**, where models predict human-assigned scores for individual speeches, (2) **Speech Ranking**, where the model predicts the speaker rankings, and (3) **House Ordering**, where models rank debating teams by adjudication outcomes.

This paper makes three primary contributions:

1. **New Dataset**: We introduce, DebateBench, a long-context reasoning benchmark comprising 32 debates conducted in the British Parliamentary format, approximately having 100,000 words per debate.

2. **Novel Task:** We introduce complex multi-turn debates for reasoning and structured format argumentation over long contexts, scored against a human ground truth.

3. **Evaluation**: Top-of-the-line LLMs struggle on DebateBench, demonstrating their inability to handle dense long-context tasks that require structured argumentation. Additional details are discussed in Section 4.

2 Related Works

Reasoning in LLMs has been studied under multiple contexts such as logical reasoning, mathematical reasoning, theorem proving etc. In real life contexts, this extends to fields at the intersection of reasoning, decision-making, and communication, such as law, politics, and education. In this section we summarise the literature on natural language reasoning which is the closest related to our task.

2.1 Natural Langauge Reasoning

Natural Language datasets like **HellaSwag** (Zellers et al., 2019), **ARC** (Clark et al., 2018), and **MMLU** (Hendrycks et al., 2021a) test LLMs ability to understand natural language and ground their answers in reality. HellaSwag tests models' ability to complete sentence by choosing the most likely option from 4 setence provided, while ARC and MMLU test models on either domain dependent information or common sense reasoning by asking questions on biology, law, economics, etc. The **TruthfulQA** (Lin et al., 2022) benchmark is especially designed to test models' grounding in reality by evaluating them on questions who's answer are prone to be misinformation or conspiratorial.

Other benchmarks like **SuperGLUE** (Wang et al., 2020) and **WinoGrande** (Sakaguchi et al., 2019) test models' comprehension of natural languages by testing them on confusing or ambiguous passages.

More closely aligned to our work are argument evaluation benchmarks like **VivesDebate-Speech** (Ruiz-Dolz and Iranzo-Sánchez, 2024), which is a dataset of 29 debates from the 2019 university debate tournament organised by the "Xarxa Vives d'universitats". However, the debates in this benchmark are not originally English, and have been machine translated from Catalan. The credibility of machine translations in preserving complicated arguments is low. Moreover, DebateBench includes debates from renowned competitions hence the judges' scores are more credible and the debates are of a higher quality. Other "debating" datasets like **USElecDeb60To16** (Haddadan et al., 2019), and **ETHIC** (Lee et al., 2024) benchmark deal with political debates between U.S. Presidential candidates and in the British Parliament respectively. These debates are significantly different from competitive debates since the main focus is on rhetoric and not logical argumentation. These debates also don't have a quantifiable metric of evaluation. The **DebateSum** (Roush and Balaji, 2020) deals with *Policy Debates* wherein the topics are released as much a year ago and the competition focuses on the presentation of evidence and data instead of principled arguments.

2.2 Long-Context Modeling Techniques

Recent advancements in LLMs have integrated sophisticated long-context modeling techniques. For instance, LLaMa 2 employs Rotary Position Embedding (RoPE) (Touvron et al., 2023), while Vicuna 1.5 (Zhang et al., 2023) fine-tunes LLaMa 2 to extend context lengths to 16,000 tokens. Similarly, ChatGLM2-32k achieves a 32,000-token context window, demonstrating the scalability of these methods. State-of-the-art models like GPT-4-Turbo (128,000 tokens) and Claude-3.5-Sonnet (200,000 tokens) further push the boundaries of context length, enabling the processing of extensive information. Despite these advancements, there is a notable scarcity of human-aligned benchmarks designed to evaluate performance at such scale.

3 The DebateBench Dataset

The dataset comprises 32 debates, each consisting of 8 speeches approximately 7 minutes long, curated from prestigious tournaments such as

