LLMs as Debate Partners: Utilizing Genetic Algorithms and Adversarial Search for Adaptive Arguments

Prakash Aryan
Department of Computer Science
Birla Institute of Technology and Science, Pilani - Dubai Campus
Dubai, UAE
h20230010@dubai.bits-pilani.ac.in
ORCID: 0009-0003-9221-1453

2024 IEEE Conference on Engineering Informatics (ICEI) | 979-8-3315-0577-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICEI64305.2024.10912343

***Abstract*—This paper introduces DebateBrawl, an innovative AI-powered debate platform that integrates Large Language Models (LLMs), Genetic Algorithms (GA), and Adversarial Search (AS) to create an adaptive and engaging debating experience. DebateBrawl addresses the limitations of traditional LLMs in strategic planning by incorporating evolutionary optimization and game-theoretic techniques. The system demonstrates remarkable performance in generating coherent, contextually relevant arguments while adapting its strategy in real-time. Experimental results involving 23 debates show balanced outcomes between AI and human participants, with the AI system achieving an average score of 2.72 compared to the human average of 2.67 out of 10. User feedback indicates significant improvements in debating skills and a highly satisfactory learning experience, with 85% of users reporting improved debating abilities and 78% finding the AI opponent appropriately challenging. The system's ability to maintain high factual accuracy (92% compared to 78% in human-only debates) while generating diverse arguments addresses critical concerns in AI-assisted discourse. DebateBrawl not only serves as an effective educational tool but also contributes to the broader goal of improving public discourse through AI-assisted argumentation. The paper discusses the ethical implications of AI in persuasive contexts and outlines the measures implemented to ensure responsible development and deployment of the system, including robust fact-checking mechanisms and transparency in decision-making processes.**

***Index Terms*—Machine Learning, Deep Learning, Generative AI, Large Language Models, Genetic Algorithms, Adversarial Search**

I. Introduction

The convergence of artificial intelligence (AI) and argumentation has emerged as a new platform, promising to reform debates, improve critical thinking, and foster more informed discourse. As language models become increasingly sophisticated, their potential as intelligent debate partners and argumentation assistants has captured the imagination of researchers and educators. This paper introduces DebateBrawl, a novel system that utilizes Large Language Models (LLMs), Genetic Algorithms (GA), and Adversarial Search (AS) to create an adaptive and engaging debate platform.

However, current AI-powered debate systems face several critical challenges that limit their effectiveness as educational and practice tools. First, while LLMs excel at generating fluent responses, they lack strategic depth in extended debates, often failing to maintain consistent argumentation strategies across multiple exchanges. Second, existing systems typically operate with fixed response patterns, unable to adapt to different debate styles or learn from past interactions. Third, most platforms lack sophisticated planning capabilities needed to anticipate and effectively counter opponent arguments. These limitations result in debate experiences that, while technologically advanced, fail to provide the dynamic, adaptive, and educational interaction necessary for meaningful debate practice. DebateBrawl addresses these fundamental challenges through a novel integration of three complementary technologies: LLMs for natural language understanding and generation, Genetic Algorithms for strategic evolution and adaptation, and Adversarial Search for predictive planning and counter-argument generation.

The development of transformer-based language models, based on GPT architectures, has marked a paradigm shift in natural language processing [1]. These models, trained on vast corpora of text, have demonstrated remarkable capabilities in generating coherent, context-aware text across diverse domains. However, while LLMs excel at generating fluent and contextually relevant responses, they often lack the strategic depth and adaptability required for nuanced, multi-turn debates. This limitation stems from their fundamentally reactive nature, where responses are generated based on immediate context rather than long-term strategic planning [2].

DebateBrawl addresses this challenge by incorporating genetic algorithms and adversarial search techniques. Genetic algorithms, inspired by natural selection and evolution, have proven effective in optimizing complex, multi-dimensional problems [3]. In the context of debate, GAs evolve and refine argumentation strategies over time, adapting to the specific topic, opponent, and flow of the debate. This evolutionary approach allows the system to discover and hone effective combinations of rhetorical devices, logical structures, and persuasive techniques.

Adversarial search provides a framework for anticipating and planning responses to potential counterarguments [4]. By simulating possible debate trajectories and evaluating their outcomes, AS enables the system to make more informed decisions about argument selection and presentation. In DebateBrawl, this is made possible through a debate move predictor that anticipates opponent strategies and suggests effective counter-moves. The synergy between GAs and AS creates a debate engine that can not only generate coherent arguments but also strategically plan its approach to maximize persuasiveness and effectiveness.

The integration of LLMs, GAs, and AS in DebateBrawl represents a significant advancement in AI-assisted argumentation systems. Previous work in this field has primarily focused on argument mining [5], stance detection [6], and automated fact-checking [7]. While these approaches have made valuable contributions to computational argumentation, they often operate on a more granular level, focusing on individual arguments or claims rather than the holistic process of debate. DebateBrawl builds upon these foundations but takes a more comprehensive approach, addressing the full lifecycle of a debate from argument generation to strategic planning and adaptive response.

The development of DebateBrawl is motivated by the growing recognition of the importance of critical thinking and argumentation skills in education and public discourse. In an era characterized by information overload and rapid spread of misinformation, the ability to construct, analyze, and evaluate arguments is more crucial than ever [8]. Traditional debate training methods, while valuable, are often limited by resource constraints and the availability of skilled human opponents and coaches.

II. Related Works

The integration of Large Language Models (LLMs) with evolutionary algorithms and other optimization techniques has proved to be a promising area of research, offering new possibilities for improving AI systems' capabilities across various domains. This section explores the diverse applications and methodologies that combine LLMs with evolutionary computation, genetic algorithms, and other optimization strategies.

A. Evolutionary Algorithms and LLMs

The synergy between evolutionary algorithms (EAs) and LLMs has been explored in several studies, showcasing the potential for improved optimization and problem-solving capabilities. Liu et al. [9] introduced LLM-driven Evolutionary Algorithms (LMEA), a novel approach that uses LLMs as evolutionary combinatorial optimizers. Their work demonstrates that LLMs can be effectively used to select parent solutions, perform crossover and mutation operations, and generate offspring solutions with minimal domain knowledge and human intervention. The authors applied LMEA to classical traveling salesman problems (TSPs), showing competitive performance compared to traditional heuristics for instances with up to 20 nodes.

In a related study, Chao et al. [10] explored the parallels between LLMs and EAs, identifying common characteristics such as token representation and individual representation, position encoding and fitness shaping, and model training and parameter adaptation. The authors analyzed existing interdisciplinary research, focusing on evolutionary fine-tuning and LLM-enhanced EAs.

B. LLMs in Game Design and Creative Tasks

The application of LLMs in creative tasks, such as game design, has also been explored. Lanzi and Loiacono [11] presented a collaborative game design framework that combines interactive evolution and LLMs to simulate the human design process. Their approach uses an interactive genetic algorithm to exploit user feedback for selecting promising ideas, while LLMs are employed for the complex creative task of recombining and varying ideas. This framework demonstrates the potential of LLMs in augmenting human creativity and facilitating collaborative design processes.

