OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Evaluating Original Oratory with Large Language Models: A Comparative Study of AI and Human Judging

Yatharth Sathya
yatharth.sathya@gmail.com

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

Speech and debate is a long-standing educational activity that improves many skills, such as communication, confidence and teamwork. Across high schools in the United States, events like speech tournaments and debate competitions serve as valuable platforms for students to express their ideas, argue their viewpoints, and develop real-world skills. According to the National Speech & Debate Association (NSDA), more than 140,000 students participate annually in speech and debate.

Among the many events offered, Original Oratory is a very popular speech category where students craft and deliver persuasive, memorized speeches on topics of their choice. This study focuses specifically on the Original Oratory event as a starting point to explore whether a pre-existing large language model (LLM) can evaluate student speeches with the same level of accuracy and consistency as human judges.

The goal is to determine whether AI can support or even enhance the judging process, particularly in environments where experienced human judges are lacking.

While this project centers on Original Oratory, the underlying approach has the potential to expand into debate events, offering broader benefits across the entire speech and debate community. By using an LLM to assist in judging, we can save countless hours of human labor and reduce the ongoing need to train new judges, often saving a large amount of time, energy and money.

Furthermore, integrating AI in this field can address one of the most persistent challenges in speech and debate competitions: human subjectivity. Even among experienced and qualified judges, evaluations can be influenced by unconscious personal biases, such as preferences for certain speaking styles, topics, or delivery methods. Inconsistent scoring may additionally result from differences in judges' backgrounds, expectations, or fatigue during long tournament days. These subjective factors can lead to unfair outcomes and discourage student participation. On the other hand, an LLM has the potential to judge with less biases on predefined criteria, reducing variability and ensuring a more objective, equitable, and transparent evaluation process. While we recognize that these LLMs might be trained with some personal biases, as they might encode societal and data biases, they may reduce certain forms of inconsistency inherent in human judging, such as judge fatigue or stylistic preferences. This could ultimately make speech and debate more accessible and enjoyable for all students.

One of the main references we're looking at in this project is a research paper called *Advances in Debating Technologies: Building AI That Can Debate Humans*. It was written by Roy Bar-Haim, Liat Ein-Dor, Matan Orbach, Elad Venezian, and Noam Slonim as part of IBM's research program. The paper focuses on the idea of developing artificial intelligence (AI) that is able to debate against its human counterparts. As AI continues to grow and improve, especially in the area of natural language processing (NLP), researchers have started exploring something called *computational argumentation*. This is a subfield of NLP that looks at how computers can be trained to understand, analyze, and even create arguments like a human. This reference paper further goes on to prove that people have already started tackling the big question of whether AI can truly engage in argument-based tasks, not just by understanding language, but also using logic and reasoning in ways comparable to humans. This makes it an important reference for our own project, which focuses on using AI to evaluate speech and debate performances.

Our secondary reference will be the paper *SELF-REFINE: Iterative Refinement with Self-Feedback*, completed by many different researchers (from institutions such as Carnegie Mellon University, University of Washington, NVIDIA, UC San Diego, and more). This paper discusses a large language model (LLM) that can refine itself after multiple attempts. This technology was tried during the experiment, as we ran a feedback loop using multiple models, to ensure better feedback for the user, therefore increasing user experience. Only through trial and error can the large language models (LLMs) produce optimal results. The paper then goes on to present a product called *Self-Refine*, which is a large language model that relies on an iterative self-refinement algorithm. This algorithm alternates between two stages, *Feedback* and *Refine*. In our project, we explore the concept of a similar feedback loop, with

one model producing text and the second refining the text of the primary model. The first model will then integrate the feedback and output an edited response, which ensures better user experience.

