AN AI-DRIVEN DEBATE JUDGING SYSTEM USING EMOTIONAL AND CONTENT ANALYSIS BASED ON ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Leo Zhang [1], Carlos Gonzalez [2]

[1] Sage Hill High School, 20402 Newport Coast Dr, Newport Coast, CA 92657
[2] Computer Science Department, California State Polytechnic University, Pomona, CA 91768

ABSTRACT

Evaluating debates is a challenging task requiring nuanced understanding of abstract reasoning. Current AI systems struggle with these complexities, often providing shallow or biased feedback. To address this, we developed Blitz Debate, a Retrieval-Augmented Generation (RAG) system that combines large language models (LLMs) with semantic search capabilities [1][2]. Blitz Debate retrieves relevant external knowledge to evaluate debate arguments with depth and accuracy, offering structured, real-time feedback. Our experiments demonstrated 90.5% accuracy in identifying winners and superior interpretative responses compared to vanilla ChatGPT, highlighting its ability to provide evidence-based and nuanced analysis. Challenges included limited real-time reasoning and contextual depth, which we addressed through enhanced context modeling and adaptive argument generation. By offering scalable, unbiased, and context-aware feedback, Blitz Debate makes debate evaluation more effective and accessible, fostering critical thinking and argumentation skills for students, educators, and competitive debaters alike.

KEYWORDS

Retrieval-Augmented, System, Semantic Search, Language Models

1. INTRODUCTION

Debates are inherently complex, often requiring intricate arguments and moral considerations centered on philosophy and ethics. Evaluating debates involves understanding abstract concepts, moral reasoning, and value-driven arguments, which makes it challenging for AI systems to do so with limited contextual awareness [3]. This is a current problem we're experiencing with standard AI systems: these models often struggle to grasp the nuances of ethical arguments or subjective value judgments. In the context of debate performance evaluation, unbiased feedback is essential for providing clarity in assessments; a lack thereof risks participants receiving poor feedback and misguided responses. The challenges extend beyond AI systems, reflecting broader societal struggles with critical thinking. According to a survey by the Reboot Foundation, while 94% of people believe critical thinking is "extremely" or "very important," 86% find such skills lacking in the general public. Alarmingly, 60% of respondents reported never having studied critical thinking in school, and nearly a quarter indicated their skills had deteriorated since high school (Critical Thinking Survey Report, Reboot Foundation). This deficit underscores the need

David C. Wyld et al. (Eds): SIPO, BDHI, NET, SOEA, CSML, AISCA, MLIOB, DNLP – 2025
pp. 217-230, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.150413

for tools that enhance reasoning and argumentation, making robust, unbiased evaluation systems all the more essential in debate contexts.

Formal discussions through debates foster critical thinking and communication skills, which are essential for personal and professional growth. Research highlights the profound benefits of debate participation: a study of 3,500 students in the Boston Debate League found that involvement in debate improved ELA achievement by 68% of the average ninth-grade year's progress, with gains particularly evident in reading skills related to analysis and argumentation (Morrison, 2023). These benefits underscore the importance of making debate more accessible and effective for students. A well-designed system that plays into these benefits can help students identify areas for improvement and receive impartial assessments of their arguments. Leveraging external knowledge sources and sophisticated language models, such a system has the potential to enhance students' ability to construct coherent and persuasive arguments.

Slonim et al. (2021) introduced Project Debater, a system for engaging in competitive debates by retrieving evidence and delivering structured arguments. Despite its groundbreaking achievements, it struggles with real-time reasoning, creative rebuttals, and nuanced cultural or emotional contexts. Blitz Debate addresses these issues by integrating contextual sentiment analysis, adaptive reasoning mechanisms, and refined evidence selection, enabling more human-like and flexible analyses.

Rinott et al. (2015) focused on automated evidence detection, relying on supervised models trained on curated datasets to identify evidence supporting specific claims. While effective, this approach is limited by its dataset's quality and inability to handle nuanced, domain-specific arguments or logical structures. Blitz Debate builds on this by incorporating richer context modeling, domain adaptation, and advanced argumentation techniques, improving evidence retrieval and analysis depth.

Bar-Haim et al. (2021) developed a comprehensive debating framework, emphasizing argument mining, stance classification, and narrative generation. However, their system struggles with real-time rebuttals and implicit arguments, relying heavily on pre-defined taxonomies and structured corpora. Blitz Debate overcomes these constraints by integrating dynamic argument generation and contextual adaptability, producing more nuanced and flexible debate evaluations.

Introducing Blitz Debate, a Retrieval-Augmented Generation (RAG) model that evaluates debate arguments by combining pre-trained language models (LLMs) with semantic search capabilities from a vector database [4].

The RAG system aims to address the challenges of providing accurate and nuanced feedback by integrating external knowledge sources with AI capabilities. Semantic search features allow the system to retrieve contextual data relevant to the debate topic, ensuring that evaluations are rooted in comprehensive and precise information. The retrieved data is processed by OpenAI, which generates structured, context-aware feedback in real-time [5].

