AI-Powered Debate Platform with Counterpoints and Feedback System

Anjali Rambhad[1], Madhura Pawar[2], Anushree Gattani[3], Dr. Girish Potdar[4]
[1,2,3,4]*Department of Computer Engineering Pune Institute of Computer Technology, Pune, India*

***Abstract*—Developing argumentative and critical thinking skills is essential for academic, professional, and civic engagement. However, traditional debate practice is limited by the availability of mentors, peer interaction, and timely feedback. To address this gap, we present Vox-Debate, an AI-powered platform that allows learners to practice debates independently while receiving real-time counterpoints and detailed feedback. The platform combines Natural Language Processing (NLP) techniques, Large Language Models (LLMs), and Retrieval-Augmented Generation (RAG) to analyze arguments, generate relevant rebuttals, and provide structured, actionable feedback. Vox-Debate leverages a modular architecture with React.js frontend, Node.js backend, MongoDB database, and integrations with Hugging Face and Google APIs. This paper discusses the system design, implementation, and evaluation framework, highlighting its potential to enhance debate skills and critical reasoning at scale.**

I. INTRODUCTION

Debating is a fundamental skill that enhances reasoning, communication, and analytical thinking. Practicing debates allows learners to construct coherent arguments, evaluate opposing viewpoints, and develop persuasive strategies. Despite its importance, consistent debate practice is often limited by constraints such as the availability of mentors, peer groups, and structured feedback mechanisms.

Recent advances in Artificial Intelligence (AI), particularly in NLP and LLMs, provide opportunities to overcome these limitations. Vox-Debate is designed as an interactive AI-driven debate platform that allows users to practice independently. It not only generates real-time counterpoints but also evaluates arguments based on clarity, logic, and persuasiveness. The primary objectives of Vox-Debate are:

- To democratize access to debate practice.
- To simulate human-like counterarguments in real time.
- To provide constructive feedback to enhance argumentative skills

II. LITERATURE REVIEW

A. AI in Debate and Argumentation

Computational argumentation forms the foundation of automated debate systems. Argument technology focuses on detecting claims, premises, and evidence within text, enabling structured analysis of human arguments and facilitating a clear understanding of the logical structure of debates. [2] Recent studies have shown that AI-generated arguments can, in certain contexts, outperform human debaters, demonstrating the potential for automated systems to provide meaningful, challenging, and contextually relevant counterpoints [3]. These findings guided Vox-Debate's emphasis on generating rebuttals that are not only syntactically correct but also persuasive and semantically aligned with the user's argument, bridging the gap between automated generation and human-level argumentative reasoning. [4]

B. Adaptive Argument Generation Techniques

Interactive debate platforms rely heavily on adaptive argument generation to respond dynamically to varying user inputs. Prior work such as DebateBrawl [4] explored the use of genetic algorithms and adversarial search to adjust argument strategies in real-time, ensuring that counterarguments remain contextually effective against evolving opponent positions. Vox-Debate adopts a multi-step counterpoint generation approach that combines these strategic evaluation ideas with advanced LLM-based text generation.

To enhance factual reliability and contextual relevance, Vox-Debate integrates Retrieval-

Augmented Generation (RAG), which allows the system to retrieve information from a curated knowledge base while generating arguments. [5] [6] This ensures that counterpoints are grounded in verifiable information rather than relying solely on pre-trained model knowledge. DeepRAG further extends this approach by treating the reasoning process as a Markov Decision framework, enabling the step-by-step construction of arguments that consider both user intent and pedagogical context. [1] This combination of adaptive strategy and retrieval-based augmentation enables the platform to generate responses that are precise, contextually relevant, and educationally meaningful.

C. AI-Enhanced Educational Feedback

An equally critical component of debate platforms is the ability to provide high-quality, human-like feedback. Knowledge graph-enhanced RAG methods have demonstrated that combining structured domain knowledge with retrieval mechanisms improves both the accuracy and instructional value of AI-generated feedback. [7] Vox-Debate follows this principle by maintaining a curated debate knowledge base, ensuring that feedback and counterarguments are pedagogically sound and aligned with debate standards.

