SocrAItic Circle: Enhancing LLM Reasoning Through Multi-Agent Debate Frameworks

Tanay Grover
Northeastern University
grover.t@northeastern.edu

Joel Markapudi
Northeastern University
markapudi.j@northeastern.edu

Gaurav Kothamachu Harish
Northeastern University
kothamachuharish.g@northeastern.edu

Pengkun Ma
Northeastern University
ma.peng@northeastern.edu

Easha Meher Koppisetty
Northeastern University
koppisetty.e@northeastern.edu

***Abstract*—Large Language Models (LLMs) often struggle with sustained logical reasoning and factual consistency in complex argumentative scenarios. This research presents SocrAItic Circle, a novel multi-agent debate framework that significantly enhances LLM reasoning capabilities through structured adversarial interaction and iterative feedback-driven cycles. Unlike traditional self-reflection approaches, our system implements specialized debate phases (preparation, critique, rebuttal) alongside a theory-grounded evaluation system featuring seven distinct judge types—each analyzing different dimensions of argument quality from logical structure to audience impact. Through controlled experimentation across multiple debate topics and model architectures, we demonstrate that this approach produces measurably stronger arguments with reduced logical fallacies, improved evidence integration, and enhanced persuasive coherence. Quantitative evaluation reveals our multi-agent framework consistently outperforms single-agent approaches across key quality metrics, with particularly strong improvements in factual accuracy and strategic framing. This work establishes a foundation for human-AI collaborative argumentation systems with applications in education, policy analysis, and critical reasoning development.**

***Index Terms*—Large Language Models (LLMs),**

I. Introduction

A. The Challenge of LLM Argumentation

Large Language Models (LLMs) have demonstrated remarkable capabilities in text generation, yet struggle with sustained logical reasoning in complex argumentative scenarios. Despite producing fluent responses, these models often fail to maintain coherence, factual accuracy, and persuasive strength across extended exchanges.

Can structured debate environments help LLMs develop better reasoning capabilities?

B. Research Focus and Educational Applications

Our research investigates whether multi-agent debate frameworks with specialized feedback mechanisms can improve LLM argument quality compared to single-agent approaches. This question has significant implications for:

- **Educational contexts:** Supporting critical thought through AI-assisted debate practice and feedback
- **Argumentative skill building:** Providing students with practice partners that can model and analyze different debate techniques
- **Instructor support:** Offering customizable debate scenarios and structured evaluation frameworks
- **Self-directed learning:** Enabling students to engage with complex topics through guided dialectical exploration

C. Research Question and Challenges

Primary Question: *Does a multi-agent debate framework with hybrid human-AI feedback improve the logical consistency, factual accuracy, and persuasive strength of arguments generated by LLMs compared to single-agent self-reflection methods?*

This exploration faces several significant challenges:

- Maintaining context across multiple conversation rounds
- Designing reliable, theory-grounded feedback
- Ensuring high-quality evaluation across diverse domains
- Mitigating API cost, latency, and compute constraints

D. The SocrAItic Circle Approach

Our system deploys a debate architecture featuring:

1) **Dual Agent Design:** Two LLM-based agents (Agent A and Agent B) representing opposing viewpoints
2) **Three-Phase Debate Structure:**
   - *Preparation:* Strategic evidence gathering, argument planning, and fallacy avoidance
   - *Rebuttal:* Targeted counterargument deployment, using structured opposition techniques
   - *Evaluation:* Multi-dimensional quality assessment
3) **Theory-Informed Evaluation:** A multi-judge assessment system drawing from established theories in argumentation, rhetoric, and persuasion

E. Research Hypotheses

We formulate three primary hypotheses:

H1: A multi-judge evaluation system enhances scoring reliability and provides more nuanced feedback.

H2: Exposure to structured adversarial challenges reduces logical fallacies and improves argument persuasiveness.

H3: Feedback-driven iterative argumentation produces measurably better debates than systems without such feedback.

II. Related Works

Recent work in multi-agent collaboration has demonstrated that LLMs can improve their performance **through interaction** rather than parameter finetuning.

