

Debating with More Persuasive LLMs Leads to More Truthful Answers

Author: Khan et al.

Uploaded to arxiv: 2024-02-09

<https://arxiv.org/abs/2402.06782>

Executive Summary

- The paper investigates if weaker models (non-experts) can assess the correctness of stronger models (experts) in a debate setting.
- Experiments conducted on the QuALITY dataset show that debate helps both non-expert models and humans improve their accuracy.
- Optimizing debaters for persuasiveness improves non-expert judges' ability to identify the truth in debates.
- Results imply that debate can effectively align models in the absence of ground truth.

Introduction

- Current LLM alignment relies heavily on human-labelled data.
- As LLM sophistication increases, human expertise alone will not suffice for alignment.
- Need scalable oversight where weaker models or non-experts can evaluate stronger models or experts.
- Investigate if debates between models can serve this purpose.

Motivation

- Assessing AI by humans becomes impractical as models surpass human capabilities.
- Explore an alignment method where weaker systems judge stronger, information-rich debaters.
- Debate setup: Two expert models argue for different answers; a non-expert judge selects the likely correct answer.

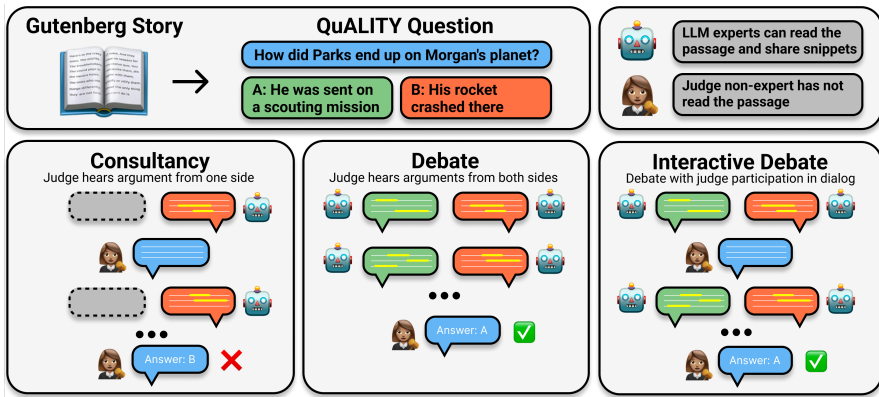


Figure: Overview of Debate Protocols including Debate, Interactive Debate, and Consultancy.

Experimental Details

- Task: QuALITY reading comprehension subset with hard questions.
- Information asymmetry: Judges are non-experts without full text access.
- All models have access to a quote verification tool.
- Debating models include GPT-4-Turbo, GPT-3.5-Turbo, Claude 2.1, Claude 1.3, Mixtral $8\times 7B$.

Metrics for Evaluation

- Win Rate: Frequency of the judge selecting a specific debater's answer.
- Elo Rating: Aggregate score depicting the persuasiveness of debaters.
- Judge Accuracy: Accuracy of judge in selecting the correct answer.

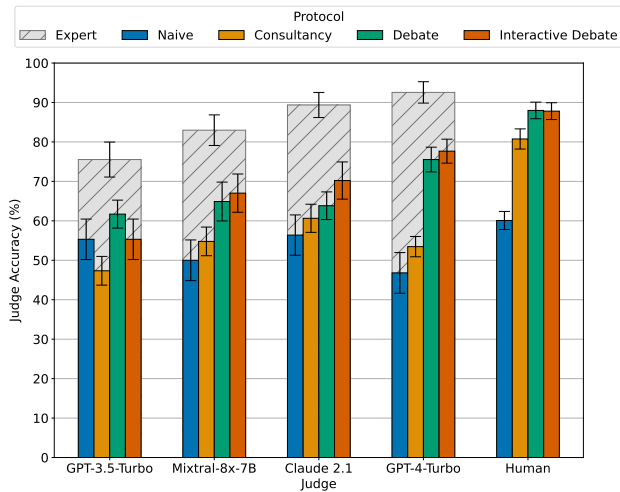


Figure: Performance of different protocols including baseline, consultancy, naive, and debate protocols.

Results with LLM Judges

- Debate improves non-expert judge accuracy significantly (76%) vs naive (48%).
- Elo and aggregate ratings show stronger correlation for more persuasive debaters.
- Consultancy degrades judge accuracy when optimizing consultants, unlike debate.

Results with Human Judges

- Debate protocols produce higher accuracy (88%) vs consultancy (78%).
- Static and interactive debate show equivalent performance.
- Human judges are better calibrated in debate than in consultancy.
- Debaters are bottleneck to further accuracy improvement.

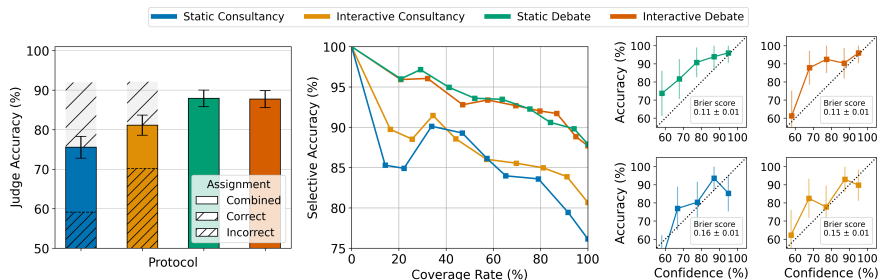


Figure: Comparison of judge accuracy, selective accuracy, and calibration performance for human judges across different protocols.

Insight on Debate Effectiveness

- Judges perform better with debates due to adversarial nature providing clearer correct information.
- Debate mechanisms are less susceptible to misleading information as compared to consultancy.
- Adding human interaction had no significant effect on judge accuracy.

Conclusion

- Debate has potential for scalable oversight without ground truth labels.
- Effective in aligning models by optimizing for persuasiveness.
- Further opportunities in enhancing debater capabilities and seamless integration of human and AI decision-making processes.

Future Directions

- Explore debate protocols in other domains with diverse evidence substantiation.
- Investigate applying debate to more complex tasks requiring stronger model reasoning abilities.
- Develop advanced quoting and verification tools to bridge gaps in evidence evaluation.