Our final reference comes from a program conducted by a leading Australian university, the University of Adelaide, which tested an AI-driven writing feedback tool with students in its program. Students submitted reflective pieces on their experience using the tool. The majority of the reflections were positive, noting the tool's ability to enhance academic writing, particularly in promoting a more analytical and rigorous style. The students felt that the AI's feedback helped them understand how to improve their work. Some suggested adding a rubric to help the tool better judge assignment expectations. In conclusion, the study concluded that the AI tool is technically proficient and capable of supporting long-term learning. This supports the premise of our project, where LLMs are used to evaluate Original Oratory performances: just as LLMs can aid in refining academic writing through targeted feedback, these same LLMs can play a similar role in speech and debate, offering constructive analysis that encourages long-term speaker improvement.

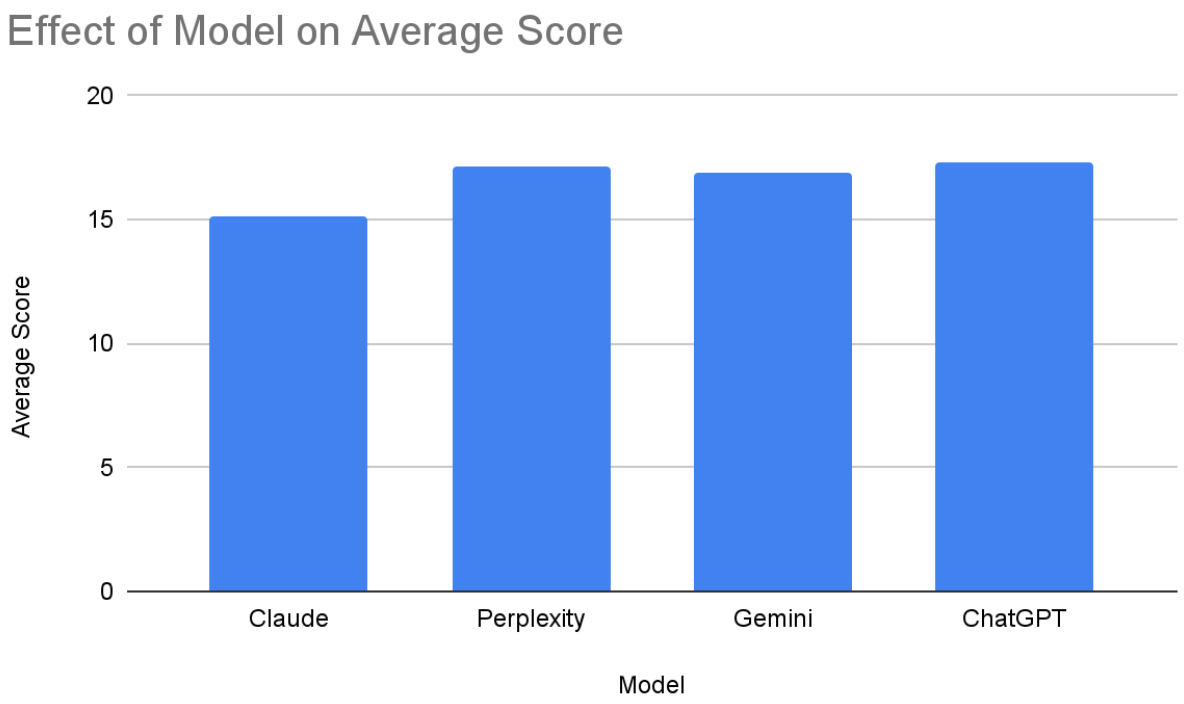
Our hypothesis for this project is that if an LLM model is deployed into the field of speech and debate, then it will be able to reduce subjectivity, as well as save time, money and energy (therefore making judging in the field more efficient), because it is trained without the same personal bias as its human counterparts, and it can evaluate arguments based on consistent criteria, process large amounts of information rapidly, and deliver objective feedback at scale.

RESULTS

This study compared evaluations of Original Oratory (OO) speeches generated by four large language models (LLMs), ChatGPT, Claude, Gemini, and Perplexity, to those provided by human judges. Each LLM was asked to rate the speeches using a predefined rubric (Message/Purpose, Content Development, Emotional Appeal/Personal Anecdotes, and Structure and Organization), with scores ranging from 1 (lowest) to 5 (highest) per category (maximum of 20 points per speech).