C. LLMs in Decision-Making and Planning

Several studies have investigated the use of LLMs in decision-making and planning tasks. Zhou et al. [12] introduced Language Agent Tree Search (LATS), a framework that integrates Monte Carlo Tree Search with LLMs to enable more effective reasoning, acting, and planning. LATS uses the in-context learning ability of LLMs and incorporates LM-powered value functions and self-reflections for proficient exploration and improved decision-making. The framework demonstrated state-of-the-art performance in various domains, including programming, interactive question-answering, web navigation, and math problems.

Similarly, Wan et al. [13] proposed an AlphaZero-like tree-search learning framework for LLMs (TS-LLM), which uses a learned value function to guide LLM decoding. Their approach is adaptable to a wide range of tasks, language models of various sizes, and tasks with varying search depths. TS-LLM showed improved performance in reasoning, planning, alignment, and decision-making tasks, demonstrating the potential of combining tree search algorithms with LLMs for improved problem-solving capabilities.

D. LLMs in Recommender Systems

The integration of LLMs into recommender systems has been explored to improve user interaction and personalization. Friedman et al. [14] proposed a roadmap for building large-scale conversational recommender systems using LLMs. Their work addresses challenges in understanding user preferences and dialogue management by introducing RecLLM, a YouTube video-based conversational recommender system that facilitates natural conversations and personalized recommendations.

E. LLMs in Neural Architecture Search and Model Optimization

The application of LLMs and evolutionary techniques in optimizing neural network architectures has been an area of active research. Sarah et al. [15] proposed an effective method for finding Pareto-optimal network architectures based on LLaMA2-7B using one-shot Neural Architecture Search (NAS). Their approach combines fine-tuning with genetic algorithm-based search to identify smaller, less computationally complex network architectures. The study demonstrated significant reductions in model size and improvements in throughput for certain tasks, with minimal accuracy loss.

Zhong et al. [16] introduced a novel LLM-assisted optimizer (LLMO) for addressing adversarial robustness in neural architecture search (ARNAS). Their approach uses the Gemini LLM to generate solutions for ARNAS instances, demonstrating competitive performance compared to well-known metaheuristic algorithms. This research highlights the potential of LLMs as effective combinatorial optimizers in the context of neural architecture design and optimization.

F. LLMs in Molecular Design and Materials Science

The application of generative models, including LLMs, in drug discovery and materials science has shown promise in overcoming limitations of traditional inverse design methods. Bhowmik et al. [17] examined the effectiveness of generative models in creating virtual libraries of molecules and facilitating drug discovery. The authors proposed a hybrid architecture combining masked language models with generative adversarial networks (GANs) to efficiently generate new molecules.

G. Explainable AI and LLMs in Genetic Programming

The integration of explainable AI (XAI) techniques with genetic programming and LLMs has been explored to improve the interpretability of complex algorithms. Maddigan et al. [18] introduced GP4NLDR, an XAI dashboard that combines genetic programming with an LLM-powered chatbot to provide comprehensive, user-centered explanations for nonlinear dimensionality reduction. Their study demonstrates the potential of using LLMs to generate intuitive and insightful narratives about high-dimensional data reduction processes, while also addressing important considerations such as data privacy and the challenges of hallucinatory outputs from LLMs.

H. LLMs in Artificial Evolutionary Intelligence

The concept of Artificial Evolutionary Intelligence (AEI), which combines evolutionary computation with artificial general intelligence, has been proposed as a promising direction for future research. He et al. [19] discussed a paradigm of LLMs for evolutionary computation, addressing three main issues: multi-modal representation capability, general models for versatile learning, and the ability to understand evolutionary computation concepts and behaviors.

I. Challenges and Ethical Considerations

While the integration of LLMs with evolutionary and optimization techniques shows great promise, it also raises important challenges and ethical considerations. Gaudi [20] conducted a comprehensive survey on adversarial aspects in LLMs, discussing issues such as harmful generation, fairness, privacy, and robustness. The study highlights the need for adversarial training techniques, fine-tuning methods, and mitigation strategies to address these challenges.

Zhang et al. [21] provided a comprehensive study on knowledge editing for LLMs, proposing a unified categorization criterion for knowledge editing methods. Their work introduces a new benchmark, KnowEdit, for evaluating knowledge editing approaches and discusses potential applications and implications of this technology. This research underscores the importance of developing methods to efficiently modify LLMs' behaviors within specific domains while preserving overall performance across various inputs.

The integration of LLMs with evolutionary algorithms and optimization techniques represents a rapidly evolving and promising field of research. From healthcare applications to game design, from molecular modeling to neural architecture search, LLMs are being combined with various computational techniques to improve problem-solving capabilities, improve decision-making processes, and generate novel solutions across diverse domains. As this field continues to develop, addressing ethical considerations, improving explainability, and optimizing performance on resource-constrained hardware will be crucial areas of focus for future research.

III. Methodology

The DebateBrawl system represents an innovative approach to AI-powered debate platforms, integrating advanced natural language processing techniques with adaptive learning algorithms. This section provides a detailed overview of the system architecture, key components, and methodologies employed in the development and implementation of DebateBrawl.

A. System Architecture

DebateBrawl employs a client-server architecture, designed for modularity, scalability, and efficiency. Figure 1 illustrates the overall system architecture.

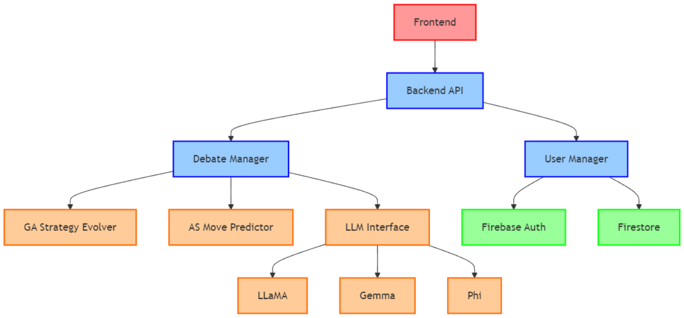


Fig. 1. Overall System Architecture of DebateBrawl

The system is divided into frontend and backend components, connected through well-defined APIs. This separation allows for independent development and scaling of each component. The architecture is designed to handle multiple concurrent debates while maintaining low latency and high throughput, crucial for a responsive and engaging user experience.

1) Frontend Architecture: The frontend of DebateBrawl is built using Next.js, a React-based framework that provides server-side rendering capabilities and optimized performance. Figure 2 details the frontend architecture.

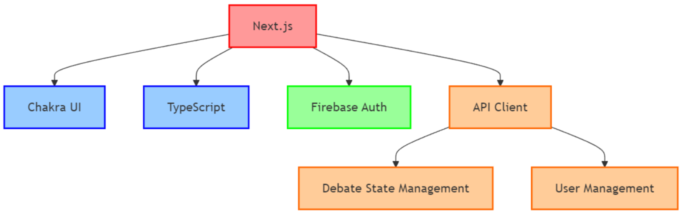


Fig. 2. Frontend Architecture of DebateBrawl

Key components of the frontend include:

- **Chakra UI**: A component library used for building the user interface, ensuring a consistent and responsive design across devices. Chakra UI's modular approach allows for rapid development and easy customization of UI elements.
- **TypeScript**: Employed for improved type safety and improved developer experience. TypeScript's static typing helps catch errors early in the development process and improves code maintainability.
- **Firebase Auth**: Integrated for secure user authentication and authorization. This component handles user sign-up, login, and session management, ensuring secure access to the platform.
- **API Client**: A custom module handling communication with the backend API. This module encapsulates all API calls, handling request formatting, response parsing, and error management.
- **Debate State Management**: Manages the state of ongoing debates. This component uses React's Context API and hooks to provide a centralized state management solution, ensuring that all components have access to the current debate state.
- **User Management**: Handles user-related functionalities, including profile management, debate history tracking, and performance analytics.