CoEvol [2] presents a co-evolutionary framework where LLMs iteratively generate, critique, and revise instruction-following responses. By simulating peer review cycles and enabling role switching, CoEvol constructs stronger and more context-aware outputs through multi-agent cooperation.

The Multi-Agent Debate (MAD) framework [3] allows LLMs to engage in structured debates, where each model presents its argument, critiques the opposition, and adapts based on previous responses, encouraging robust reasoning.

The formal structure of our debate protocol draws from Bench-Capon and Dunne's work on computational argumentation frameworks [5], which provides models for structured exchanges between opposing agents. The feedback mechanisms in our system are informed by persuasive systems design research [7], which emphasizes the importance of timely, targeted feedback for behavioral change.

III. Methodologies

A. Theoretical Foundations of Design

Our system design draws upon established theories in argumentation, persuasion, and cognitive processing. Toulmin's Model of Argumentation [8] informed our structured debate prompts, requiring agents to provide claims supported by evidence, warrants linking them, and acknowledgment of rebuttals.

To evaluate persuasive elements, we incorporated both classical rhetorical appeals and Cialdini's principles of influence [4]—specifically using his authority principle to assess factual strength, while applying scarcity and social proof principles to evaluate rhetorical effectiveness. The judge evaluation component implements aspects of the Elaboration Likelihood Model [9]. This hybrid theoretical approach ensures comprehensive evaluation across multiple dimensions of argument quality.

B. System Architecture

Our multi-agent debate framework implements a structured approach to improve LLM argumentative capabilities through adversarial training and specialized evaluation.

Adversarial Approach: Multi-Agent Debate Flow

- **Agents Involved:** Two agents (Agent A and Agent B) independently generate responses.
- **Preparation Phase:**
  - Agent performs a **Self-Check**.
  - Agent enters **Critique Phase**, where it critiques the argument of the opposing agent.
- Agent gives a **Rebuttal** to defend, refine their stance.
- Rebuttals and critiques are written to a **Data Persistence** module for cyclic, or downstream work.

Judgement Pipeline

- **Saved Results** → Load critiques, rebuttals, and agent logs.
- **Parse and Format** → Prepare data for evaluation pipeline.
- **Run Judge Prompts** → LLM-based evaluators assess arguments.
- **Score Extraction** → Capture evaluation metrics and ranks.
- **Final Table** → Aggregate scores into a presentable format.

1) Overview of Architecture: The SocrAItic Circle framework consists of three primary components: the Prompt Management System, the Multi-Agent Debate Engine, and the Judge Evaluation Pipeline.

2) Prompt Management System: Our prompt engineering approach uses a YAML-based configuration system that separates debate instructions from implementation logic. This design enables:

- **Modularity**: Different prompt strategies can be tested without modifying core code
- **Theoretical grounding**: Prompts incorporate Toulmin's Argumentation Model and Cialdini's Influence Principles
- **Phase-specific guidance**: Tailored instructions for opening, rebuttal, and closing phases

The specificed Prompt Manager class handles the loading, formatting, and delivery of these structured prompts, ensuring consistent application of said frameworks across the stages.

3) Formal Debate Structure: Our framework implements a structured three-round debate format inspired by competitive academic practices:

- **Opening Round**: Debaters establish positions with defined claims, conceptual framing, and evidence
- **Rebuttal Round**: Agents directly address opponents' arguments while advancing their own position through a structured four-part response
- **Closing Round**: Debaters synthesize the exchange, highlighting key points of clash and framing the debate's conclusion

This rigid structure guides LLMs toward argumentation patterns that would be difficult to achieve in free-form exchanges.

4) Multi-Agent Debate Engine: The debate engine orchestrates structured interactions between opposing LLM agents. Unlike single-agent approaches, our system implements:

- **Preparation**: Agents perform self-checks on evidence quality and anticipate counterarguments
- **Critique**: Agents analyze opposing arguments for fallacies and factual weaknesses
- **Rebuttal**: Responses incorporate critique findings and targeted refutations

This approach mirrors the cognitive processes of skilled human debaters, who analyze and prepare before responding—a critical advantage over simple turn-taking architectures.