Overall Scoring:

	Claude	Gemini	ChatGPT	Perplexity	Human
Average Score (/20)	15.14	16.85	17.32	17.18	16.32
Highest Score	18	18.5	19	18.5	18.25
Lowest Score	8.25	14.5	11	8.75	9.5



	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/Personal Anecdotes	2.83	3.14	3.68	3.79	3.71
Structure/Organization	3.54	4.1	4.46	4.71	4.36

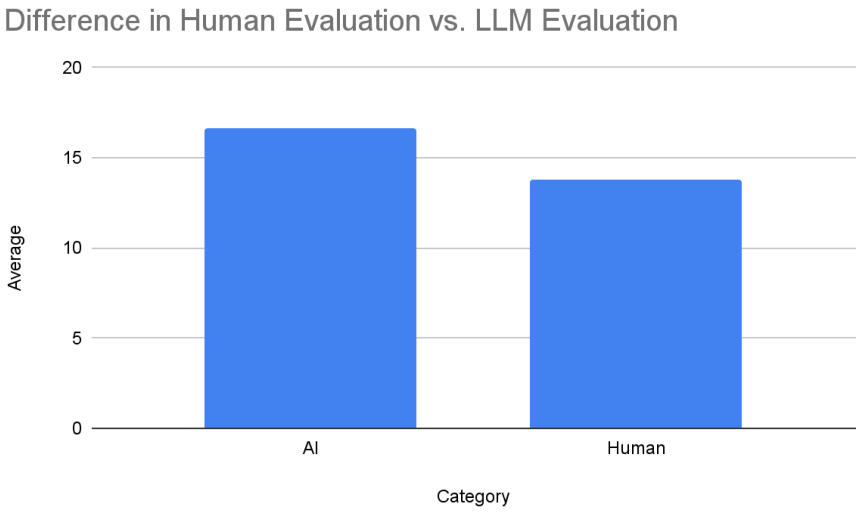
Claude provided the most conservative feedback, frequently noting issues with emotional resonance and clarity, which aligned closely with stricter human judge standards, who have previously judged circuit tournaments (such as the National Individual Event Tournament of Champions, also known as NIETOC). Its individual ratings by category were the highest observed compared to any other model.

Gemini was the most generous, often praising speeches without sufficient critique, which could mislead less experienced speakers. This suggests that Gemini will most likely not work practically. One note is that Gemini flagged two OOs for inappropriate content, which is why it has a lower average score

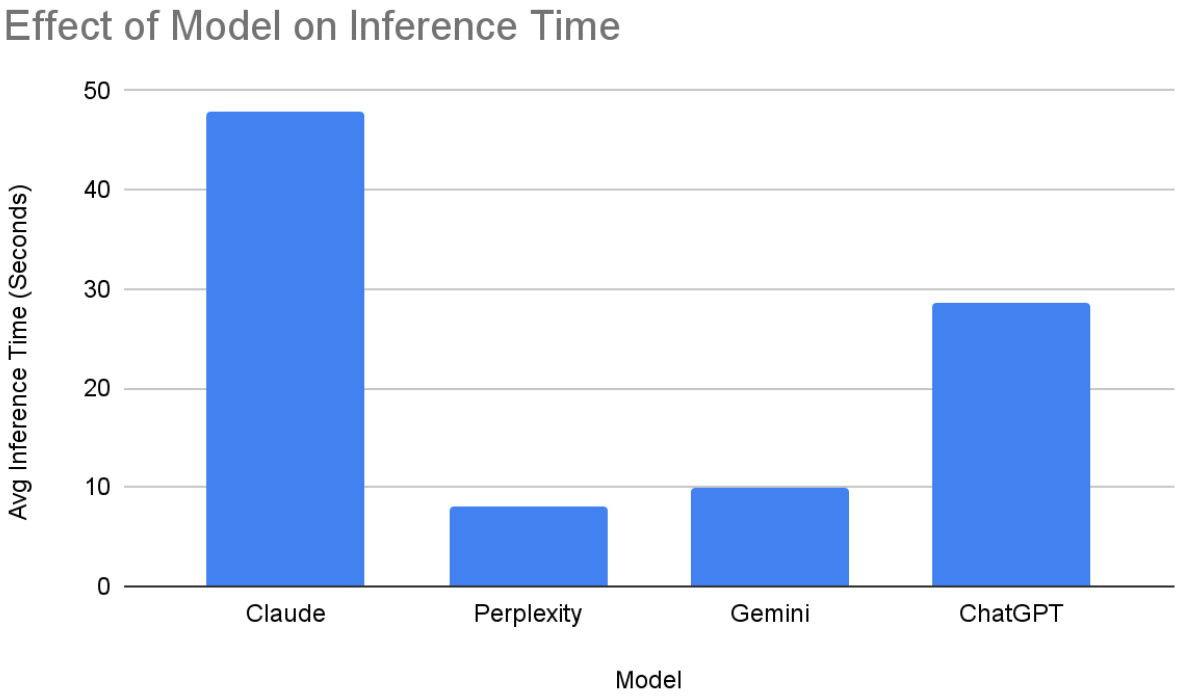
(because the OOs that it flagged received higher scores from each of the other models individually).