This approach is effective for debate evaluation due to its ability to combine scalability and depth. Unlike standalone AI tools, which often rely solely on LLMs and may produce shallow or inaccurate feedback, the RAG framework enhances the depth of analysis by incorporating external knowledge. Additionally, the real-time feedback mechanism makes the system especially useful for students preparing for debates, providing immediate feedback upon execution.

In comparison to manual judging, which can be thorough but is time-intensive and subjective, the RAG-based method offers a more robust and impartial solution. Its architecture ensures an understanding of both abstract and practical arguments, making it well-suited for complex debate formats like Lincoln-Douglas. Moreover, the RAG framework allows for continuous updates to the knowledge base, enabling adaptability to various debate topics and ensuring the system remains relevant and effective over time.

The experiments aimed to evaluate the effectiveness of the Blitz Debate system in two dimensions: its accuracy in identifying debate winners and its ability to provide nuanced textual responses compared to ChatGPT.

The first experiment tested the RAG model's accuracy by using a dataset of Lincoln-Douglas debates. The system was evaluated on its ability to predict winners correctly, achieving a 90.5% accuracy rate, suggesting that it effectively leverages external knowledge and contextual cues. The results highlighted the robustness of the RAG architecture and its integration with high-quality data.

The second experiment compared textual responses from Blitz Debate and ChatGPT. Blitz Debate consistently outperformed ChatGPT by providing more detailed and evidence-based justifications. It referenced specific arguments from the transcripts and addressed counter arguments effectively, while ChatGPT often provided broader but less precise overviews.

These results demonstrate Blitz Debate's superior analytical depth and contextual awareness, attributable to its integration of external knowledge and advanced reasoning frameworks.

2. CHALLENGES

In order to build the project, a few challenges have been identified as follows.

2.1. Design A Machine Learning Framework

One of the primary challenges when it comes to designing a machine learning framework to make inferential decisions on debates is tackling the non-deterministic nature of providing results to users. Our previous implementation yielded inconsistent results of debates when fed the same transcripts as input, which questions the performance of the model to provide consistent feedback on complex data. We aim to design a structure for the prompt we feed through LangChain that suggests a more pragmatic decision-making process that will provide more evaluation metrics for the LLM to analyze when making a decision [6]. Concrete details which can be analyzed will make the performance of the chatbot more deterministic, and results will be more consistent with one another.

2.2. Time Complexity

Another significant challenge is considering the time complexity of such an application. Our input audio data is converted to textual data, and then afterwards, sent to a backend server via HTTP requests to analyze the structural contents [7]. It is then afterwardsreturned back to the frontend, where the user can interpret its results. Ways in which we can reduce the runtime of this application is by making use of a server and serverless architecture, where the audio gets transcribed to textual data via a serverless system. This should improve performance since the code will be running closer to the user, reducing latency. Our server also scales down overtime, which is another area we can seek to improve upon. Instead of booting up the server immediately

upon submission, we instead launch the application prior to submission, so that the user experiences less latency and a smoother experience on the application.

2.3. User Experience

A major component of the program is creating a customizable experience for the user. One challenge to consider is that the current system operates on a simple input/output structure, where users receive responses based on what they pass in. This structure may limit the program's ability to provide personalized feedback or adapt to varying user needs. To address this, we could implement a more dynamic framework that allows users to ask specific questions about their performance and receive targeted suggestions for improvement. For instance, users could ask about strategies to strengthen their rebuttals or enhance their argument structure, creating a more interactive and tailored experience. Additionally, we could enable the chatbot to parameterize different scoring metrics or adapt to various debate styles, ensuring flexibility across different debate formats. Another potential solution could involve offering example prompts that users can respond to and have analyzed by the system, helping them practice and refine their skills in a structured way.

3. SOLUTION

The main structure of the program is centered around a pipeline that links user input, semantic search and retrieval, and real-time feedback generation. These three components work together seamlessly to provide assessments of debate arguments.

The program begins with the user providing two audio prompts as input, representing the arguments to be evaluated. This input is processed by the system, initiating the framework's operations. The input data is received by the server, which activates the framework. The framework organizes the data flow and prepares the system for semantic search and retrieval. A semantic search is conducted within a vector database using a pre-trained BAAI embedding model. This search retrieves contextual data relevant to the debate topic, ensuring that the system can evaluate the arguments accurately. The knowledge database is specifically tailored towards debate formats, storing valuable information to enhance the system’s understanding of the prompts. The vector database then returns the retrieved contextual information, providing the AI with a comprehensive understanding of the debate’s context. The retrieved data is sent to the OpenAI API, generates responses, evaluates the arguments, and highlights strengths and areas for improvement. LangChain manages the system’s real-time feedback, providing the user with immediate responses.

We use OpenAI for natural language understanding and response generation; the BAAI embedding model for semantic search within a vector database; LangChain for orchestrating the feedback and ensuring seamless integration between components; and a cloud-based Flask-Render server to handle the framework and provide real-time processing and scalability.