5) Adversarial Critique Mechanism: Central to our approach is a sophisticated adversarial critique system that challenges arguments through targeted analysis:

- **Structured Analytical Framework**: The critique prompt designates the agent as a "critical thinking expert" tasked with identifying specific weaknesses in the opponent's argument, creating a focused adversarial mindset
- **Five-Dimension Weakness Analysis**: Prompts instruct agents to systematically evaluate: (1) logical gaps in reasoning, (2) factual errors or misrepresentations, (3) unstated assumptions without justification, (4) vulnerability to counterarguments, and (5) rhetorical weaknesses or missed persuasive opportunities
- **Quote-Based Evidence**: Critiques must include direct quotes from the argument when identifying problems, ensuring specificity rather than general criticisms
- **Constructive Alternatives**: For each weakness identified, the system requires providing specific examples of stronger alternatives (e.g., "A stronger connection would show that X leads to Y by explaining Z")
- **Strategic Prioritization**: Prompts explicitly instruct agents to focus 80% of analysis on the 3-4 most damaging weaknesses rather than listing many minor issues. This ensures the ethical norm of dialectical fairness is upheld.
- **Metacognitive Framing**: Critiques are presented to the original agent as external expert analysis to encourage objective consideration rather than defensive dismissal

6) Evidence Self-Check Mechanism: A key innovation in our framework is the evidence self-check mechanism that operates before each debate response. This system prompts agents to critically examine their own arguments for potential weaknesses:

- **Claim Validation**: The agent identifies main claims and checks if each has adequate supporting evidence
- **Source Assessment**: The agent evaluates cited sources and identifies where stronger evidence is needed
- **Opponent Analysis**: The agent examines the opponent's argument for factual errors or logical weaknesses
- **Citation Integrity**: The agent verifies that citations are accurate and appropriately contextualized

This self-reflection process mirrors expert debater preparation and significantly reduces the **tendency of LLMs to fabricate sources** or make unsupported claims. Importantly, this mechanism operates as a preparation step rather than simply filtering output, allowing the agent to strategically strengthen areas of evidential weakness before generating its rebuttal.

7) Persistent Memory and Preparation Architecture: Our system implements a sophisticated storage architecture that manages both the visible debate transcript and the invisible preparation steps:

- **Debate Storage Layer**: A specialized Debate Storage class maintains a hierarchical structure of rounds, exchanges, and preparation materials, enabling agents to reference previous arguments while maintaining a clean separation between public debate content and private preparation work
- **Preparation Pipeline**: Each debate round incorporates three distinct preparation steps that remain invisible in the final transcript:
  - *Evidence Check*: Self-examination of claim support and citation integrity
  - *Adversarial Critique*: Analysis of opponent's argument for weaknesses
  - *Enhanced Prompt Formulation*: Combining base debate prompt with preparation materials
- **Context Management**: The system carefully manages context windows by selectively incorporating only relevant preparation materials and previous arguments, enabling deeper analysis without exceeding token limits

This architecture allows for significantly more sophisticated agent behavior than simple prompt-response systems, as **each visible debate exchange is the product of multiple preparatory reasoning steps** that mirror human debate preparation processes.

C. Evaluation Methodology

1) Multi-Dimensional Judge Framework: Our evaluation system employs specialized LLM-based judges, each focusing on different aspects of argument quality. This multi-judge approach is inspired by:

- **Aristotelian Rhetorical Appeals**: The rhetorical judge evaluates the classic trio of ethos (credibility), pathos (emotional appeal), and logos (logical reasoning) as interconnected components of persuasive communication.
- **Anthropic's Persuasiveness Methodology**: Belief-shift judge implements Anthropic's approach to measuring persuasive impact on audience perspective. Estimates potential shifts in audience stance before and after exposure
- **Elaboration Likelihood Model (ELM)**: Assessing both central (logical) and peripheral (rhetorical) routes to persuasion
- **Toulmin's Model**: Evaluating claims, evidence, and warrants independently
- **Audience Segmentation Theory**: The audience judge evaluates reception across multiple audience types (neutral, skeptical, and supportive).
- **Meta-Evaluation Consensus Theory**: A meta-judge implements advanced consensus forming methodology, weighting the evaluations and resolving divergent assessments through principled aggregation techniques.
- **Pragma-Dialectical Theory**: Ethical and strategic judges assess adherence to Van Eemeren and Grootendorst's pragma-dialectical norms, evaluating conduct, fairness in representation, and strategic maneuvering
- **Cognitive Dissonance Theory**: Addresses how the judge evaluates arguments' **ability to overcome resistance to position changes** by resolving psychological inconsistencies between prior beliefs and new information.