ChatGPT delivered balanced feedback, generally matching human feedback in tone and content, but occasionally overestimated the emotional impact. One factor at this specific point is that ChatGPT's scores generally offered a middle ground between the other models.

Perplexity offered feedback, and had very quick inference times, but the feedback often lacked quality and necessity that was seen by the other models.



The human evaluation was taken from three judges, who are experienced in the field of speech and debate judging with over 20 years of experience between all three of them. This difference in judging, with all of them having judged in the activity for quite some time prior, led to less inter-rater reliability, which showcased the inconsistencies in subjective human judging.



One note is that all of the models were evaluated to produce around the same word counts.

DISCUSSION

The findings from our study highlight both the potential and the limitations of using large language models (LLMs) to evaluate Original Oratory (OO) speeches (although this can expand to include other speech and debate events). While the models were able to consistently identify structural elements and provide constructive feedback, they varied significantly in their scoring tendencies and demonstrated notable weaknesses in evaluating subjective components, particularly emotional appeal. Also, an observation is that the amount of feedback that the models gave varied a lot, leading to many discrepancies in inference time.

Across all four models tested, Claude, Gemini, ChatGPT, and Perplexity, there were many similarities in the core parts of feedback that they provided. Categories such as *Message/Purpose* and *Structure/Organization* were generally evaluated with reliability and consistency, with few discrepancies from model to model. This suggests that LLMs are capable of recognizing logical flow, argumentative coherence, and speech organization, which are necessary in speech judging.

Moreover, all models produced usable written feedback that could help speakers identify areas for improvement, especially in the context of writing their speeches. This shows promise for using LLMs as supplementary tools, especially for preliminary rounds, speaker coaching, or judge training in environments where experienced human judges are limited.

The weakest category across all models was Emotional Appeal/Personal Anecdotes. While human judges could identify nuanced emotional moments and assess their authenticity and impact, LLMs often misinterpreted or overlooked these elements. They also failed to provide adequate feedback for these areas of the speech. This is probably due to the lack of true emotions in these AI models. LLMs rely entirely on text patterns and they struggle to evaluate genuine pathos or delivery-based persuasion.

Also, these models had a tendency to overrate speeches, especially Gemini. These inflated scores highlight a challenge in aligning AI evaluation standards with the stricter expectations common in competitive environments, especially high-level tournaments with very experienced judges. Claude's lower scoring average suggests it may be better suited for real judging, though further testing is needed to provide a surefire conclusion.

While we recognize that only four closed source LLM and no open-source LLM models like Llama, Qwen, DeepSeek family models were used, we would like to share that this was intentional. Although these models might be more popular or open-source, we intentionally chose these models as primary items. In future studies, this can expand to several models, yet we only chose these models as they were accessible to us during the experiment.