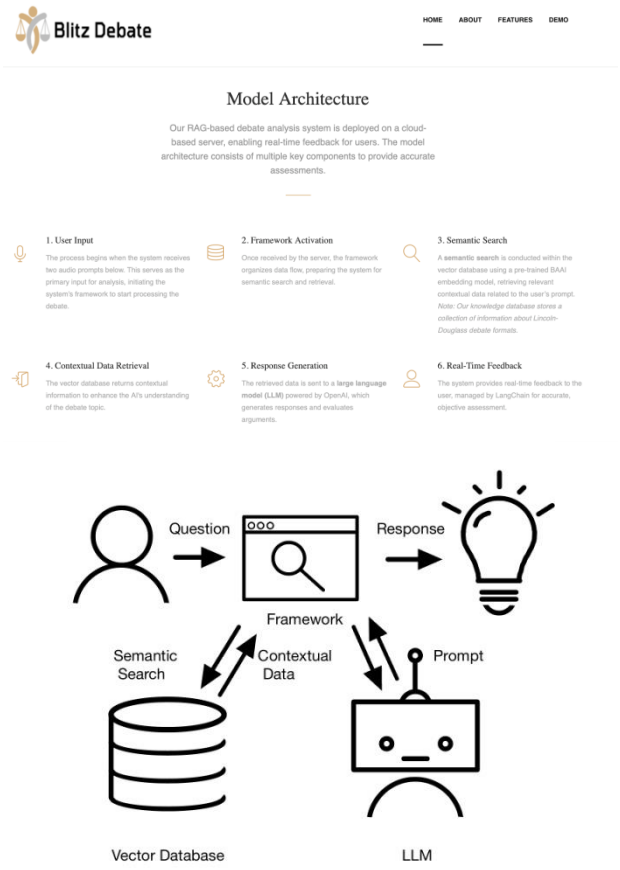


Figure 1. Overview of the solution

One of the primary components involved in our system is retrieving the necessary input data in order to begin inference. This requires the user(s) to record their audio through the website. The Lincoln-Douglas debate format sees an affirmative and negative side, where both sides argue their parts in a debate of their choosing [8]. The interface offers an easy experience for multiple users to upload their audio data. If the users would like a predetermined topic for discussion, the website has a feature which will select a random topic that the users can expand off of in their debate.

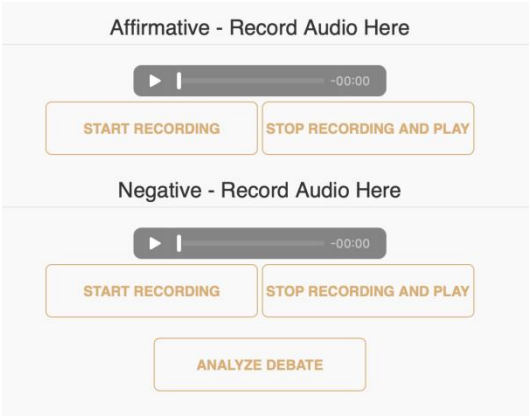


Figure 2. Screenshot of record audio function

```
<div class="row">
        <!-- Affirmative Recording Modal -->
        <div class="col-xs-12 col-sm-12 col-md-8 col-md-offset-2 text-center">
                <fieldset>
                        <legend>Affirmative - Record Audio Here</legend>
                        <audio id="affirmative-audio" controls></audio>
                        <button onclick="startRecording('affirmative')">Start Recording</button>
                        <button onclick="stopRecording('affirmative')">Stop Recording and Play</button>
                </fieldset>
        </div>

        <!-- Negative Recording Modal -->
        <div class="col-xs-12 col-sm-12 col-md-8 col-md-offset-2 text-center">
                <fieldset>
                        <legend>Negative - Record Audio Here</legend>
                        <audio id="negative-audio" controls></audio>
                        <button onclick="startRecording('negative')">Start Recording</button>
                        <button onclick="stopRecording('negative')">Stop Recording and Play</button>
                </fieldset>
        </div>

        <!-- Analyze Debate Button -->
        <div class="col-xs-12 col-sm-12 col-md-8 col-md-offset-2 text-center">
                <button onclick="analyzeDebate()">Analyze Debate using Generative AI</button>
        </div>

        <!-- Display Results -->
        <div class="col-xs-12 col-sm-12 col-md-8 col-md-offset-2 text-center">
                <div id="transcription-result"></div>
                <div id="gpt-response"></div>
        </div>
</div>
```

Figure 3. Screenshot of code 1

The audio tag is where the user will be recording their snippet for the debate. There are buttons to start and stop the recording, and there is an extra feature that allows for the user to listen back to the recording, in the event that they would like to re-record their code snippet. Once the affirmative and negative sides of the debate finish their recording, the users are able to submit their response by clicking on the “Analyze Debate using Generative AI” button, which will begin the inference by sending a http request to our servers [10]. The backend server is responsible for taking that audio data, converting to a textual format, and then sending this information to our RAG system so that we can extract important details about the structure of the affirmative and negative debate. Once this is completed, the information is then returned and sent back to the client, through the “gpt-result” tag.