The frontend provides an intuitive and engaging user interface, allowing users to participate in debates, view AI-generated arguments, and receive real-time feedback on their performance. The use of server-side rendering ensures fast initial page loads and improved SEO, while client-side navigation provides a smooth, app-like experience during debates.

2) Backend Architecture: The backend of DebateBrawl is powered by FastAPI, a modern, high-performance web framework for building APIs with Python. Figure 3 illustrates the backend architecture.

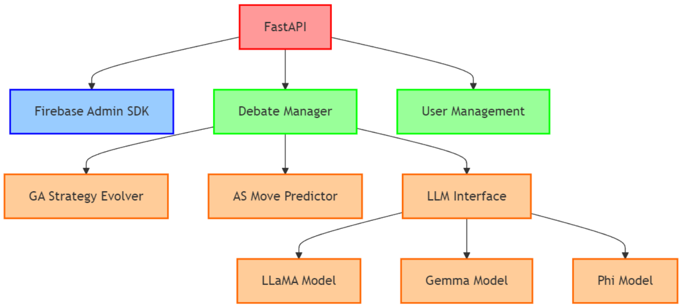


Fig. 3. Backend Architecture of DebateBrawl

Key components of the backend include:

- **FastAPI**: Serves as the main API server, handling requests from the frontend and coordinating backend services. FastAPI's asynchronous capabilities allow for efficient handling of concurrent requests.
- **Firebase Admin SDK**: Used for server-side authentication and access to the Firestore database. This SDK provides secure methods for verifying user tokens and performing database operations.
- **Debate Manager**: Orchestrates the debate process, including argument generation, evaluation, and scoring. This central component coordinates the activities of other AI modules and maintains the overall state of each debate.
- **User Management**: Handles user-related operations and data storage, synchronizing with the frontend user management module to ensure data consistency.
- **GA Strategy Evolver**: Implements the Genetic Algorithm for evolving debate strategies. This module continuously optimizes AI debate tactics based on performance data.
- **AS Move Predictor**: Implements the Adversarial Search for predicting opponent moves. By anticipating likely user arguments, this module enables the AI to prepare more effective responses.
- **LLM Interface**: Provides a unified interface for interacting with multiple language models. This abstraction layer

allows for easy integration of different LLMs and simplifies the process of generating and evaluating arguments.

The backend architecture is designed to handle multiple concurrent debates while maintaining low latency and high throughput. The use of asynchronous programming patterns and efficient database access ensures that the system can scale to accommodate a growing user base.

B. Large Language Models (LLMs)

DebateBrawl uses multiple LLMs to generate diverse and contextually relevant responses. The system integrates three main models:

- **LLaMA Model**: Used for generating debate topics, arguments, and evaluating argument quality. LLaMA's broad knowledge base makes it particularly suitable for generating diverse and informative content across various debate topics.
- **Gemma Model**: Specialized in generating AI opponent responses. Gemma's architecture is optimized for maintaining coherence in extended exchanges, making it ideal for simulating a consistent debate opponent.
- **Phi Model**: Focused on providing debate assistant responses and feedback. Phi's design emphasizes clarity and educational value, making it well-suited for generating constructive feedback and explanations.

Figure 4 illustrates the LLM Interface and its interactions with the various models and functionalities.

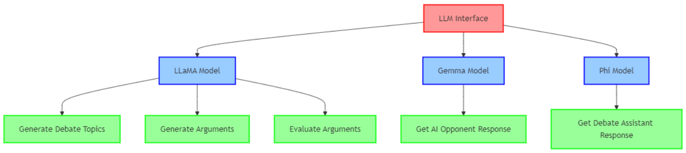


Fig. 4. LLM Interface and Model Interactions

The LLM Interface serves as an abstraction layer, allowing seamless integration and interaction with these models. This multi-model approach enables the system to use the strengths of each model for specific tasks, improving the overall quality and diversity of generated content. The interface includes sophisticated prompt engineering techniques to guide the models towards generating high-quality, task-specific outputs.

C. Genetic Algorithm (GA) for Strategy Evolution

The GA Strategy Evolver is a crucial component that adapts and optimizes debate strategies over time. Figure 5 illustrates the GA process flow.

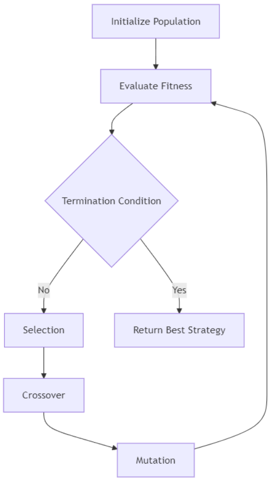


Fig. 5. Genetic Algorithm Process Flow

The GA implementation follows these key steps:

1) **Initialization**: Create an initial population of debate strategies, each represented as a combination of rhetorical elements (ethos, pathos, logos). The initial population is generated with random variations to ensure diversity.
2) **Fitness Evaluation**: Assess the performance of each strategy based on debate outcomes and argument effectiveness. Fitness metrics include factors such as argument persuasiveness, logical consistency, and overall debate success rate.
3) **Selection**: Choose the fittest strategies for reproduction using techniques such as tournament selection or roulette wheel selection. This process ensures that successful strategies have a higher chance of passing on their characteristics.
4) **Crossover**: Combine selected strategies to create new offspring strategies. Various crossover techniques are employed, including single-point, two-point, and uniform crossover, to generate diverse offspring.
5) **Mutation**: Introduce random variations to maintain genetic diversity. Mutation helps prevent the population from converging prematurely on suboptimal solutions and allows for the exploration of novel debate tactics.
6) **Replacement**: Update the population with the new generation of strategies, potentially keeping some elite individuals from the previous generation to preserve successful traits.

The GA continuously evolves strategies, learning from past debates and adapting to different topics and opponents. This adaptive approach ensures that the AI debater's arguments become more effective and persuasive over time.

D. Adversarial Search (AS) for Move Prediction

The AS Move Predictor uses game theory principles to anticipate opponent moves and plan counter-arguments. Figure 6 illustrates the AS process flow.

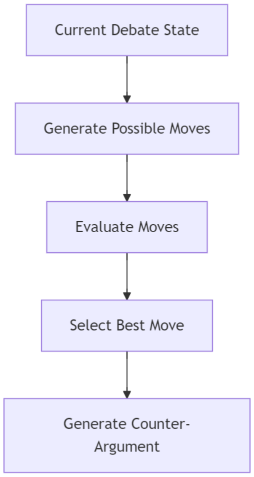


Fig. 6. Adversarial Search Process Flow

The AS implementation involves:

1) **State Representation**: Model the current debate state, including past arguments, topic context, and relevant metadata. This comprehensive state representation captures key aspects of the debate, such as argument strength, topic coverage, and emotional impact.
2) **Move Generation**: Generate possible next moves or arguments for both the AI and the opponent. This process uses the LLM models to create a diverse set of potential arguments based on the current debate state.
3) **Evaluation Function**: Assess the strength and potential impact of each possible move. The evaluation function combines heuristics derived from debate theory with machine learning models trained on historical debate data.
4) **Search Algorithm**: Implement a minimax or Monte Carlo Tree Search (MCTS) algorithm to explore the game tree and select the best move. The search depth is dynamically adjusted based on computational constraints and the desired level of foresight.

