

Abstract

This paper explores the potential of pre-existing large language models (LLMs), such as ChatGPT, Claude, and Gemini, to assist in judging Original Oratory (OO), a category in competitive speech and debate, although in the future it should expand into other speech and debate events. As judging in speech events often involves subjective assessments of content, structure, delivery, and rhetorical impact, this study evaluates whether current LLMs can offer reliable and consistent feedback aligned with their human counterparts. The experiment involved inputting transcripts of OO speeches into various LLMs and comparing their evaluations against those of humans who participate in the activity, using a standardized rubric to evaluate the speeches. The results show that while LLMs can effectively identify key structural elements, their performance is less consistent than their human counterparts, tending to overrate the speeches and struggling to accurately judge emotional appeal. The findings suggest that, with refinement, LLMs could serve as valuable tools for preliminary judge feedback or judge training, although they are not ready to be used as substitutes for human adjudication in high-stakes competition settings. The outcome of this paper illustrates that if an LLM model is deployed into the field of speech and debate, then it will be able to reduce subjectivity, as well as make the field much more efficient, because it is trained without the same personal biases as its human counterparts.

Introduction

PARTICIPATION

**140,000
Students
Nationally**

JUDGING BURDEN

**Increasing
Parental
Fatigue**

FEEDBACK GAP

**Lack of
Relevant
Feedback**

RESEARCH QUESTION

If an LLM model is deployed into the field of speech and debate, then will it be able to reduce subjectivity, as well as save time, money and energy (therefore making judging in the field more efficient)?

BACKGROUND LITERATURE

Computational Argumentation¹

Self-Refine³

Methods

01 INPUT

7 Speeches
Original Oratory
Transcripts

02 EVALUATE

Fixed Prompt
Same NIETOC
Judging Prompt
for Each Model

03 COMPARE

**Scores &
Feedback**
*Against Human
Judging*

MODELS TESTED

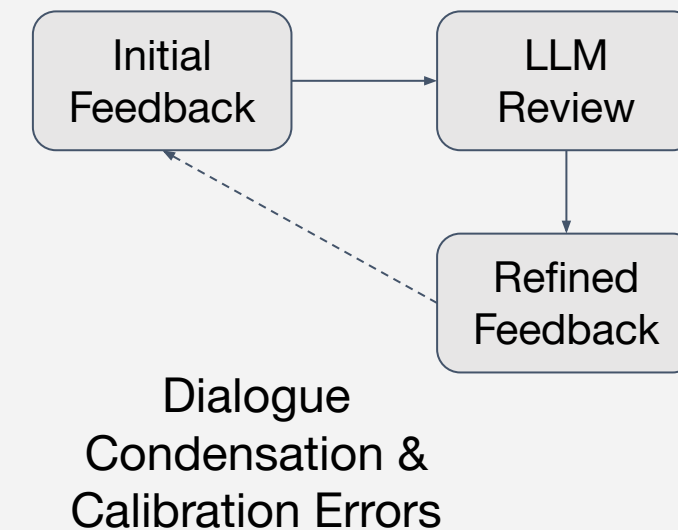
Claude
Sonnet 4

ChatGPT
GPT-4.1

Gemini
Flash 2.5

Perplexity
Sonar

FEEDBACK LOOP



4 CATEGORY RUBRIC

Message / Purpose

**Content
Development**

**Structure /
Organization**

**Emotional Appeal /
Personal Anecdotes**

Scores From

1 (Poor)
...
5 (Excellent)