2) Judge Types and Evaluation Criteria: The system implements seven specialized judge types:

Judge Type	Theoretical Foundation	Evaluation Focus
Factual	Cialdini's Authority Principle	Accuracy and evidence quality
Logical	Central Route Processing (ELM)	Reasoning structure and fallacy detection
Rhetorical	Cialdini's Liking & Scarcity Principles	Persuasive language and stylistic effectiveness
Belief-Shift	Anthropic's Persuasion Methodology	Potential to change opinions
Audience	Cialdini's Social Proof Principle	Resonance with different audience segments
Strategic	Pragma-Dialectical Tactics	Argument selection and adaptive responses
Ethical	Pragma-Dialectical Fairness	Intellectual honesty and fair representation

TABLE I
JUDGE TYPES WITH THEIR THEORETICAL FOUNDATIONS AND EVALUATION FOCUS AREAS

D. Experimental Design

To evaluate our hypotheses, we conducted debates across multiple topics using the following conditions:

- **Control Condition**: Single-agent self-reflection approach (no adversarial critique)
- **Experimental Condition A**: Multi-agent debate with critique phase, single judge
- **Experimental Condition B**: Multi-agent debate with critique phase, multi-judge evaluation
- **Topics**: Covering ethical, scientific, and policy domains

Performance was measured across logical consistency, factual accuracy, and persuasive strength using our multi-dimensional judge framework, with **some additional human evaluation** at times, to validate system assessments.

E. Integration of Methodological Variants

The framework described thus far establishes the foundation of our multi-agent debate system through structured agent interactions and theory-informed evaluation.

In the following sections, we present **more methodological variants**, evaluation experimentals that extend and enhance this core framework.

F. Methodological Extensions - Further Variants

1) Judge-Guided Improvement Cycle: We experimented on a judge-guided improvement cycle that enables iterative refinement of arguments based on specialized feedback:

- **Real-time Revision Loop**: After each debate round, arguments are evaluated by specialized judges whose feedback is incorporated into subsequent rounds
- **Semantic Change Tracking**: We measured the degree of difference between original and revised arguments using Sentence-BERT embeddings
- **Feedback Alignment Analysis**: We quantified how closely revisions followed judge suggestions, tracking implementation of specific recommendations

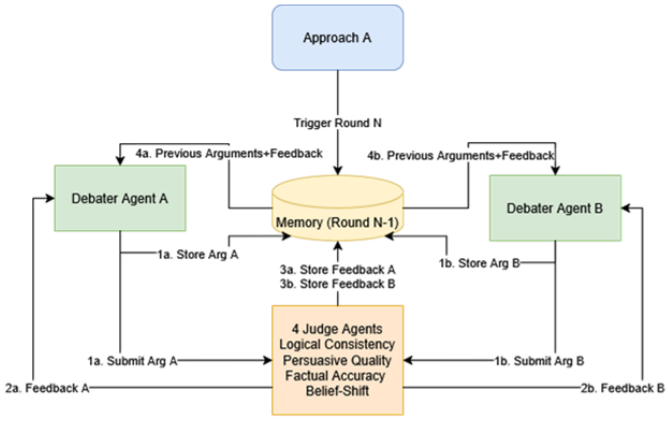

Fig. 1. Approach A Flowchart

Our analysis shows semantic change scores ranging from 0.32 to 0.48 across debate rounds, indicating meaningful yet measured revisions. Feedback alignment scores between 0.66 and 0.81 demonstrate that revisions effectively incorporated judge recommendations, with alignment increasing in later rounds as speakers became more responsive to structured feedback.