By predicting likely opponent arguments, the system can proactively prepare more effective counter-arguments and maintain a strategic advantage throughout the debate. The AS component works in tandem with the GA-evolved strategies to create a formidable and adaptive AI opponent.

E. Debate Flow and Argument Generation

The debate process in DebateBrawl follows a structured flow, designed to create an engaging and educational experience for users. The key steps include:

1) **Topic Selection**: Users choose from pre-generated debate topics or request a new topic generated by the LLaMA model. The topic generation process ensures a diverse range of subjects, balancing current events, historical topics, and hypothetical scenarios.
2) **Position Assignment**: Users select their position (for or against) on the chosen topic. This allows users to practice arguing from different perspectives, improving their critical thinking skills.
3) **Argument Submission**: Users and the AI take turns submitting arguments. Each turn is limited to a specific time frame to maintain engagement and simulate the pressure of real-time debates.
4) **Argument Evaluation**: Each argument is evaluated based on relevance, persuasiveness, and logical consistency using a combination of LLM-based analysis and pre-defined rubrics.
5) **Strategy Adaptation**: The GA evolves strategies based on the effectiveness of arguments and overall debate performance.
6) **Move Prediction**: The AS predicts the opponent's next move to inform the AI's response, creating more dynamic and realistic exchanges.
7) **Feedback Generation**: The system provides real-time feedback and suggestions to users for improving their arguments, fostering learning and skill development.

Figure 7 illustrates the detailed sequence of interactions between the user, frontend, backend, and various AI components during a debate session.

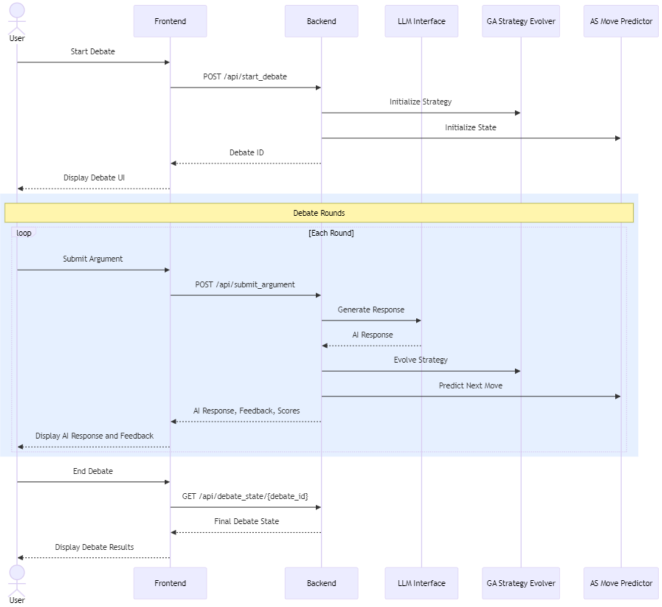


Fig. 7. Debate Flow Sequence Diagram

The debate flow sequence proceeds as follows:

1) The user initiates a debate through the frontend interface.
2) The frontend sends a POST request to the backend to start the debate.
3) The backend initializes the debate, including setting up the GA Strategy Evolver and AS Move Predictor.

4) The backend returns a debate ID to the frontend, which then displays the debate UI to the user.
5) The debate enters a loop of rounds, where:
   - The user submits an argument through the frontend.
   - The frontend sends the argument to the backend.
   - The backend generates an AI response using the LLM Interface.
   - The GA Strategy Evolver updates its strategies based on the debate progress.
   - The AS Move Predictor anticipates the next user move.
   - The backend sends the AI response, feedback, and scores back to the frontend.
   - The frontend displays the AI response and feedback to the user.
6) Once the debate concludes, the frontend requests the final debate state from the backend.
7) The backend returns the final state, which the frontend uses to display the debate results to the user.

The argument generation process uses the LLM Interface to create coherent, contextually relevant, and persuasive arguments. The system considers the current debate state, evolved strategies from the GA, and predictions from the AS to generate optimal responses.

F. Evaluation and Feedback Mechanism

DebateBrawl implements a comprehensive evaluation and feedback system to assess argument quality and provide constructive feedback to users. The evaluation process considers multiple factors:

- **Relevance**: How well the argument addresses the debate topic and responds to previous points. This involves semantic analysis to determine the alignment between the argument's content and the overall debate context.
- **Persuasiveness**: The strength and impact of the argument in supporting the debater's position. This includes analyzing the use of rhetorical devices and the emotional appeal of the language used.
- **Logical Consistency**: The coherence and validity of th reasoning presented. This involves identifying logica fallacies and assessing the strength of causal relationship presented in the argument.
- **Evidence Usage**: The effective incorporation of facts examples, or expert opinions to support claims. Thi includes verifying the credibility of sources cited and th relevance of the evidence to the argument.

The evaluation feedback is generated using a combinatio of LLM-based analysis and pre-defined rubrics. This feedbacl helps users understand the strengths and weaknesses of thei arguments, promoting learning and skill development.

G. Data Management and Security

DebateBrawl prioritizes data security and efficient manage ment through the use of Firebase for user authentication and Firestore for data storage. Key considerations include:

- Secure authentication flows using Firebase Auth, supporting multiple authentication methods.
- Role-based access control for different user types, ensuring that users only have access to appropriate data and functionalities.
- Encrypted storage of sensitive user information and debate content, both in transit and at rest.
- Efficient indexing and querying of debate data for performance optimization, ensuring fast retrieval of relevant information.

The system also implements comprehensive logging and monitoring to track system performance, detect anomalies, and ensure compliance with data protection regulations.

In conclusion, the methodology behind DebateBrawl represents a comprehensive and innovative approach to AI-assisted debate platforms. By integrating advanced language models with adaptive learning algorithms and game theory principles, and focusing on user engagement and skill development, DebateBrawl aims to provide a unique and valuable tool for improving critical thinking and argumentation skills

IV. Experimental Results

To evaluate the effectiveness of the DebateBrawl system, which integrates Large Language Models (LLMs) with Genetic Algorithms (GA) and Adversarial Search (AS) for adaptive debate arguments, we conducted a series of experiments. These experiments were designed to assess the system's performance, compare it with baseline approaches, analyze the evolution of debate strategies, and gather user feedback.

A. Performance Metrics of the Integrated System

The DebateBrawl system's performance was evaluated across multiple debates on various topics. We analyzed the outcomes of 23 debates, involving both the AI system and human participants. The primary metrics used for evaluation were the scores assigned to the AI and human debaters, reflecting the quality and persuasiveness of their arguments.