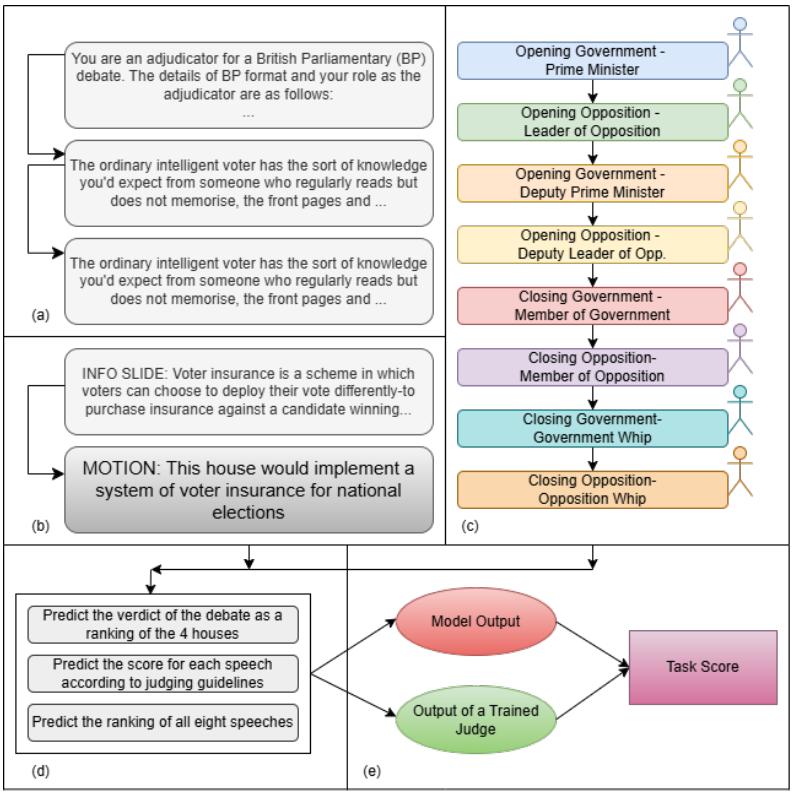

Figure 2: The system prompt explaining the format of the debate as well as the metrics of judgment (a) along with the information slide (if present) and the motion (b) and the transcript of the debate (c) which contains 8 speeches by 4 teams (or houses) is passed to the model. The model is tested on 3 tasks (d) and the output is compared to the results given by trained judges to compute the task scores for the model (e).

the Doxbridge Worlds Schools Debating Championships, LSE Opens, and past editions of the World Universities Debating Championship (WUDC). These tournaments are recognized for their high-quality debates, ensuring a robust benchmark. A full list of debates is provided in the Appendix.

Debates follow the British Parliamentary format, featuring four teams (or houses): Opening Government, Opening Opposition, Closing Government, and Closing Opposition. Each team competes in a structured round, with detailed format specifications provided in Figure 2.

The debate topic, referred to as the "motion," is framed as a proposal for "the house." For example, a motion from the Cambridge IV 2020 debate states: "This house believes that protest movements should actively integrate religious figures and institutions in opposition to authoritarian regimes." Government teams support the motion, while opposition teams oppose it. Teams are allotted 15 minutes of preparation time without internet access, discouraging reliance on statistics or specialized knowledge. Instead, speakers are expected to construct arguments based on principles and plausible scenarios, logically extending them. Judges, guided by the WUDC judging manual[1], are instructed to evaluate debates as "ordinary intelligent voters" or "informed global citizens." They discount appeals to highly specialized concepts unless clearly explained, ensuring arguments remain accessible. While complex claims are permissible, they must be articulated in jargon-free, understandable terms. During the evaluation, LLMs are provided with the judging manual and instructed to adjudicate accordingly (see Appendix for the full system prompt).

The judging manual also outlines heuristics for speech objectives. For instance, the first two speeches are expected to contextualize the debate and clarify ambiguities, while the final two speeches must identify major clashes and justify their team's success. The system prompt, adapted from the WUDC manual, exceeds 18,000 tokens and includes details on debate format, speaker roles, and adjudication heuristics. Additionally, most motions include an "Info Slide" explaining key terms, as illustrated in Figure 2. Consequently, perform-

[1] https://shorturl.at/QnjKe

ing well on DebateBench requires models to excel at in-context learning (Dong et al., 2024), further underscoring the benchmark's complexity.

3.1 Dataset Collection

We first curate debating videos from YouTube. These videos are recorded with prior consent of all speakers and publicly shared. We remove the videos which contain less than 8 speeches, and those in which large sections are unintelligible due to bad audio quality. We then use GPT-Whisper (Verma, 2024) to generate transcripts. However, these transcripts are of low quality and need to be processed before being used. We then use GPT-4o mini to correct any grammatical errors and spelling mistakes. The timestamps are preserved in this step. It was observed, that this step was able to correctly infer punctuation and correct grammar and spelling.

Finally, all the transcripts are manually verified to ensure a high quality dataset. Speaker tags and other tokens are added to differentiate speakers in the debate.

Statistic	Value
Total number of speeches	256
Total number of words	884,395
Total number of tokens	1,326,592
Average number of words per speech	11,904
Total hours of content	∼36
Average speaker score	80.25
Speaker score standard deviation	2.69

Table 1: Statistics of the dataset. The tokens are calculated assuming the standard 1 word ≈ 1.5 tokens.