2) Alternative Evaluation Approaches: While our primary evaluation framework relies on specialized judges, we explored several complementary measurement techniques:

- **Semantic Similarity Metrics**: Lightweight embedding-based measures proved most compatible with our real-time feedback requirements
- **Automated Readability Analysis**: Standard readability metrics (Flesch-Kincaid) were tested but ultimately rejected as they penalized domain-appropriate technical language without capturing persuasive quality
- **Advanced NLP Techniques**: We investigated BLEURT, natural language inference models, and lexical overlap metrics (ROUGE, BERTScore), but these either required extensive fine-tuning or failed to capture the deeper logical revisions we aimed to evaluate

These explorations informed our decision to prioritize lightweight semantic similarity techniques, or LLM-based evaluation.

G. Approach A: Structured Debate with Delayed Feedback

1) Architecture and Debate Flow: Approach A employs a structured two-agent debate framework (Fig. 1) inspired from the structure of a traditional debate. Two LLM-based agents, Agent A and Agent B, are assigned opposing positions on a given topic. Each debate is organized into sequential rounds:

1) *Opening Round:* Agents present initial positions.
2) *Rebuttal Rounds:* Multiple rounds where agents refute opponent points and reinforce their own arguments. Agent A argues, then Agent B responds within each round.
3) *Closing Round:* Agents summarize their case and argue for its overall strength.

Both agents have persistent conversational memory, meaning they consider all arguments from **prior rounds** when formulating responses. They are prompted with role-specific context (e.g. "You are a Policy Advocate arguing *for* regulation" vs "You are a Free Speech Advocate arguing *against*") to ensure diametrically opposed viewpoints. Prior to the live debate, each agent is instructed to outline its claims, warrants, and potential rebuttals. However, in Approach A this preparation is one-shot; once the debate starts, agents do not receive external feedback on their arguments until the round is over.

2) Round-wise Judge Feedback: We use a **Delayed Feedback Loop** here. Crucially, the scores and qualitative feedback from the judges are **not revealed** to the agents during the round in which the argument was made. Instead, this feedback is recorded and provided to the agent *before* they formulate their argument for the *next* round.

We also ensure **Independent Evaluation**. Judges operate independently for each turn, ensuring evaluation does not interfere with the ongoing debate dynamics.

3) Agent Behavior and Adaptation: In Approach A, agent adaptation relies on:

- **Initial Role Prompting:** Defining their stance and argumentative style.
- **Dialogical Interaction:** Responding directly to the opponent's points from the current round.
- **Retrospective Feedback:** Incorporating judge feedback received from the *previous* round's performance.
- **Prompt Guidance:** Specific instructions encourage direct refutation and reinforcement during rebuttal rounds (e.g., "Address your opponent's strongest points and reinforce your key arguments").

H. Approach B: Self-Improving Debate with Integrated Feedback

Approach B builds on the same two-agent debate structure over four rounds on the same topic but introduces an **iterative self-improvement cycle** within each round. The core idea is to make the agents "learn" from the judges' feedback in real time, effectively turning each round into a two-step process(Fig. 2): an initial attempt and a refined attempt.

- **Two-Step Argument Generation:** Each agent's turn within a round involves:
  1) *Initial Argument Generation:* The agent produces its argument as in Approach A.
  2) *Immediate Judge Feedback:* The multi-faceted judging system evaluates this initial argument immediately.
  3) *Feedback Integration:* This feedback is instantly provided back to the generating agent.
  4) *Argument Revision:* The agent revises its argument specifically to address the judge's critique.
  5) *Final Submission:* The revised ("improved") argument becomes the official contribution for that turn in the debate.
- **Iterative Refinement:** This feedback-and-revision loop occurs for both agents in every round, promoting continuous improvement throughout the debate.