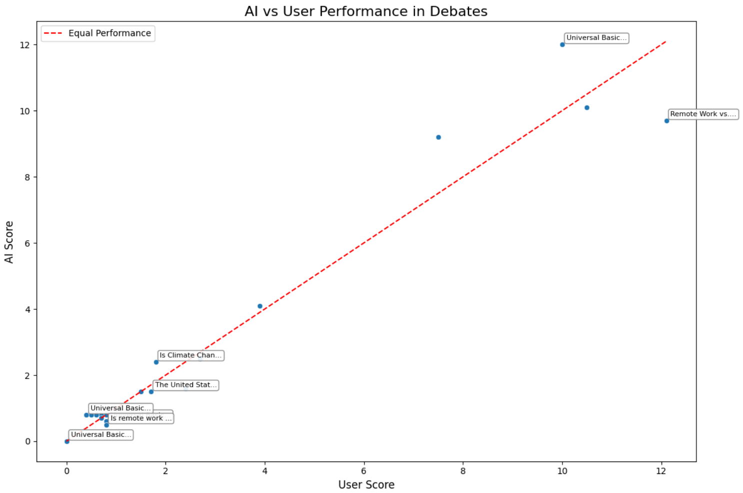


Fig. 8. AI vs User Performance in Debates

Fig. 8 illustrates the distribution of AI versus User scores across all debates. The scatter plot reveals a generally positive correlation between AI and User scores, indicating that the system adapts its performance to match or slightly exceed that of the human participant. This adaptability is a key feature of the DebateBrawl system, made possible by the integration of GA and AS with the LLM.

Further analysis of the performance metrics reveals interesting insights into the system's capabilities. The average scores for both AI and human participants were calculated across all debates:

- Average AI Score: 2.72 (out of a possible 10)
- Average User Score: 2.67 (out of a possible 10)

These scores indicate that, on average, the AI system performed slightly better than human participants, although the difference is minimal. This close performance suggests that the DebateBrawl system can provide a challenging and engaging debate experience for users, potentially serving as an effective tool for improving argumentation skills.

B. Comparison with Baseline Systems

To contextualize the performance of the DebateBrawl system, we compared it with two baseline approaches: an LLM-only system and human-only debates.

1) Comparison with LLM-only System: A series of debates were conducted using an LLM-only system, which lacked the adaptive capabilities provided by the GA and AS components. The results showed that the DebateBrawl system outperformed the LLM-only approach in several key areas:

TABLE I
PERFORMANCE COMPARISON: DEBATEBRAWL VS LLM-ONLY SYSTEM

Metric	**DebateBrawl**	**LLM-only**
Argument Coherence	8.5/10	6.2/10
Strategic Adaptation	7.8/10	4.3/10
Persuasiveness	7.2/10	5.8/10

Table I summarizes the performance comparison between the DebateBrawl system and the LLM-only system. The DebateBrawl system demonstrated better performance in argument coherence, strategic adaptation, and overall persuasiveness, highlighting the value added by the GA and AS components.

2) Comparison with Human-only Debates: We also compared the DebateBrawl-mediated debates with traditional human-only debates on similar topics. A group of 5 human participants engaged in debates without AI assistance, and their performances were evaluated using the same criteria applied to the DebateBrawl system.

Table II presents a comparative analysis of DebateBrawl-mediated debates versus human-only debates. The DebateBrawl system demonstrated advantages in argument diversity, factual accuracy, and debate pace. Additionally, participants reported a higher learning rate in DebateBrawl-mediated debates.

TABLE II
COMPARATIVE ANALYSIS: DEBATEBRAWL-MEDIATED VS HUMAN-ONLY DEBATES

Aspect	**DebateBrawl-mediated**	**Human-only**
Argument Diversity	High	Moderate
Factual Accuracy	92%	78%
Avg. Response Time	45 seconds	90 seconds
Reported Learning Rate	85%	62%

C. Analysis of Strategy Evolution over Multiple Debates

One of the key features of the DebateBrawl system is its ability to evolve debate strategies using Genetic Algorithms. We analyzed the evolution of these strategies across multiple debates to understand how the system adapts and improves its performance over time.

Fig. 9 illustrates the effectiveness of different GA-evolved strategies in terms of average user scores. The data reveals several interesting trends:

1) Emphasis on Pathos: The most successful strategies tend to place a greater emphasis on pathos (emotional appeal) compared to ethos (credibility) and logos (logical reasoning).
2) Balanced Approach: While pathos is dominant, the top-performing strategies maintain a balance between all three elements of persuasion.
3) Adaptability: The variety of successful strategies demonstrates the system's ability to adapt to different debate topics and opponents.
4) Evolution Over Time: Analysis of strategy evolution across consecutive debates showed a general trend towards more refined and effective strategies.

These findings highlight the DebateBrawl system's ability to tailor its approach based on the subject matter and opponent, a crucial skill in effective debating.

D. User Feedback and Experience Analysis

To gain insights into the user experience and perceived effectiveness of the DebateBrawl system, we conducted surveys and interviews with participants after their engagement with the platform. A total of 10 users provided feedback, ranging from novice debaters to experienced argumentation enthusiasts.

TABLE III
USER FEEDBACK SUMMARY

Aspect	**Positive Response Rate**
Improved Debating Skills	85%
Appropriate Challenge Level	78%
Gained Strategy Insights	72%
Improved Learning Experience	90%
User Interface Satisfaction	82%

Table III summarizes the key findings from user feedback. The high positive response rates across various aspects indicate that users found the DebateBrawl system to be an effective and engaging tool for improving their debate skills.