3.2 Evaluation Metrics

We propose three tasks for evaluating the performance of large language models (LLMs) on DebateBench. For each task, the models are provided with the judging manual along with the debate transcript:

- **Verdict Prediction**: The models are tasked with predicting the ranking of the four houses. The predicted ranking is compared to the ground truth ranking generated by the trained judges by computing the absolute difference at each position. Let the predicted ranking be denoted as $r_{\text{pred}} = (r_1, r_2, r_3, r_4)$ and the ground truth ranking as $r_{\text{gt}} = (g_1, g_2, g_3, g_4)$, where r_i and g_i are the positions of the houses in the predicted and ground truth rankings, respectively. The verdict prediction score, Δ, is defined as:

$$\Delta = \sum_{i=1}^{4} |r_i - g_i|$$

  If the model predicts the ranking correctly, then $\Delta = 0$.

- **Speaker Scores**: Each of the 8 speeches is assigned a score between 50 and 100 by the judges. The guidelines for assigning these scores are provided in the judging manual and passed as a system prompt. This task is formulated as a regression problem, where the model predicts a score $s_i \in [50, 100]$ for each speech i. The model's prediction $\hat{s}_i$ is compared to the true score s_i by taking absolute difference. Additionally, we introduce tolerance ϵ, allowing the predicted score to fall within an acceptable range of the true score. We ablate on multiple tolerance values, ranging from 2 to 9.

- **Speaker Ranks**: Given the challenging nature of the speaker scoring task, we introduce a simpler extension by asking the model to rank the 8 speeches. Let $r_{\text{pred}} = (r_1, r_2, \ldots, r_8)$ represent the predicted ranks, where r_i denotes the rank of speech i, and $r_{\text{gt}} = (g_1, g_2, \ldots, g_8)$ denotes the true ranks. We evaluate model performance using the absolute difference metric, defined as:

$$\Delta_{\text{rank}} = \sum_{i=1}^{8} |r_i - g_i|$$

  A lower Δ_{rank} indicates a better match with the ground truth ranking. Models that perform well on the speaker scoring task should also perform well on this ranking task. However, this task offers greater differentiation in cases where models perform poorly on speaker scores.

4 Experiments

4.1 Experimental Setup

Settings We utilize the official WUDC judging manual as context and pass the debate transcripts for evaluation. We apply the corresponding chat

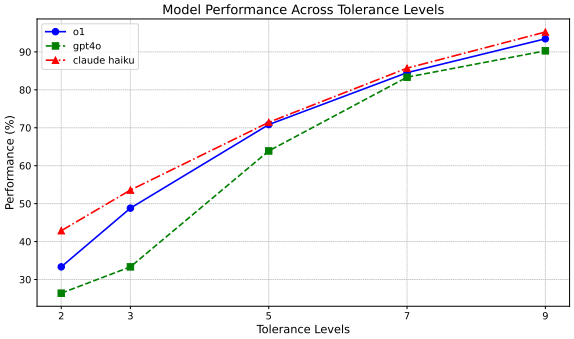

Figure 3: Model accuracy for speaker score prediction at varying delta windows from ground truth

and system prompts for each LLM and keep the temperature set to 0.0 while retaining all other sampling parameters as the standard configurations.

Models We evaluate 3 LLMs, namely OpenAI's GPT4o(OpenAI et al., 2024) and o1(Jaech et al., 2024), and Anthropic's Claude Haiku 3.5.

4.2 Main Results

From 1 we can see that o1 performs the best across the three models for ranking tasks but is still considerably unreliable, with high errors for the verdict and speaker ranking prediction. gpt4o and Haiku 3.5 have comparable results regarding verdict prediction, with all three having an average error of 1. However, they have varying results on speaker scores, with Haiku performing the best and gpt4o performing the worst. This demonstrates that current LLMs are unable to handle contexts of this size when detailed reasoning is involved.

4.3 Speaker Score Accuracy

To evaluate how close LLMs are in predicting speaker scores, which is well demarcated in the WUDC judging manual, we evaluate their performance by checking whether they were a certain tolerance away from the ground truth. From Figure 3 we see that Claude Haiku performs well, however even at a tolerance of 5, models are about 70% accurate for score prediction. While this may seem good, it is worth noting that the speaker scores have a standard deviation of 2.69 (as shown in Table 1), which implies that a tolerance of 5 is considerably high.

5 Future Work and Conclusion

In this work, we introduce DebateBench, a novel dataset comprising high-quality transcripts and metadata from prestigious competitive debates. This dataset is designed to evaluate language models on their ability to perform in-context learning and logical reasoning over long-form, natural language discourse.

Evaluation: We also present a preliminary evaluation of o1, GPT4o, and Claude Haiku 3.5 on three newly formulated tasks. Our findings indicate that current models struggle to achieve high accuracy on DebateBench, highlighting the challenges posed by its complex reasoning and extensive context requirements.