The next component of our system makes use of our RAG system, which contains documents which encompasses a wide variety of documents about Lincoln-Douglas debates. In order for this to work, we set up an OpenAI RAG Model using LangChain and uploaded several different documents, and retrieved the vectorstore for later use. This vectorstore is saved internally in our backend server, so that we can ask a range of questions regarding the stored documents and the audio data that the user sends. This approach is more effective in providing more sophisticated results, as the AI will have the necessary context in order to make sound conclusions regarding different structures in debates, such as, the clarity of both sides, the effectiveness of their arguments, and more.

```
logger.info('Searching FAISS index.')
D, I = vectorstore.index.search(query_embedding_np, k=5)
logger.info('FAISS index search complete.')

# Retrieve documents based on the FAISS index results as Document objects
retrieved_docs = []
for i in I[0]:
    doc_id = vectorstore.index_to_docstore_id[i]
    logger.info(f'Retrieving document with id {doc_id}')
    doc = vectorstore.docstore.search(doc_id)
    if isinstance(doc, Document):
        retrieved_docs.append(doc)
        logger.info(f'Document {doc_id} retrieved.')
    else:
        logger.warning(f"Unexpected object type: {type(doc)} for doc_id: {doc_id}")

# Using the retrieved documents as the context
context = "\n\n".join([doc.page_content for doc in retrieved_docs])
logger.info('Context for LLM created.')

prompt = ChatPromptTemplate.from_template(template)
doc_chain = create_stuff_documents_chain(llm, prompt)
logger.info('LLM chain created.')

# Invoke the chain with the transcription and context
logger.info('Invoking LLM chain.')
response = doc_chain.invoke({'context': context, 'input': combined_transcription})
logger.info('LLM chain invocation complete.')
```

Figure 4. Screenshot of code 2

We can see in the above code snippet the variable named “vectorstore,” which serves as our vector database where all of the textual embeddings are located. We use this to then create a document chain, which is dependent on the audio data that the user sends to the program. Once this is performed, we then create a document chain, where relevant information is retrieved from the collection of documents in our vector store before being used to generate a response. In this example, the response will be dependent on the prompt we feed to the AI. We ask several questions, one of which regards an evaluation of the performance of both sides of the debate. This is then returned once we invoke the chain with the transcription and context, and we can then return back to the user. This approach can more effectively enhance the context and accuracy of the generated text based on the retrieved document data, since our AI is catered towards analyzing the results of Lincoln-Douglas debates.

The final step in this approach involves finishing the http request and sending the information back to the client so that the user can interpret the results inferred by the RAG system. The information is presented to the client in json format, and the javascript converts the results into a more interpretable, well-formatted response for the user.

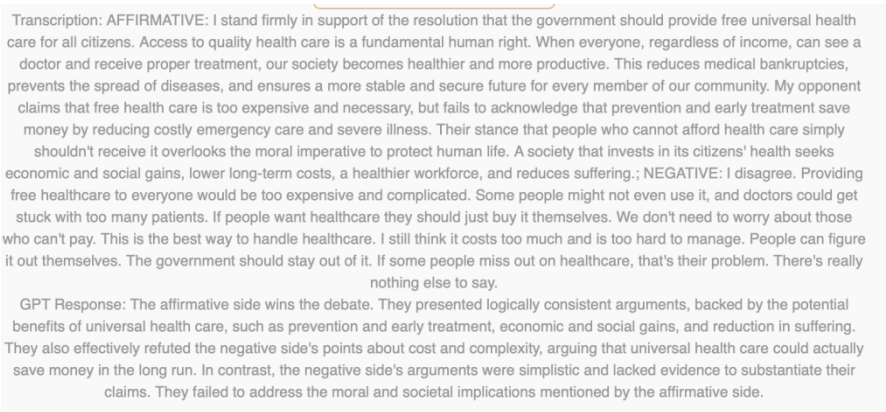


Figure 5. Screenshot of transctiption

```
    .then(response => {
        console.log('Response status:', response.status, response.statusText);
        if (!response.ok) {
            throw new Error('Network response was not ok');
        }
        return response.json();
    })
    .then(data => {
        console.log('Server response received:', data);
        // Display the server response
        document.getElementById('transcription-result').innerText = 'Transcription: ' + data.transcription;
        document.getElementById('gpt-response').innerText = 'GPT Response: ' + data.winner;
    })
    .catch(error => {
        console.error('Error in fetch:', error);
        alert('An error occurred while analyzing the debate. Check the console for details.');
    });
}).catch(function(error) {
    console.error('Error exporting audio:', error);
    alert('An error occurred while exporting the recordings. Check the console for details.');
});
```

Figure 6. Screenshot of code 3

For the following example, the resolution was as follows: “The government should provide free, universal healthcare for all citizens.” The following transcript was used:

Affirmative Constructive:

“Good evening. I stand firmly in support of the resolution that the government should provide free, universal healthcare for all citizens. Access to quality healthcare is a fundamental human right. When everyone, regardless of income, can see a doctor and receive proper treatment, our society becomes healthier and more productive. This reduces medical bankruptcies, prevents the spread of disease, and ensures a more stable and secure future for every member of our community. Ultimately, a universal healthcare system is both morally justified and practically beneficial.”

Negative Constructive:

“I disagree. Providing free healthcare to everyone would be too expensive and complicated. Some people might not even use it, and doctors could get stuck with too many patients. If people want healthcare, they should just buy it themselves. We don’t need to worry about those who can’t pay. This is the best way to handle healthcare.”