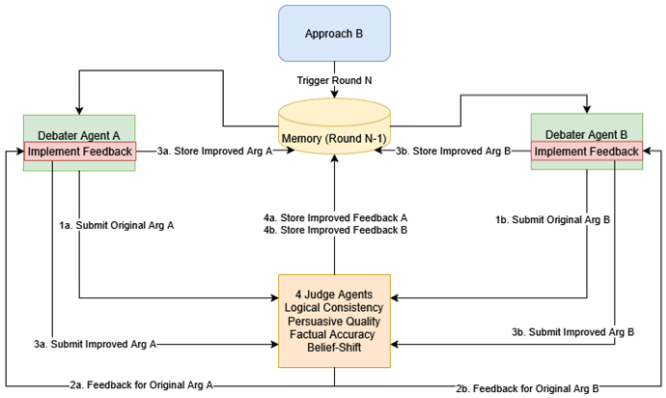

Fig. 2. Approach B Flowchart

1) Agent and Judge Roles in Refinement: The roles adapt slightly to accommodate the new loop:

- **Dual Judge Role:** The judge acts as both an **evaluator** (scoring initial and revised arguments on the same criteria as Approach A) and a **coach** (providing constructive feedback aimed at guiding improvement). The tone of feedback is slightly adjusted to be constructive, pointing out not just what was weak but also hints at how to improve. For instance, the judge might note that a debater's claim is logically sound but **lacks empirical evidence**, or that the rhetoric is passionate but **fails to acknowledge counter-arguments**.
- **Agent Adaptation:** Agents receive immediate constructive criticism and attempt to rectify weaknesses. They respond to the opponent's *final, improved* arguments from previous turns. Agents are instructed to refine presentation and reasoning without fundamentally altering their stance or introducing drastically new facts. Only the final improved arguments are stored in the main debate history visible to the opponent.

2) Mechanism for Online Learning: Approach B is designed to simulate an online learning process:

- **Immediate Application:** Agents apply feedback instantly, potentially reinforcing learning more effectively than delayed feedback.
- **Persistent Learning Potential:** Over multiple rounds, agents might internalize recurring critiques and preemptively address them, learning beyond single-turn edits.
- **Process vs. Model:** Using the same underlying LLM for both approaches ensures observed performance differences stem from the interaction framework, not the base model capability.
- **Computational Cost:** This approach roughly doubles the number of LLM calls per debate turn, increasing computational overhead but enabling targeted refinement.

TABLE II
DEBATE JUDGMENT SCORES BY MODEL AND JUDGE TYPE

Judge Model	Stance	Judge Type	deepseek_r1	gemma3	phi4	qwen2.5	yi
0	AGAINST	AVERAGE	6.4/10	6.6/10	6.1/10	7.1/10	3.3/10
1	AGAINST	Audience	7.5/10	7.0/10	7.0/10	7.5/10	0.0/10
2	AGAINST	Belief_shift	7.0/10	6.5/10	6.0/10	7.0/10	0.0/10
3	AGAINST	Ethical	7.0/10	6.0/10	6.0/10	7.0/10	0.0/10
4	AGAINST	Factual	4.0/10	6.0/10	5.0/10	6.0/10	6.0/10
5	AGAINST	Logical	7.0/10	6.0/10	5.0/10	7.0/10	8.0/10
6	AGAINST	Rhetorical	6.0/10	7.5/10	7.0/10	7.0/10	9.0/10
7	AGAINST	Strategic	6.0/10	7.0/10	7.0/10	8.0/10	0.0/10
8	FOR	AVERAGE	7.4/10	7.6/10	7.5/10	7.9/10	3.6/10
9	FOR	Audience	8.5/10	8.0/10	7.5/10	8.0/10	9.5/10
10	FOR	Belief_shift	8.0/10	7.0/10	7.0/10	8.0/10	0.0/10
11	FOR	Ethical	7.0/10	7.5/10	8.0/10	8.0/10	9.0/10
12	FOR	Factual	6.0/10	7.0/10	7.0/10	7.0/10	7.0/10
13	FOR	Logical	7.0/10	7.0/10	7.0/10	8.0/10	8.0/10
14	FOR	Rhetorical	8.0/10	8.5/10	7.0/10	8.0/10	9.0/10
15	FOR	Strategic	7.0/10	8.0/10	8.0/10	8.0/10	0.0/10