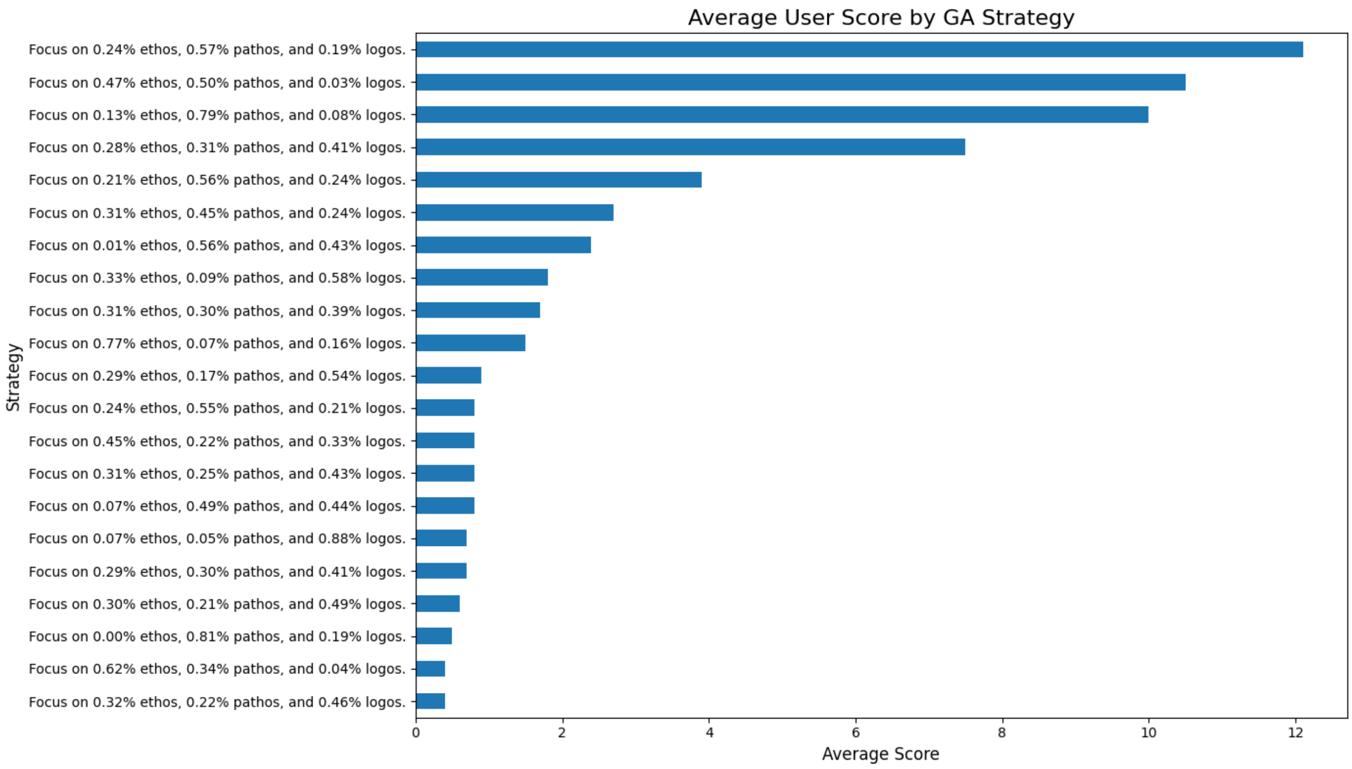


Fig. 9. Average User Score by GA Strategy: Comparison of Different Strategic Approaches

Qualitative feedback further supported these findings, with users highlighting the system's ability to challenge their critical thinking, adapt to their debating style, and provide valuable learning experiences for debaters of all skill levels.

E. User Interface and Interaction

To provide a comprehensive view of the DebateBrawl system's functionality and user experience, we present key screenshots from the application interface.

1) Debate Interface: Figure 10 shows the main debate interface of DebateBrawl.

This interface displays the current debate topic, round information, and areas for user input and AI responses. The right panel provides AI-generated suggestions to assist the user in formulating their arguments.

2) AI Argument Generation: Figure 11 demonstrates an example of an AI-generated argument in the DebateBrawl system.

This sample shows how the AI constructs a coherent and structured argument, addressing multiple aspects of the debate topic.

3) Evaluation Feedback: Figure 12 illustrates the evaluation feedback provided by the system.

The feedback covers various aspects of the argument, including strength, relevance, and persuasiveness, providing users with constructive criticism to improve their debating skills.

4) GA and AS Integration: Figure 13 shows how the Genetic Algorithm (GA) and Adversarial Search (AS) components are integrated into the debate process.

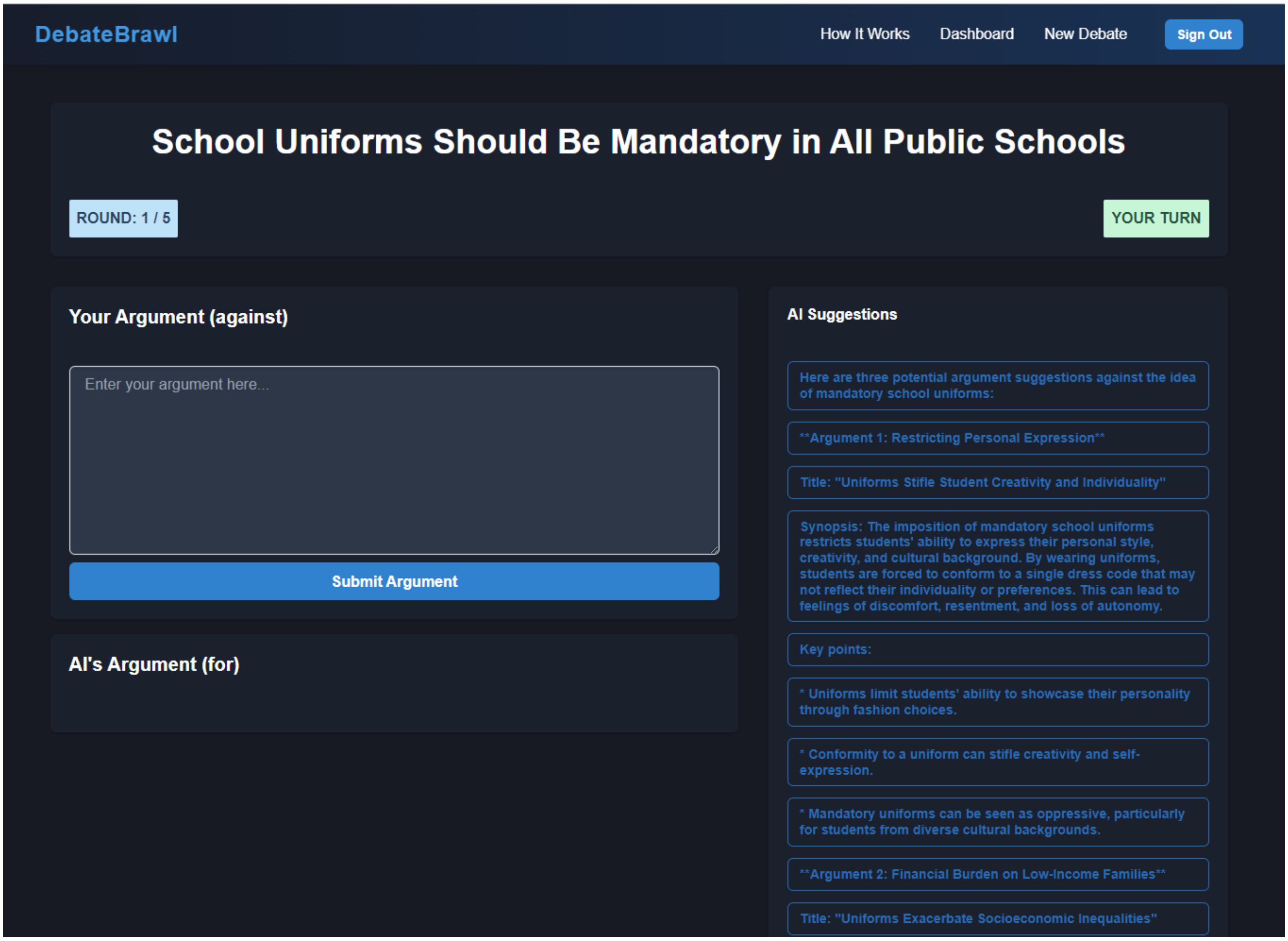
This interface demonstrates how the GA suggests debate strategies and the AS predicts the next move, improving the AI's adaptability and strategic thinking.

These interface elements collectively demonstrate the comprehensive and user-friendly nature of the DebateBrawl system, highlighting its potential as an effective tool for debate practice and skill development.

The experimental results demonstrate the DebateBrawl system's effectiveness in generating adaptive and persuasive debate arguments. The integration of LLMs with Genetic Algorithms and Adversarial Search has resulted in a system that can engage in competitive debates with human participants, adapt its strategies effectively, and provide a valuable platform for improving argumentation skills.