Affirmative Rebuttal:

“My opponent claims free healthcare is too expensive and unnecessary, but fails to acknowledge that prevention and early treatment save money by reducing costly emergency care and severe illness. Their stance that people who cannot afford healthcare simply shouldn’t receive it overlooks the moral imperative to protect human life. A society that invests in its citizens’ health sees economic and social gains—lower long-term costs, a healthier workforce, and reduced suffering. My opponent’s argument ignores these undeniable benefits and instead relies on the flawed notion that healthcare should only belong to those who can afford it. That stance is morally unsound and economically short-sighted.”

Negative Rebuttal:

“I still think it costs too much and is too hard to manage. People can figure it out themselves. The government should stay out of it. If some people miss out on healthcare, that’s their problem. There’s really nothing else to say.”

As we can see in the above example, the audio data is successfully transcribed into textual data, and the response is then returned back to the user. In this example, the affirmative side poses a significantly stronger argument compared to the negative, which looks to be the RAG model's interpretation in the below response [9]:

"The affirmative side wins the debate. They presented logically consistent arguments, backed by the potential benefits of universal health care, such as prevention and early treatment, economic and social gains, and reduction in suffering. They also effectively refuted the negative side's points about cost and complexity, arguing that universal health care could actually save money in the long run. In contrast, the negative side's arguments were simplistic and lacked evidence to substantiate their claims. They failed to address the moral and societal implications mentioned by the affirmative side."

4. EXPERIMENT

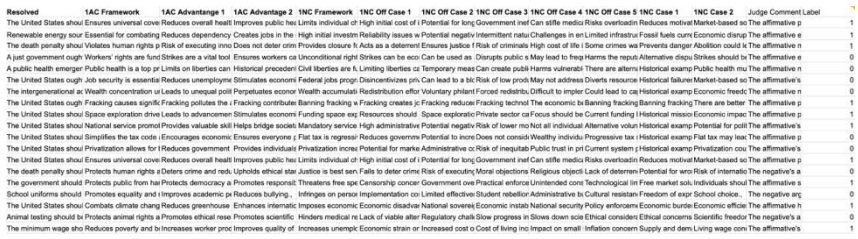
4.1. Experiment 1

One thing we aim to consider with Blitz Debate is the accuracy of the RAG model's responses. This is crucial in determining whether or not the model is effective in its analysis.

We create a dataset of Lincoln-Douglas debates, where there are several columns: the affirmative and negative transcripts, and the winner of the debate. We will perform a simple binary classification test, where we evaluate the model based on the accuracy, defined as the proportion of correct guesses out of the total number of debates. Accuracy is the most intuitive metric to consider, especially since our main goal is to determine how often the RAG model's inference about the winner matches the ground truth label (the known winner). Accuracy can directly give us a clear, simple measure of the model's correctness. The control data was collected from openCaselist, a platform which provides debate transcripts from previous high-school and college competitions, as administered from the National Speech & Debate Association (NSDA). The data is structured as follows:

Resolved: The topic for the debate
AC: The affirmative case
NC: The negative case
Judge Comments: The notes written from judges observing and rating the performance of the debate

Label: The winner of the debate, as instructed from the 'Judge Comments' column. A label of '1' indicates the affirmative won the debate, whereas a label of '0' means the negative won.