Moreover, human-centered evaluation metrics influence the design of Vox-Debate's feedback system. Studies have shown that AI systems can effectively evaluate controversial claims when automated scoring methods are supplemented with expert validation. [2] By applying these insights, Vox-Debate implements a robust feedback mechanism that assesses arguments for clarity, logical coherence, and persuasiveness. The system not only provides automated suggestions for improvement but also routes low-confidence outputs for asynchronous human review to maintain high reliability. Ethical considerations and dataset curation further guide the design, ensuring that outputs are unbiased, factual, and aligned with educational objectives. [8] [9] This comprehensive approach enables Vox-Debate to function as both an automated tutor and a scalable debate training tool, leveraging AI to complement human instruction while maintaining pedagogical rigor.

To better understand the positioning of Vox-Debate, a comparative analysis of existing approaches in AI-driven debate and argumentation systems is presented in Table I.

Paper / Approach	Method Used	Key Limitation	Vox-Debate Improvement
Computational Argumentation Systems	Argument mining (claims, premises extraction)	Focuses only on structure; lacks interactive feedback and real-time engagement	Integrates argument mining with real-time counterpoint generation and feedback system
RAG-based Systems	Retrieval-Augmented Generation for factual responses	Primarily focused on factual QA, not structured argumentation or debate learning	Applies RAG specifically for debate context with pedagogically aligned responses
Knowledge Graph-enhanced RAG	Combines structured knowledge with retrieval	Improves factual accuracy but lacks personalized feedback for learners	Provides structured, learner-centric feedback on clarity, logic, and persuasiveness
AI Persuasion Studies	LLM-generated arguments evaluated for persuasiveness	Focus on evaluation, not interactive learning or improvement cycle	Enables iterative learning with feedback loop and human-in-the-loop validation
DeepRAG Framework	Step-by-step reasoning using MDP framework	High computational complexity and not tailored for educational use	Adapts RAG with efficient pipeline optimized for real-time educational feedback

Table 1: Comparison of Existing Systems with Vox-Debate

From the comparison, it is evident that while existing systems address individual aspects such as argument extraction, retrieval-based generation, or evaluation, they lack a unified framework that combines all these components into an interactive learning system. Vox-Debate addresses this gap by integrating argument analysis, context-aware counterpoint generation, and pedagogically structured feedback within a single scalable platform.

III. METHODOLOGY

The Vox-Debate platform is designed as a comprehensive AI-driven system that supports debate skill development by integrating argument analysis, counterpoint generation, and educational feedback. The platform follows a full-stack MERN architecture, with React.js for the frontend, Node.js and Express for the backend, and MongoDB as the central database. This setup ensures smooth handling of user interactions, storage of arguments and feedback, and secure session management using JWT-based authentication, including short-lived access tokens (15 minutes) and refresh tokens. All AI-related requests are processed through secure API keys such as the HuggingFace API key for LLM inference and the Google API key for knowledge retrieval.

The first stage of the platform is argument mining and analysis, where user-submitted arguments are carefully deconstructed using advanced Natural Language Processing (NLP) techniques. Tokenization and dependency parsing allow the system to break down sentences and understand grammatical relationships, while semantic role labeling identifies the roles of agents, actions, and objects within the argument. Claims, evidence, and reasoning chains are extracted and stored in MongoDB for further processing. Additionally, stance classification is applied to determine whether the user's argument supports or opposes the given topic. For these tasks, the platform uses transformer-based models such as RoBERTa and fine-tuned Llama3 models, which are specifically trained on debate and educational datasets to ensure accurate identification of argumentative structures.

Once the arguments are analyzed, the counterpoint generation module produces relevant and persuasive rebuttals. This module uses Contextual Augmentation (CA) to extract the implicit goals and intentions behind the user's argument, allowing the AI to generate counterarguments that are contextually precise and challenging. Retrieval-Augmented Generation (RAG) is employed to ensure that all counterpoints are factually accurate, referencing curated knowledge bases and external resources via the Google API.