MATERIALS AND METHODS

The concept of a self-feedback loop, particularly in the area of LLMs and NLP, refers to the cycle in which a speaker listens to, processes, and adjusts their own speech output in real time. It plays a fundamental role in effective communication, speech correction, and public speaking. Similarly, in the field of LLMs and NLP, the LLM will produce the feedback, which will be evaluated and reviewed by another LLM model or itself in a separate prompt window, and be received by the first LLM. This, in turn, allows feedback to be generated much better, although it will take a longer time to produce.

	1 - Poor	2 - Fair	3 - Good	4 - Very Good	5 - Excellent

Message/Purpose					
Content Development					
Emotional Appeal/Personal Anecdotes					
Structure and Organization					

Just like self-feedback loops in LLMs improve clarity through evaluation, the rubric for Original Oratory provides a structured framework. The rubric is designed to provide a clear, objective framework for evaluating speeches in the Original Oratory event, minimizing subjectivity in the judging process. It focuses on four key categories: message/purpose, content development, emotional appeal/personal anecdotes, and structure/organization. Each category is rated on a scale from 1 (Poor) to 5 (Excellent), allowing the AI to assess the clarity of the speaker's central message, the strength and logic of their arguments (ethos and logos), the effectiveness of their emotional engagement and personal storytelling (pathos), and the overall organization and flow of the speech. This structured evaluation aligns with the National Speech & Debate Association's goals for Original Oratory, which emphasize persuasive communication through a balance of logic, emotion, and clear structure. Ultimately, the rubric ensures consistency and fairness in competition while guiding students toward impactful and well-crafted speeches.

Because speech evaluation naturally involves some subjectivity, especially when it comes to emotional impact, delivery style, and personal resonance, the rubric serves to reduce this subjectivity by grounding judgments in specific, consistent criteria.

Another key portion of this paper is dialogue condensation. This refers to the process of reducing the length and complexity of dialogue by extracting some parts and retaining only the most necessary parts of the conversation. This process is particularly critical when working with large language models (LLMs), many of which are constrained by input token limits and computational efficiency requirements. Inference time is directly impacted by the size and complexity of the input data. Longer dialogues, especially those filled with redundant or irrelevant information, increase the computational burden on the model. This not only slows down response generation but also places a strain on the model, which can worsen user experience.

Also, some models cannot handle more textual input than a predefined max token length. If the dialogue exceeds the pre-existing max token length, the model will not be able to effectively analyze the text. So, dialogue condensation will allow the core portions of the dialogue to be preserved, while eliminating excess text, and allowing the inference times of the model to decrease. In summary, dialogue condensation is a crucial optimization technique that improves the operational performance of language

models. By minimizing the size of input while preserving essential content, it ensures that inference remains fast and efficient, meeting both the technical and user experience requirements of modern AI systems. An observation while completing this was that dialogue condensation can be flawed, as it might eliminate key components from the speech, resulting in lower scores for the speakers. However, it is our best bet to ensure a greater experience for our users.
Calibration error is the difference between a recorded reading and the true value of the measured quantity. While calibration error was another metric that we considered implementing in this paper, the field of automated text scoring is varied between different judges and cannot be constrained by a "true value", so we decided to not use it.

In addition to calibration errors, we considered the role of confidence intervals when interpreting the results of our model comparisons. A confidence interval provides a statistical range within which the true mean score is likely to fall at a given level of certainty. These intervals can help determine whether observed differences between models are statistically meaningful or simply due to random variation. However, we ultimately decided not to include confidence intervals in our analysis because our sample size was relatively limited. The focus of this study was to examine practical scoring tendencies across models under consistent prompting conditions, not to generalize findings to a broader population of performances. As a result, reporting averages and qualitative trends was more aligned with the scope of our project, while a full statistical inference framework (similar to utilizing confidence intervals as a metric) would be better suited for a larger-scale, more controlled study.