Resolved	1AC Framework	1AC Advantange 1	1AC Advantage 2	1NC Framework	1NC Off Case 1	1NC Off Case 2	1NC Off Case 3	1NC Off Case 4	1NC Off Case 5	1NC Case 1	1NC Case 2	Judge Comment	Label
The United States shoul	Ensures universal cove	Reduces overall healt	Improves public hea	Limits individual ch	High initial cost of i	Potential for long	Government inef	Can stifle medica	Risks overloadin	Reduces motivat	Market-based so	The affirmative p	1
Renewable energy sour	Essential for combating	Reduces dependency	Creates jobs in the	High initial investm	Reliability issues w	Potential negativ	Intermittent natu	Challenges in en	Limited infrastruc	Fossil fuels curre	Economic disrup	The affirmative e	1
The death penalty shoul	Violates human rights p	Risk of executing inno	Does not deter crim	Provides closure fa	Acts as a deterrent	Ensures justice f	Risk of criminals	High cost of life i	Some crimes wa	Prevents danger	Abolition could le	The affirmative n	1
A just government ough	Workers' rights are fund	Strikes are a vital tool	Ensures workers ca	Unconditional right	Strikes can be eco	Can be used as	Disrupts public s	May lead to freq	Harms the reputa	Alternative disp	Strikes should be	The affirmative e	0
A public health emerger	Public health is a top pr	Limits on liberties can	Historical precedent	Civil liberties are fu	Limiting liberties ca	Temporary meas	Can create publi	Harms vulnerabl	There are alterna	Historical examp	Public health mu	The affirmative n	0
The United States ough	Job security is essentia	Reduces unemployme	Stimulates economi	Federal jobs progr	Disincentivizes priv	Can lead to a blc	Risk of low prod	May not address	Diverts resource	Historical failure	Market-based so	The affirmative's	0
The intergenerational ac	Wealth concentration ur	Leads to unequal polit	Perpetuates econor	Wealth accumulati	Redistribution effor	Voluntary philant	Forced redistribu	Difficult to imple	Could lead to cap	Historical examp	Economic freedc	The affirmative n	0
The United States ough	Fracking causes signific	Fracking pollutes the a	Fracking contribute	Banning fracking w	Fracking creates jc	Fracking reduces	Fracking technol	The economic be	Banning fracking	Banning fracking	There are better	The affirmative p	1
The United States shoul	Space exploration drive	Leads to advancemen	Stimulates economi	Funding space exp	Resources should	Space exploratio	Private sector ca	Focus should be	Current funding l	Historical missio	Economic impac	The affirmative p	1
The United States shoul	National service promot	Provides valuable skill	Helps bridge societi	Mandatory service	High administrative	Potential negativ	Risk of lower mo	Not all individuals	Alternative volun	Historical examp	Potential for poli	The affirmative's	1
The United States shoul	Simplifies the tax code	Encourages economic	Ensures everyone p	Flat tax is regressiv	Reduces governme	Potential to incre	Does not consid	Wealthy individu	Progressive tax s	Historical examp	Flat tax may lead	The affirmative p	0
The United States shoul	Privatization allows for t	Reduces government	Provides individuals	Privatization increa	Potential for marke	Administrative co	Risk of inequitab	Public trust in pri	Current system p	Historical examp	Privatization cou	The affirmative's	0
The United States shoul	Ensures universal cove	Reduces overall healt	Improves public hea	Limits individual ch	High initial cost of i	Potential for long	Government inef	Can stifle medica	Risks overloadin	Reduces motivat	Market-based so	The affirmative p	1
The death penalty shoul	Protects human rights a	Deters crime and redu	Upholds ethical star	Justice is best serv	Fails to deter crime	Risk of executing	Moral objections	Religious objecti	Lack of deterren	Potential for wro	Risk of internatio	The negative's a	0
The government should	Protects public from hai	Protects democracy a	Promotes responsit	Threatens free spe	Censorship concer	Government ove	Practical enforce	Unintended cons	Technological lin	Free market solu	Individuals shoul	The affirmative s	1
School uniforms should	Promotes equality and i	Improves academic pe	Reduces bullying,	Infringes on perso	Implementation co	Limited effective	Student rebellion	Administrative bu	Cultural resistan	Freedom of expr	School choice,	The negative arg	0
The United States shoul	Combats climate chang	Reduces greenhouse	Enhances internatic	Imposes economic	Economic disadvar	National sovereig	Economic instab	National security	Policy enforcem	Economic burde	Economic efficie	The affirmative h	1
Animal testing should be	Protects animal rights a	Promotes ethical rese	Promotes scientific	Hinders medical re	Lack of viable alter	Regulatory challe	Slow progress in	Slows down scie	Ethical considera	Ethical concerns	Scientific freedor	The negative's a	0
The minimum wage sho	Reduces poverty and bu	Increases worker proc	Improves quality of	Increases unemplc	Economic strain or	Increased cost o	Cost of living inc	Impact on small	Inflation concern	Supply and dem	Living wage con	The affirmative's	1

Figure 7. Figure of experiment 1

Ground Truth:

1-1-1-0-0-0-0-1-1-1-0-0-1-0-1-0-1-0-1-0-1-0-1-0-1-0-1-0-1-0-1-0-1-0-1-0-1-0-1-0-1-0-1-0

Prediction:
1-1-1-0-0-0-0-0-1-1-1-0-1-0-0-0-1-0-1-0-1-0-1-0-1-0-1-1-1-0-1-0-1-0-1-0-1-0-1-0-1-0

Accuracy = (Total Correct Predictions)/(Total Predictions) = 38/42 ≈ 90.5%

We converted each debate's prediction into a binary correctness value, where "1" indicates a correct guess (matching the ground truth winner) and "0" indicates an incorrect guess. With 38 correct predictions out of 42, the mean value of correctness (i.e., average) is 38/42 ≈ 0.905. Because the majority of predictions are correct, the median correctness value, when ordering all correctness points from lowest to highest, is also 1. The lowest correctness value is 0 (representing the few instances where the model chose the wrong winner), and the highest correctness value is 1 (all correct predictions).

What's somewhat surprising is the relatively high accuracy—over 90%—which may exceed initial expectations given the complexity of debates and subtlety of argument quality. This could indicate that the model is leveraging contextual cues effectively. The biggest factor influencing these results may be the robustness of the underlying retrieval-augmented generation approach, combined with well-structured debate data, allowing the model to accurately interpret and reflect the ground truth labels.

4.2. Experiment 2

A potential blind spot is the model's effectiveness in analysis. Ensuring proper presentation of information is necessary to ensure trustworthiness of the model's assessments and increases confidence in the final overall outcomes.