20 Points Total

Results

Effect of Model on Average Score

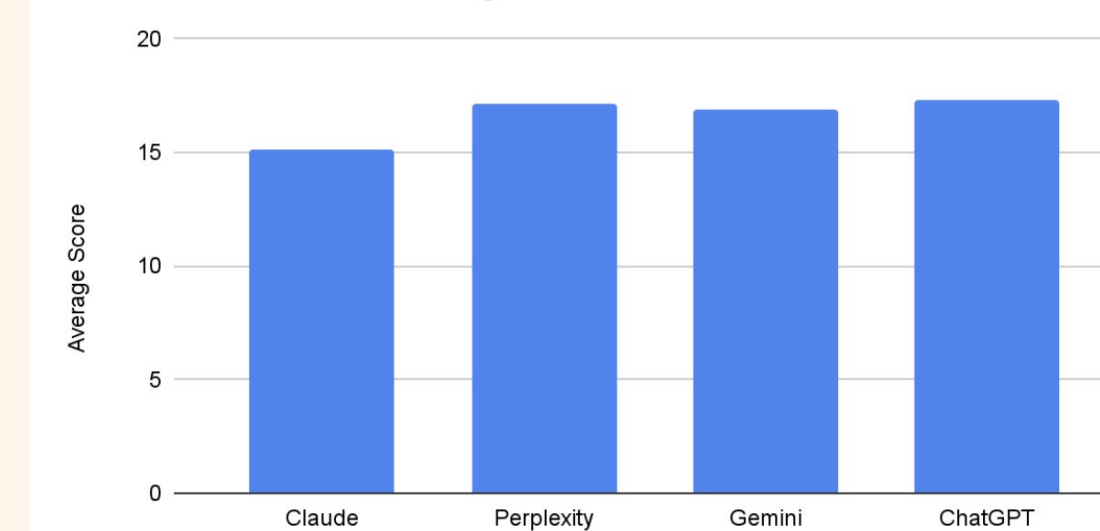


Figure 1: Effect of Model on Average Score

Difference in Human Evaluation vs. LLM Evaluation

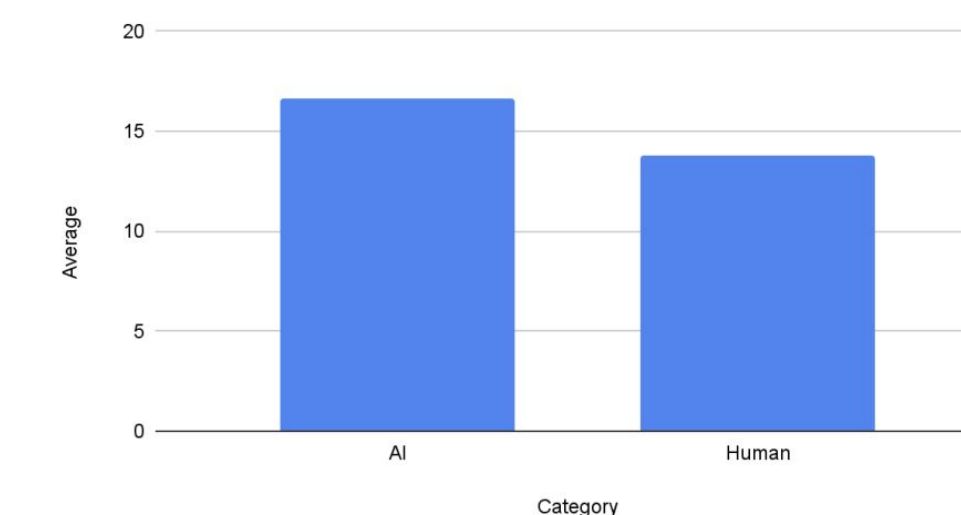


Figure 2: Difference in Human Evaluation vs. LLM Evaluation

Effect of Model on Inference Time

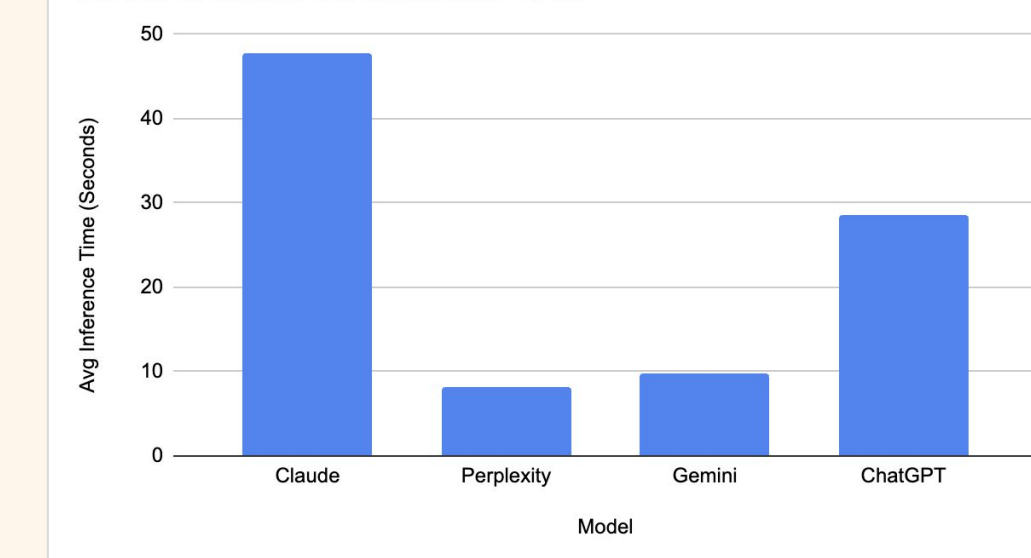


Figure 3: Effect of Model On Inference Time

SPEECH EXCERPT WITH POSSIBLE RUBRIC EVIDENCE

The statistics are heartbreaking. According to the National Alliance on Mental Illness, more than 60% of people with mental health conditions don't seek help, because to them? Silence feels easier than vulnerability. It's easier to say, "Hey, I'm fine" than to admit, "I'm not okay." I know I said, 'self-imposing', but...what's hard to admit is that...this silence inevitably hurts others, too. Every year, 64% of bullying goes unreported on school grounds. An 8 year old girl was one out of millions that fell victim to this statistic. That girl was from my elementary school...

Highlighted passages are illustrative cues, not recorded explanations from the models.

Message / purpose

Central claim about
silence and vulnerability

Content development

Two statistics are used
as supporting evidence

Emotional appeal

Quoted inner thoughts and
a personal connection

Structure / organization

Broad statistics lead to
a personal example

SUPPLIED EXAMPLE SCORES

	Human avg	Claude	ChatGPT	Gemini	Perplexity
● Message / purpose	4.78	3.93	4.36	4.44	4.07