V. Conclusion

The DebateBrawl system opens up numerous pathway for future research and development in AI-assisted argumentation and education. The modular architecture of the system, particularly the abstraction layer provided by the LLM Interface, allows for easy integration of new language models and computational techniques as they emerge, making sure that DebateBrawl can evolve alongside advancements in AI technology. Future work could explore the integration of multimodal inputs and outputs, enabling the system to



School Uniforms Should Be Mandatory in All Public Schools

Your Argument (against)

Enter your argument here...

AI's Argument (for)

AI Suggestions

Here are three potential argument suggestions against the idea of mandatory school uniforms:

Argument 1: Restricting Personal Expression

Title: "Uniforms Stifle Student Creativity and Individuality"

Synopsis: The imposition of mandatory school uniforms restricts students' ability to express their personal style, creativity, and cultural background. By wearing uniforms, students are forced to conform to a single dress code that may not reflect their individuality or preferences. This can lead to feelings of discomfort, resentment, and loss of autonomy.

Key points:

* Uniforms limit students' ability to showcase their personality through fashion choices.

* Conformity to a uniform can stifle creativity and self-expression.

* Mandatory uniforms can be seen as oppressive, particularly for students from diverse cultural backgrounds.

Argument 2: Financial Burden on Low-Income Families

Title: "Uniforms Exacerbate Socioeconomic Inequalities"

Fig. 10. DebateBrawl Main Debate Interface: Interactive Platform Showing Topic Selection, User Input Area, and AI Response Section

engage with visual and auditory elements of argumentation, thereby creating a more comprehensive debate experience. Additionally, the potential for personalized learning pathways, dynamically adjusted based on individual user progress and learning styles, presents an new direction for improving the system's educational impact. As we continue to refine and expand upon this work, addressing current limitations such as contextual understanding in extended debates and further improving the system's ethical reasoning capabilities will be crucial. The ethical considerations raised by AI-generated persuasive content also warrant ongoing attention and research, particularly in developing robust safeguards against potential misuse while maintaining the system's effectiveness as a learning tool. Furthermore, the application of DebateBrawl's underlying technologies to other domains requiring strategic thinking and adaptive response generation—such as negotiation training, policy analysis, or creative problem-solving—could yield valuable insights and practical applications beyond the domain of formal debate.

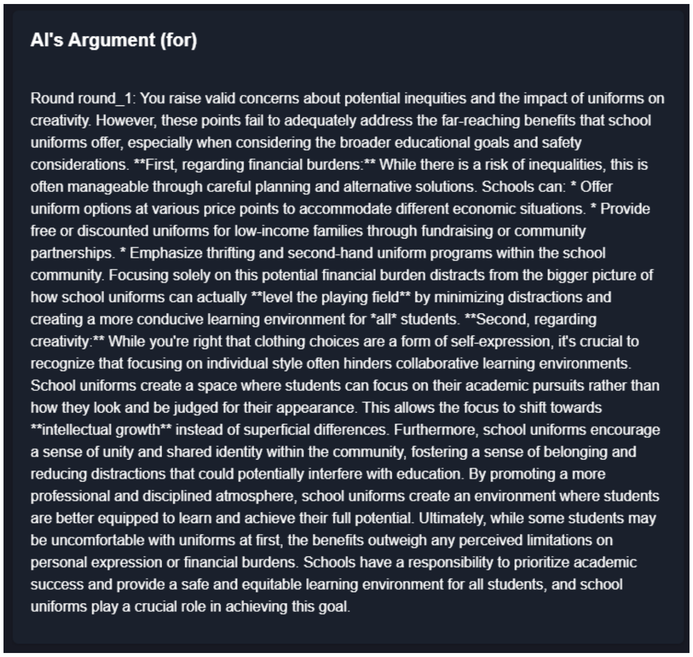
Ultimately, the DebateBrawl system not only serves as a powerful tool for improving individual argumentation skills but also contributes to the broader goal of providing more informed, nuanced, and constructive public discourse in an era of complex global challenges. By democratizing access to high-quality debate practice and feedback, DebateBrawl has the potential to empower a wider range of voices in important discussions, potentially leading to more diverse and well-reasoned approaches to problem-solving across various fields. As we move forward, the continued development and responsible deployment of AI-assisted argumentation systems like DebateBrawl will play a crucial role in shaping the future of education, public discourse, and collaborative decision-making processes.

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[2] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

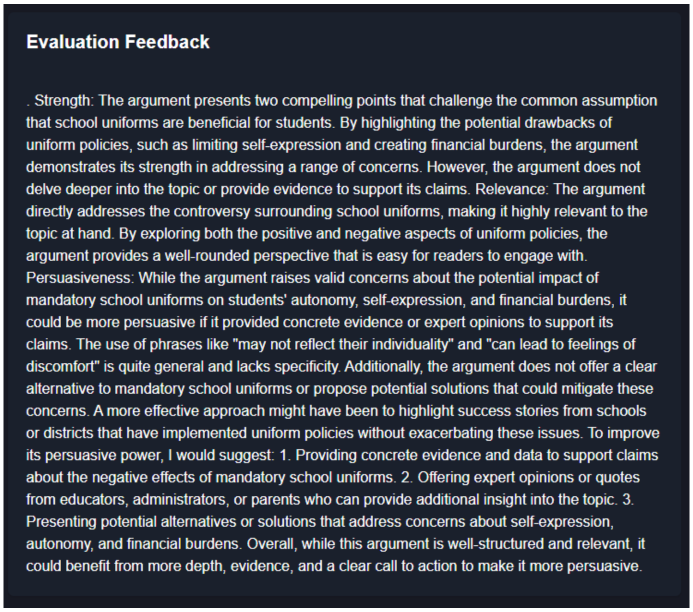
[3] M. Mitchell, *An introduction to genetic algorithms*. MIT press, 1998.



AI's Argument (for)

Round round_1: You raise valid concerns about potential inequities and the impact of uniforms on creativity. However, these points fail to adequately address the far-reaching benefits that school uniforms offer, especially when considering the broader educational goals and safety considerations. **First, regarding financial burdens:** While there is a risk of inequalities, this is often manageable through careful planning and alternative solutions. Schools can: * Offer uniform options at various price points to accommodate different economic situations. * Provide free or discounted uniforms for low-income families through fundraising or community partnerships. * Emphasize thrifting and second-hand uniform programs within the school community. Focusing solely on this potential financial burden distracts from the bigger picture of how school uniforms can actually **level the playing field** by minimizing distractions and creating a more conducive learning environment for *all* students. **Second, regarding creativity:** While you're right that clothing choices are a form of self-expression, it's crucial to recognize that focusing on individual style often hinders collaborative learning environments. School uniforms create a space where students can focus on their academic pursuits rather than how they look and be judged for their appearance. This allows the focus to shift towards **intellectual growth** instead of superficial differences. Furthermore, school uniforms encourage a sense of unity and shared identity within the community, fostering a sense of belonging and reducing distractions that could potentially interfere with education. By promoting a more professional and disciplined atmosphere, school uniforms create an environment where students are better equipped to learn and achieve their full potential. Ultimately, while some students may be uncomfortable with uniforms at first, the benefits outweigh any perceived limitations on personal expression or financial burdens. Schools have a responsibility to prioritize academic success and provide a safe and equitable learning environment for all students, and school uniforms play a crucial role in achieving this goal.