We will present identical debate transcripts and known expert judgments of winners to both vanilla ChatGPT and our improved model. Instead of focusing solely on whether each model names the correct winner, we will compare how each model interprets and justifies that outcome. This involves analyzing the specificity, logical coherence, and evidence-citation in their textual responses. The experiment is structured this way to isolate differences in interpretative quality, demonstrating whether our solution offers more informed and meaningful reasoning than vanilla ChatGPT. An example below can be seen from a previous resolution observed from the JW Patterson Tournament of Champions. The transcripts themselves were created by AI for interpretability, and the results were compared between the two solutions. An example can be seen in the following data:

Resolution	Affirmative Transcript	Negative Transcript	ChatGPT Response	Blitz Debate Response
The United States should adopt a universal basic income.	I affirm the resolution. A universal basic income ensures financial security for all citizens, reducing poverty, and stabilizing the economy. By guaranteeing a baseline income, we provide a safety net that allows people to pursue education, entrepreneurship, and personal growth. In the long run, UBI stimulates economic activity and leads to a more equitable society. My opponent ignores that many people forced into low-wage work would finally have a chance to improve their skills. Data from pilot programs show that recipients use UBI to seek better opportunities, not to remain idle. The cost is	I negate. A UBI is too expensive and discourages work. By handing out money to everyone we remove the incentive to find jobs and contribute productively instead of strengthening the economy. UBI leads to dependency, laziness, and overextended government budgets. Even if some people do better, it's not worth the huge price tag. The government shouldn't hand out money to everyone without requiring work. There will always be those who take advantage and rely on free income, draining resources that could be spent elsewhere.	The Affirmative provided a clear vision of social and economic improvement through UBI and addressed cost and incentive concerns. The Negative relied heavily on fears without substantial evidence. The Affirmative is more persuasive.	The affirmative side presented stronger arguments, utilizing evidence from pilot programs to support their claims about the impact of a universal basic income (UBI). They logically argued that UBI results in reduced poverty, increased personal growth, and economic stability. The Negative side's argument lacked such evidence and was mainly based on unproven assumptions about laziness and dependency. While they raised legitimate concerns about cost and potential abuse, they failed to effectively counter the affirmative points about reduced welfare bureaucracy and increased productivity. Therefore, the Affirmative side won the debate.
	offset by reduced welfare bureaucracy and increased productivity. UBI empowers individuals rather than crippling them.			
Just governments ought to prioritize civil liberties over national security measures.	I affirm, civil liberties are the cornerstone of a just society. When governments erode personal freedoms for the sake of security, they compromise moral principles and risk authoritarianism. Upholding civil liberties ensures trust, preserves democratic values, and respects human dignity. My opponent assumes that giving up rights leads to safety, but oppressive surveillance and unchecked security powers often target innocent citizens and erode trust. A balance is possible, but civil liberties must be the priority or we lose the very	I negate. Without strong national security measures, a society cannot survive long enough to enjoy its liberties. Surveillance, preventative detentions, and other security tools protect people from threats. Without these, chaos ensues and freedoms vanish under terrorist attacks or foreign aggression. It's better to be safe than sorry. If a few people lose some privacy, that's an acceptable tradeoff to prevent disaster. Civil liberties are worthless if terrorists destroy the		Both sides presented reasonable arguments, but the affirmative side won this debate. The affirmative side's argument was logically consistent and emphasized the moral and democratic importance of civil liberties. They also effectively critiqued the negative side's argument by pointing out the potential for misuse of surveillance and other security measures. The negative side, while presenting a strong argument emphasizing the importance of security, failed to adequately address the concerns raised by the affirmative side about potential authoritarianism
	freedom we speak to protect.	society we live in.		and erosion of trust. They also relied on a fear-based argument ("better safe than sorry") without providing substantial evidence to support the claim that sacrificing privacy prevents disaster.

Figure 8. Figure of experiment 2

We will collect pairs of textual responses—one from vanilla ChatGPT and one from our enhanced model—based on the same debate material. Each response is examined along several qualitative dimensions: Does the model explain the winner's reasoning effectively? Does it reference particular arguments made by the affirmative or negative sides, and does it address counterarguments? Are the explanations logically consistent, well-supported by the given transcripts, and aligned with expert commentary? As seen from the data, we can see the Blitz Debate provides a more thorough approach in providing detailed analysis as to why one side won over the other.

Comparing the textual responses from ChatGPT to those of Blitz Debate reveals notable differences in detail, evidence usage, and nuance. ChatGPT's summaries tend to focus on broad thematic elements and overall persuasiveness, stating that the affirmative side "provided a clear vision" or "emphasized the moral and democratic importance" of certain principles. While this framing is accurate, it often lacks direct references to specific pieces of evidence or particular arguments made within the transcript. ChatGPT's reasoning can feel more like a general overview rather than an in-depth, argument-by-argument analysis.

Blitz Debate's textual response tends to incorporate more explicit reference to the debate content. For example, it mentions "evidence from pilot programs" in the UBI debate and "fears without substantial evidence" in the negative's case, or directly addresses the affirmative critique of "unchecked security powers" in the second debate.

5. RELATED WORK

Slonim et al. (2021) present Project Debater, an AI system designed to engage in competitive, human-like debates [11]. It works by retrieving relevant evidence, organizing arguments into coherent narratives, and delivering structured speeches. While highly innovative, the system's effectiveness is constrained by its reliance on pre-processed data and limited real-time reasoning. It struggles with creative rebuttals and deep contextual understanding, often ignoring subtle emotional or cultural nuances. Compared to their approach, our project provides a more adaptive reasoning mechanism, integrates contextual sentiment analysis, and refines evidence selection processes. This enables our system to produce more contextually aware, flexible, and human-like debate analyses.