● Content development	3.39	3.96	4.32	4.41	4.07
● Emotional appeal / anecdotes	2.83	3.14	3.68	3.79	3.71
● Structure / organization	3.54	4.10	4.46	4.71	4.36

Discussion

Inter-Rater Reliability

Inference Times

Differences with
Humans

Model Calibration

Conclusion

How can we ensure that AI models understand argument quality beyond surface-level features like fluency or vocabulary?

To what extent will AI models be able to evaluate spontaneous or impromptu debate formats that lack structure?

Will AI be able to condense the dialogue of longer debate formats, especially since some debates can reach up to two and a half hours?

What are the key challenges and considerations in transforming an AI-based debate evaluation system into a scalable and user-friendly product for use in educational or competitive settings?

References

- Bar, R., Liat, H., Matan, E.-D., Venezian, O., & Slonim, N. (2021). Advances in Debating Technologies: Building AI That Can Debate Humans (pp. 1–5). <https://aclanthology.org/2021.acl-tutorials.1.pdf> Lee, D. (2024b). AI-Powered Writing Support: Writing Feedback. Meaningful Approaches to Learning and
- Teaching in the Age of AI: 2024 Festival of Learning and Teaching. https://www.researchgate.net/publication/381473241_AI-Powered_Writing_Support_Writing_Feedback
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoy, S., Yang, Y., Gupta, S., Majumder, B., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (n.d.). SELF-REFINE: Iterative Refinement with Self-Feedback. <https://arxiv.org/pdf/2303.17651>