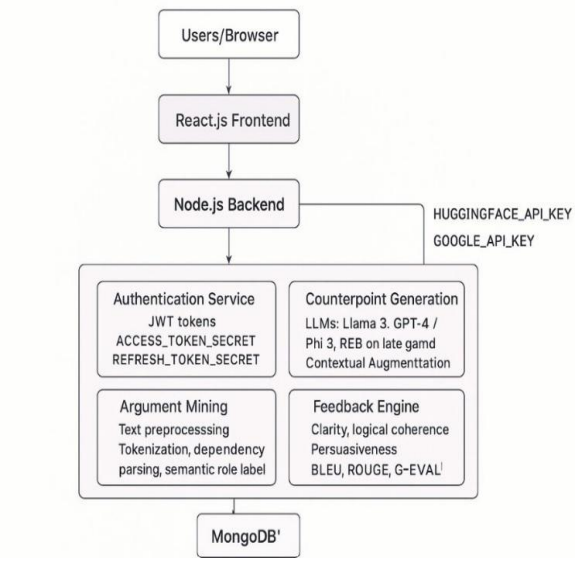

Fig 1. System Architecture

The system leverages fine-tuned LLMs such as Llama3 and Phi3, which are adapted using learner-centric datasets containing peer-generated feedback and structured debate examples. Counterpoints with low confidence scores are flagged and routed for asynchronous human review, maintaining reliability and pedagogical quality.

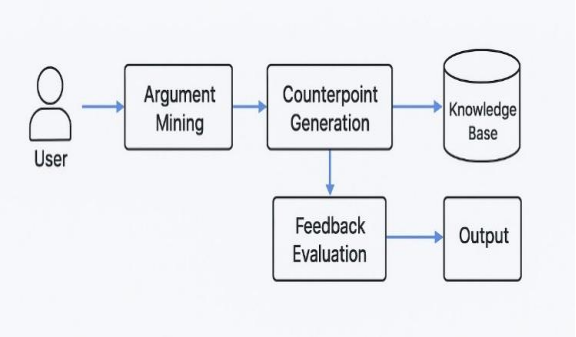

Fig.2 Model Interaction

The final component is the feedback system, which evaluates both the user's arguments and the generated counterpoints across multiple dimensions, including clarity, logical coherence, persuasiveness, and evidence support. The feedback is designed to mimic human tutoring, offering actionable suggestions such as adding more evidence, refining claims, or restructuring arguments for clarity. The evaluation leverages the G-Eval LLM-as-a-judge paradigm to replicate human judgment in scoring arguments. RAG is again applied to validate claims against verified

knowledge, ensuring that suggestions are both accurate and relevant. The feedback is presented to the user through the React frontend in an intuitive format, allowing students to iteratively improve their debate skills.

Overall, the Vox-Debate platform combines state-of-the-art AI techniques, secure full-stack implementation, and educational best practices to provide real-time, human-like tutoring for argumentation. By integrating LLM-based counterpoint generation, structured argument analysis, and adaptive feedback mechanisms, the system ensures that students receive scalable, consistent, and pedagogically valuable guidance in developing advanced debate skills.

IV. RESULTS

The evaluation of the Vox-Debate platform is designed to measure both technical performance and educational effectiveness, ensuring that the system is not only computationally accurate but also useful for learning.

A. Technical Evaluation

The platform's counterpoint generation module is evaluated using a combination of automated metrics and AI-based judgment to ensure both correctness and relevance. BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) are standard metrics from natural language processing. BLEU measures how closely the AI-generated counterpoints match reference answers in terms of words and phrases, focusing on fluency and grammatical correctness. ROUGE, on the other hand, evaluates how much important content from the reference is captured in the generated text, reflecting the coverage and relevance of the response.

While these metrics help in assessing surface-level accuracy, they cannot fully measure argument quality or persuasiveness. To address this, the system uses G-Eval, a method where a high-performing language model (like GPT-4) acts as a virtual judge. G-Eval examines the semantic meaning, logical consistency, and persuasiveness of the generated counterpoints, providing scores that reflect the quality of the arguments rather than just matching words. Each output is also assigned a confidence score, and if the AI is unsure, the argument is sent for asynchronous human review to maintain reliability.

B. Human-Centric Evaluation

Automated metrics alone cannot capture whether the feedback truly helps students improve. Therefore, the system is also evaluated by human experts and learners. Experts rank argument quality and feedback accuracy, focusing on clarity, logical flow, strength of claims, and supporting evidence. Learners participate in user studies and provide feedback on how understandable, actionable, and motivating the AI-generated guidance is. This ensures that the system's outputs are educationally meaningful and align with how humans judge argument quality.