Throughout this project, multiple different types of models were utilized to provide feedback on different types of speeches. The models used were Claude, Perplexity, ChatGPT and Gemini. All of them were chosen due to the fact that they were closed-source and easily accessible models. However, each model had notable discrepancies, such as Claude, which consistently gave lower scores than the other models, with an average score of 14.96 out of 20. Other models, such as Gemini consistently gave higher scores, with an average score of 17.3 out of 20. Our personal opinion agreed with multiple models on various occasions, most accurately coming closest to the scores and qualitative feedback we would have given based on my subjective evaluation. Yet, in the real world of speech and debate judging, qualified and experienced judges will be stricter, making it more difficult for the competitors. For this reason, Claude will be the best model for practical use, as its stricter evaluations more closely reflect the standards expected in actual judging scenarios. This is because Claude's cumulative average score of 15.14 was extremely close to the average human evaluation score of 14.54, displaying its potential usage in the field. Another note we would like to add is that the prompt that the model is provided will most likely result in a difference of scores, but the prompt we fed into the models remained constant throughout this experiment (Prompt: Grade this OO based on the below rubric (feel free to use the decimals 0.25, 0.5, 0.75 and grade as a nietoc judge)). When we added that the AI should grade as a regular league judge, instead of a NIETOC circuit judge, the scores increased drastically, showing how the content of the prompt is crucial in ensuring good results.

In total, seven Original Oratory speeches were evaluated across all four models. The speeches varied in length and topic, reflecting a range of competitive styles. Some were authentic OOs written by student

competitors, while others were synthetically generated to broaden the dataset. Each speech was submitted to every model using the same fixed prompt: 'Grade this OO based on the below rubric (feel free to use the decimals 0.25, 0.5, 0.75 and grade as a NIETOC judge).' No temperature adjustments or custom system prompts were applied beyond this input, and each speech was evaluated in a single run per model with no multi-run averaging, meaning LLM output variance across runs was not controlled for. When Gemini flagged two speeches and declined to score them, those speeches were excluded from Gemini's average calculations rather than replaced, which should be noted as a source of incomparability when interpreting Gemini's aggregate scores relative to the other models.

We intentionally did not optimize for human agreement, as the purpose of including multiple human evaluators was to capture the natural subjectivity inherent in speech judging rather than to enforce consensus. High inter-rater disagreement illustrates the very problem our study aims to address, that human judging is inconsistent, and strengthens the motivation for exploring AI-based evaluation.

CONCLUSION

This study demonstrates the potential of large language models (LLMs) like ChatGPT, Claude, Gemini, and Perplexity to assist in evaluating Original Oratory speeches. While LLMs could judge structure, content, and the themes of the speech, they couldn’t properly evaluate emotional appeal, which is an essential aspect of effective public speaking. The discrepancies between models, especially in lenience during scoring, reflect the influence of their training data and tuning objectives.

Although these models are not yet reliable enough to replace human judges in formal competition settings, they can serve as valuable tools for feedback, practice evaluation, and judge training. Claude, with its stricter and more aligned scoring, shows a lot of promise for being deployed in the real world. However, the experiment also revealed limitations, particularly in how dialogue condensation can affect the quality of AI-based evaluation.

Some questions that we might have in the future of this topic would be the following: How can we ensure that AI models understand argument quality beyond surface-level features like fluency or vocabulary? (as emotion and other statistics will add to the quality of the arguments), To what extent will AI models be able to evaluate spontaneous or impromptu debate formats that lack structure? (as the models will have less preparation on how to judge these events), 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? (refers to dialogue condensation which was aforementioned in the materials and methods section), and 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? (referring to its practicality as a real world product).

With further refinement, such as incorporating speech delivery analysis, better training on speech-specific datasets, and improved feedback loops, LLMs could play a transformative role in making speech and

debate more equitable, accessible, and scalable. Hopefully, in the future this can be used as a practical product that can act in place of a qualified judge. As AI continues to evolve, so does its potential to support educational spaces in meaningful, practical ways.

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., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, 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

ACKNOWLEDGEMENTS

Special thanks to Mark Lee and Rimi Putti for generously sharing their Original Oratory speeches, which served as essential material for training and analysis to this experiment.

Also, we are very grateful to Aarav Bavishi, Michaela Northrop and Mark Woodhead for serving as human evaluators, as their insights and scoring contributed directly to the comparative analysis of this study.