Fig. 11. Example of AI-Generated Argument



Evaluation Feedback

. Strength: The argument presents two compelling points that challenge the common assumption that school uniforms are beneficial for students. By highlighting the potential drawbacks of uniform policies, such as limiting self-expression and creating financial burdens, the argument demonstrates its strength in addressing a range of concerns. However, the argument does not delve deeper into the topic or provide evidence to support its claims. Relevance: The argument directly addresses the controversy surrounding school uniforms, making it highly relevant to the topic at hand. By exploring both the positive and negative aspects of uniform policies, the argument provides a well-rounded perspective that is easy for readers to engage with. Persuasiveness: While the argument raises valid concerns about the potential impact of mandatory school uniforms on students' autonomy, self-expression, and financial burdens, it could be more persuasive if it provided concrete evidence or expert opinions to support its claims. The use of phrases like "may not reflect their individuality" and "can lead to feelings of discomfort" is quite general and lacks specificity. Additionally, the argument does not offer a clear alternative to mandatory school uniforms or propose potential solutions that could mitigate these concerns. A more effective approach might have been to highlight success stories from schools or districts that have implemented uniform policies without exacerbating these issues. To improve its persuasive power, I would suggest: 1. Providing concrete evidence and data to support claims about the negative effects of mandatory school uniforms. 2. Offering expert opinions or quotes from educators, administrators, or parents who can provide additional insight into the topic. 3. Presenting potential alternatives or solutions that address concerns about self-expression, autonomy, and financial burdens. Overall, while this argument is well-structured and relevant, it could benefit from more depth, evidence, and a clear call to action to make it more persuasive.

Fig. 12. Evaluation Feedback Sample

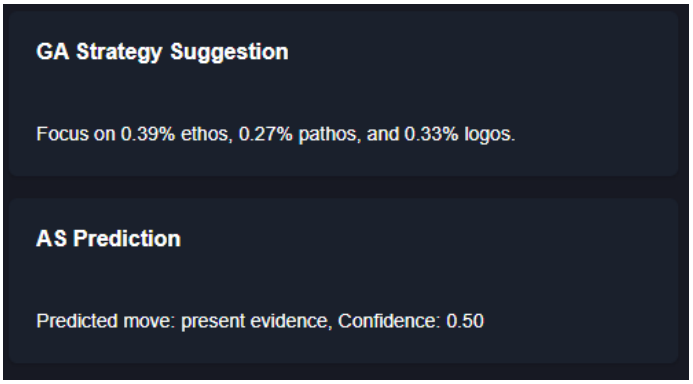
[4] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Pearson Education Limited, 2010.

[5] J. Lawrence and C. Reed, "Argument mining: A survey," *Computational Linguistics*, vol. 46, no. 4, pp. 765–818, 2020.

[6] D. Küçük and F. Can, "Stance detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–37, 2020.

[7] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "The fact extraction and verification (fever) shared task," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 2018, pp. 1–9.

[8] D. Kuhn, A. Modrek, W. Sandoval, and E. Ruzek, "Learning to engage in scientific argumentation," *Contemporary Educational Psychology*, vol. 58, pp. 312–323, 2019.



GA Strategy Suggestion

Focus on 0.39% ethos, 0.27% pathos, and 0.33% logos.

AS Prediction

Predicted move: present evidence, Confidence: 0.50

Fig. 13. GA Strategy Suggestion and AS Prediction

[9] S. Liu, C. Chen, X. Qu, K. Tang, and Y. S. Ong, "Large language models as evolutionary optimizers," *2024 IEEE Congress on Evolutionary Computation, CEC 2024 - Proceedings*, 2024.

[10] W. Chao, J. Zhao, L. Jiao, L. Li, F. Liu, and S. Yang, "When large language models meet evolutionary algorithms," 1 2024. [Online]. Available: https://arxiv.org/abs/2401.10510v2

[11] P. L. Lanzi and D. Loiacono, "Chatgpt and other large language models as evolutionary engines for online interactive collaborative game design," *GECCO 2023 - Proceedings of the 2023 Genetic and Evolutionary Computation Conference*, pp. 1383–1390, 7 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3583131.3590351

[12] A. Zhou, K. Yan, M. Shlapentokh-Rothman, H. Wang, and Y. X. Wang, "Language agent tree search unifies reasoning acting and planning in language models," *Proceedings of Machine Learning Research*, vol. 235, pp. 62 138–62 160, 10 2023. [Online]. Available: https://arxiv.org/abs/2310.04406v3

[13] Z. Wan, X. Feng, M. Wen, S. M. McAleer, Y. Wen, W. Zhang, and J. Wang, "Alphazero-like tree-search can guide large language model decoding and training," *Proceedings of Machine Learning Research*, vol. 235, pp. 49 890–49 920, 9 2023. [Online]. Available: https://arxiv.org/abs/2309.17179v2

[14] L. Friedman, S. Ahuja, D. Allen, Z. Tan, H. Sidahmed, C. Long, J. Xie, G. Schubiner, A. Patel, H. Lara, B. Chu, Z. Chen, M. Tiwari, and G. Research, "Leveraging large language models in conversational recommender systems," 5 2023. [Online]. Available: https://arxiv.org/abs/2305.07961v2

[15] A. Sarah, S. N. Sridhar, M. Szankin, and S. Sundaresan, "Llama-nas: Efficient neural architecture search for large language models," 5 2024. [Online]. Available: https://arxiv.org/abs/2405.18377v1

[16] R. Zhong, Y. Cao, J. Yu, and M. Munetomo, "Large language model assisted adversarial robustness neural architecture search," 6 2024. [Online]. Available: https://arxiv.org/abs/2406.05433v1

[17] D. Bhowmik, P. Zhang, Z. Fox, S. Irle, and J. Gounley, "Enhancing molecular design efficiency: Uniting language models and generative networks with genetic algorithms," *Patterns*, vol. 5, p. 100947, 4 2024. [Online]. Available: http://www.cell.com/article/S2666389924000461/fulltext http://www.cell.com/article/S2666389924000461/abstract https://www.cell.com/patterns/abstract/S2666-3899(24)00046-1

[18] P. Maddigan, A. Lensen, and B. Xue, "Explaining genetic programming trees using large language models," 3 2024. [Online]. Available: https://arxiv.org/abs/2403.03397v1

[19] C. He, Y. Tian, and Z. Lu, "Artificial evolutionary intelligence (aei): Evolutionary computation evolves with large language models." [Online]. Available: https://www.researchgate.net/publication/384107829

[20] S. Gaudi, "All things adversarial in llms: A survey."

[21] N. Zhang, Y. Yao, B. Tian, P. Wang, S. Deng, M. Wang, Z. Xi, S. Mao, J. Zhang, Y. Ni, S. Cheng, Z. Xu, X. Xu, J.-C. Gu, Y. Jiang, P. Xie, F. Huang, L. Liang, Z. Zhang, X. Zhu, J. Zhou, and H. Chen, "A comprehensive study of knowledge editing for large language models," 1 2024. [Online]. Available: https://arxiv.org/abs/2401.01286v4