Rinott et al. (2015) introduce a system for automatically detecting textual evidence that supports a given claim [12]. Their approach relies on supervised learning models trained on a specifically crafted benchmark dataset. The model identifies sentences likely to contain evidence and matches them to a claim, offering a structured pipeline for evidence retrieval. Although the solution shows promising experimental results, it is limited by the quality and scope of its training data and its ability to handle nuanced, domain-specific claims. It does not consider deeper logical structures, emotional appeals, or cross-document connections. Our project refines these aspects by integrating richer context modeling, domain adaptation techniques, and more advanced argumentation structures.

Bar-Haim et al. (2021) provide an in-depth tutorial on developing AI systems capable of debating humans [13]. Their approach integrates argument mining, stance classification, argument quality assessment, and narrative generation into a cohesive pipeline. While effective at structuring arguments and generating coherent rebuttals, the system faces challenges in real-time comprehension and rebuttal of nuanced human speech, especially when arguments are implicit. Additionally, the system is constrained by its reliance on pre-structured corpora and predefined taxonomies for argument retrieval. Our project improves on these limitations by introducing

adaptive contextual reasoning and dynamic argument generation, allowing for more flexible and nuanced debate interpretations.

6. CONCLUSIONS

One limitation of Blitz Debate is its reliance on the quality and breadth of external knowledge sources. While the RAG framework effectively retrieves contextual data, its performance can degrade if the vector database lacks diverse or high-quality information relevant to a specific debate topic. Additionally, while the system excels at structured and logical arguments, it struggles with deeply nuanced ethical or philosophical reasoning that requires subjective interpretation or cultural awareness [14]. Another challenge is the real-time feedback mechanism, which occasionally sacrifices depth for immediacy, leading to overly generalized feedback.

To address these issues, we would expand the knowledge base by integrating domain-specific datasets and incorporating user-feedback loops to refine data quality. Enhancing the model with fine-tuning on ethically complex debates would help to improve the RAG system's grasp of subjective arguments [15]. Finally, implementing a tiered feedback system—providing both immediate and in-depth analyses—would balance real-time responsiveness with detailed, actionable insights for users.

Blitz Debate demonstrates the potential of Retrieval-Augmented Generation models to enhance debate evaluation with contextual accuracy and nuanced feedback. By addressing current limitations and integrating improvements, this system can significantly advance how debates are analyzed and understood, fostering critical thinking, fairness, and accessibility for students and participants across diverse contexts.

REFERENCES

[1] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.

[2] Kasneci, Enkelejda, et al. "ChatGPT for good? On opportunities and challenges of large language models for education." Learning and individual differences 103 (2023): 102274.

[3] Bosch, Jan, Helena Holmström Olsson, and Ivica Crnkovic. "Engineering ai systems: A research agenda." Artificial intelligence paradigms for smart cyber-physical systems (2021): 1-19.

[4] Min, Bonan, et al. "Recent advances in natural language processing via large pre-trained language models: A survey." ACM Computing Surveys 56.2 (2023): 1-40.

[5] Roumeliotis, Konstantinos I., and Nikolaos D. Tselikas. "Chatgpt and open-ai models: A preliminary review." Future Internet 15.6 (2023): 192.

[6] Asyrofi, Rakha, et al. "Systematic Literature Review Langchain Proposed." 2023 International Electronics Symposium (IES). IEEE, 2023.

[7] Schechter, Stuart, Murali Krishnan, and Michael D. Smith. "Using path profiles to predict HTTP requests." Computer Networks and ISDN Systems 30.1-7 (1998): 457-467.

[8] Lincoln, Abraham, and Stephen A. Douglas. The complete Lincoln-Douglas debates of 1858. University of Chicago Press, 1991.

[9] Siriwardhana, Shamane, et al. "Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering." Transactions of the Association for Computational Linguistics 11 (2023): 1-17.

[10] Feuerriegel, Stefan, et al. "Generative ai." Business & Information Systems Engineering 66.1 (2024): 111-126.

[11] Liu, Ou Lydia, Lois Frankel, and Katrina Crotts Roohr. "Assessing critical thinking in higher education: Current state and directions for next-generation assessment." ETS Research Report Series 2014.1 (2014): 1-23.

[12] O'Neill, James Milton. A Manual of Debate and Oral Discussion for Schools, Societies and Clubs. Century Company, 1920.

[13] Stubbs, Charles, and Tommy Yuan. "Towards an Argument Knowledge Base for Autonomous Debating Systems." CMNA@ COMMA (2022): 51-60.

[14] Rinott, Ruty, et al. "Show me your evidence-an automatic method for context dependent evidence detection." Proceedings of the 2015 conference on empirical methods in natural language processing. 2015.

[15] Bar-Haim, Roy, et al. "Advances in debating technologies: Building AI that can debate humans." Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts. 2021.

© 2025 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.