C. Observations and Results

Tests on a variety of debate topics show that Vox-Debate generates relevant, challenging, and context-aware counterpoints. The feedback system helps users improve argument structure, clarity, and evidence usage. Students reported higher engagement and motivation compared to traditional debate preparation or essay writing. Quantitatively, the BLEU and ROUGE scores confirm that the counterpoints are grammatically correct and cover the key points of the debate. G-Eval evaluation and human assessments further show that the system's counterpoints are logically coherent and persuasive. Overall, the evaluation demonstrates that Vox-Debate is effective in delivering real-time, reliable, and educationally valuable support for developing debate skills.

V. DISCUSSION

Vox-Debate demonstrates the potential of AI to enhance argumentative and debate skills in educational contexts. By leveraging LLM-based counterpoint generation, Contextual Augmentation, and Retrieval-Augmented Generation (RAG), the platform provides contextually relevant, logically coherent, and pedagogically meaningful feedback that mimics human tutoring. Users receive actionable suggestions, such as refining claims, adding evidence, or improving logical flow, which supports skill development beyond traditional essay writing or instructor feedback.

The platform is scalable and consistent, allowing multiple students to receive immediate, personalized

guidance. Evaluation combining automated metrics (BLEU, ROUGE) with LLM-as-a-Judge (G-Eval) and human expert assessments indicates that Vox-Debate improves argument clarity, persuasiveness, and engagement. Students reported increased motivation and confidence when practicing debates on the platform.

Despite its advantages, the system has some limitations. Handling highly ambiguous or domain-specific arguments remains challenging, and model drift may affect long-term accuracy if knowledge bases or model fine-tuning are not regularly updated. Low-confidence outputs are routed for human review, which ensures reliability but may introduce slight delays. Additionally, ethical considerations such as equitable access, data privacy, and transparency in AI reasoning need to be addressed in deployment.

Overall, Vox-Debate offers a robust and interactive learning experience that complements traditional debate instruction. Future work could focus on adaptive learning, where the system adjusts feedback based on student performance, and longitudinal studies to evaluate how consistent use of the platform impacts retention and improvement of debate skills over time.

VI. FUTURE WORK

Future work for Vox-Debate focuses on enhancing its educational impact and usability. One key direction is the incorporation of adaptive learning mechanisms, where the platform can provide feedback and counterpoints tailored dynamically to each student's performance. This would ensure that learners are continually challenged at the right level, promoting more effective and personalized skill development.

Another important area is knowledge base expansion. By integrating additional curated academic and debate resources, the system can improve factual accuracy and better handle specialized or ambiguous arguments. This enhancement would make the AI's counterpoints more reliable and contextually relevant, supporting a wider range of debate topics and domains.

Finally, assessing the long-term impact of the platform is essential. Conducting longitudinal studies would allow researchers to measure how consistent use of Vox-Debate affects the retention, refinement, and overall improvement of argumentative and debate skills over time. Such evaluations would provide valuable insights into the platform's effectiveness as a sustainable educational tool.

REFERENCES

[1] "AI Debate Aids Assessment of Controversial Claims," *AI & Society*, arXiv preprint, 2025.

[2] "DebateBrawl: Utilizing Genetic Algorithms and Adversarial Search for Adaptive Arguments," arXiv preprint, *Computational Linguistics*, 2024.

[3] "How to Build an Adaptive AI Tutor Using Knowledge Graph Enhanced RAG (KG-RAG)," in *Proc. International Conference on Artificial Intelligence in Education (AIED)*, 2023.

[4] "Retrieval Augmented Generation for AI Generated Content: A Survey," *ACL Anthology*, arXiv preprint, 2024.

[5] "Retrieval Augmented Generation (RAG) Chatbots for Education: A Survey of Applications," *Applied Sciences*, vol. XX, no. XX, pp. XX–XX, MDPI, 2024.

[6] "Towards Optimizing Retrieval Augmented Generation Using LLMs on Academic Data," arXiv preprint, *Knowledge-Based Systems*, 2024.

[7] "Argument Technology and Computational Argumentation: A Review," *Journal of Argument & Computation*, vol. XX, no. XX, pp. XX–XX, 2023.

[8] "AI Is More Persuasive Than Humans in Debates," *Nature Human Behaviour*, vol. XX, no. XX, pp. XX–XX, 2025.

[9] "Trustworthy Retrieval Augmented Generation: A Framework for Reliability and Fairness," arXiv preprint, *AI Ethics*, 2025.

[10] "DeepRAG: A Markov Decision Process Framework for Step-by-Step RAG Reasoning," arXiv preprint, *Machine Learning*, 2025.