The limited class of models for evaluation was due to limitations of cost, and future works plan to include a broader range of models featuring varying context windows, different types of preference alignment, and models fine-tuned on this task to evaluate LLMs in a much deeper mannDataset er.

Dataset Extension: A natural extension of this dataset involves incorporating argument annotations into the debate transcripts. Such annotations would enable additional benchmarking tasks, including question-answering, and could serve as a valuable resource for assessing human alignment, as they would be produced by experienced debate judges.

DebateBench is currently formulated to include only those debates for which the metadata defining all three tasks was available. This constraint can be mitigated by expanding the dataset to include debates lacking certain metadata, such as speaker scores. While these debates may not support all evaluation tasks, they will be valuable for verdict prediction and increase the scope of evaluation.

Future Work: Given that debate motions frequently address contentious topics, DebateBench also provides a valuable resource for analyzing bias in language models by examining how they weigh opposing arguments on controversial issues.

Overall, DebateBench presents a challenging benchmark for language model evaluation while simultaneously facilitating future advancements in areas such as human alignment and bias mitigation.

References

Daman Arora, Himanshu Gaurav Singh, and Mausam. 2023. Have llms advanced enough? a challenging problem solving benchmark for large language models. *Preprint*, arXiv:2305.15074.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen

Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Taewhoo Lee, Chanwoong Yoon, Kyochul Jang, Donghyeon Lee, Minju Song, Hyunjae Kim, and Jaewoo Kang. 2024. Ethic: Evaluating large language models on long-context tasks with high information coverage. *Preprint*, arXiv:2410.16848.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim,

Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Allen Roush and Arvind Balaji. 2020. DebateSum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online. Association for Computational Linguistics.

Ramon Ruiz-Dolz and Javier Iranzo-Sánchez. 2024. Vivesdebate-speech: A corpus of spoken argumentation to leverage audio features for argument mining. *Preprint*, arXiv:2302.12584.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

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*.

Prateek Verma. 2024. Whisper-gpt: A hybrid representation audio large language model. *arXiv preprint arXiv:2412.11449*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems. *Preprint*, arXiv:1905.00537.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Zheng Zhang, Chen Zheng, Da Tang, Ke Sun, Yukun Ma, Yingtong Bu, Xun Zhou, and Liang Zhao. 2023. Balancing specialized and general skills in llms: The impact of modern tuning and data strategy. *arXiv preprint arXiv:2310.04945*.

A System Prompt

Below is the system prompt for the verdict prediction task:

```
You are an experienced debate judge. You
    have the following judging manual
    context:
    {judging_manual}

Below is a single text containing all 8
    speeches for this round:
    {all_speeches}

Use the judging manual to assign speaker
    scores in a fair and consistent
    manner.
Output the rankings in the following
    format:
First: <house>
Second: <house>
Third:<house>
Fourth: <house>

For example, a sample output could be:
First: OG
Second: CG
Third: OO
Fourth: CO

No additional explanation should be
    provided.
```

Figure 4: The judging manual is adapted from the WUDC judging manual and contains 15,361 words. The entire prompt, including the judging manual, can be found in the code repository.

```
Astana Open 2023 Round 3
Astana Open Round 4
Astana Open Round 5
Belgrade WUDC 2022 Round 6
Belgrade WUDC 2022 Round 7
Belgrade WUDC 2022 Round 8
Belgrade WUDC 2022 Round 9 Room 2
Cambridge IV 2020 Round 1
Cambridge IV 2020 Round 2
Doxbridge 3 Round 4
Doxbridge 4 Round 1
Doxbridge 4 Round 2
Doxbridge Worlds 2021 West Round 4
LSE Open 2023 Round 2 Room 2
LSE Open 2023 Round 3 Room 1
LSE Open 2023 Round 3
LSE Open 2024 Round 1A
LSE Open 2024 Round 3 A
LSE Open 2024 Round 3 B
Pakistan Pre ABP 2024 - Round 5
Panama WUDC 2025 Round 1
Panama WUDC 2025 Round 5
Panama WUDC 2025 Round 7
The Natolin European Round Robin
    Debating Championships Round 1
The Natolin European Round Robin
    Debating Championships Round 2
The Natolin European Round Robin
    Debating Championships Round 3
The Natolin European Round Robin
    Debating Championships Round 4
The Natolin European Round Robin
    Debating Championships Round 5
Doxbridge 3 Round 3
Doxbridge Pre WUDC 2022 Round 5
Doxbridge Worlds 2021 West Round 5
LSE Open 2023 Round 5
```

Figure 5: List of debate rounds included in the dataset.