TABLE III
COMPARISON OF APPROACHES A AND B ACROSS ROUNDS

Approach	Round	Arg Type	Avg Logic	Avg Rhetoric	Avg Evidence	Avg Belief
Approach A	1		6.5	7.0	6.0	6.5
	2		6.8	7.3	6.2	6.9
	3		6.9	7.3	6.3	6.4
	4		6.8	7.5	6.3	6.8
Approach B	1	Original	6.4	6.8	6.3	6.5
		Improved	7.0	7.4	6.5	6.8
	2	Original	7.1	7.3	6.6	6.9
		Improved	7.0	7.5	6.7	6.8
	3	Original	7.0	7.2	6.9	6.6
		Improved	7.5	7.7	7.0	7.2
	4	Original	7.0	7.3	7.1	7.3
		Improved	7.6	7.8	7.2	7.4

IV. EXPERIMENTS

A. Semantic Evaluation Results.

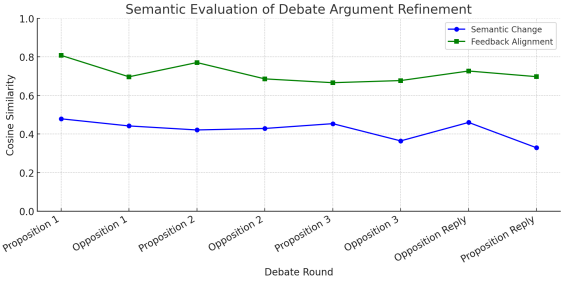

Fig. 3. Semantic evaluation metrics across debate rounds. Blue bars show Semantic Change (difference between original and revised speeches); green bars show Feedback Alignment (similarity between revisions and judge suggestions).

B. Evaluating Adversarial and Self-Evidence Check Approach

Table II presents results from our multi-agent debate experiments implementing adversarial critique and self-evidence verification mechanisms.

A notable finding emerged regarding model capability thresholds: the Yi-9B model repeatedly failed to complete evaluations when presented with judge prompts. When processing debate transcripts (15,000-18,000 tokens), it consistently deviated into irrelevant narratives about mythology rather than performing the requested argument assessment.

The performance patterns across different model sizes and architectures reveal systematic differences in argumentative capability, which we analyze in depth in the analysis section.

C. Comparative Performance For Approach - A, B

We conducted debates across four distinct topics (technology adoption, social media policy, climate regulation, education funding) under both Approach A and Approach B protocols. Argument quality was assessed after each turn using the multi-dimensional framework, yielding feedback and scores for Logic, Rhetoric, Evidence, and Belief-Shift (on a 1-10 scale).

Table III presents the average scores achieved per round under each approach. Approach A shows baseline performance with delayed feedback. Approach B results are broken down into the 'Original' argument score (before feedback) and the 'Improved' argument score (after revision based on immediate feedback). Each score is treated as an indicator of argument quality. The judges' textual feedback was analyzed to gain

qualitative insights into each argument's strengths and weaknesses. Since we do not produce a single explicit "winner" during the debate, we determine outcomes by comparing the cumulative or average scores of the two sides. For fairness, each debate topic was contested with the same initial prompt for both approaches, and the same judge configurations, so that scoring biases would be consistent across A vs B.

D. Debate Quality Analysis And Average Scores (Iterative Feedback Variants - A, B)

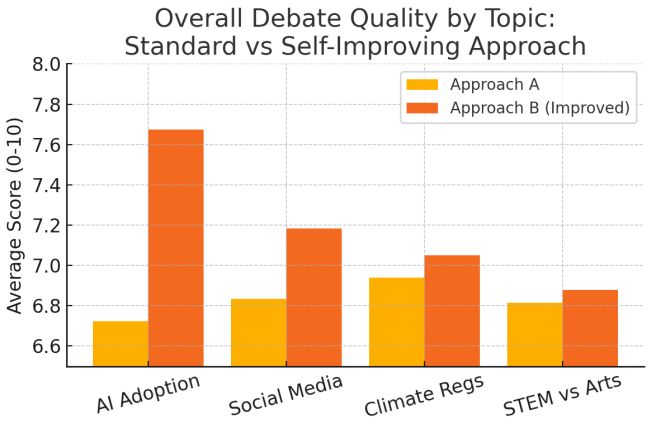

Fig. 4. Overall debate quality by topic, comparing Approach A (standard debate) and Approach B (self-improving). Bars show the average score (across all rounds and criteria) for each side combined. Approach B consistently achieves higher scores, with the largest gains on the AI adoption topic.

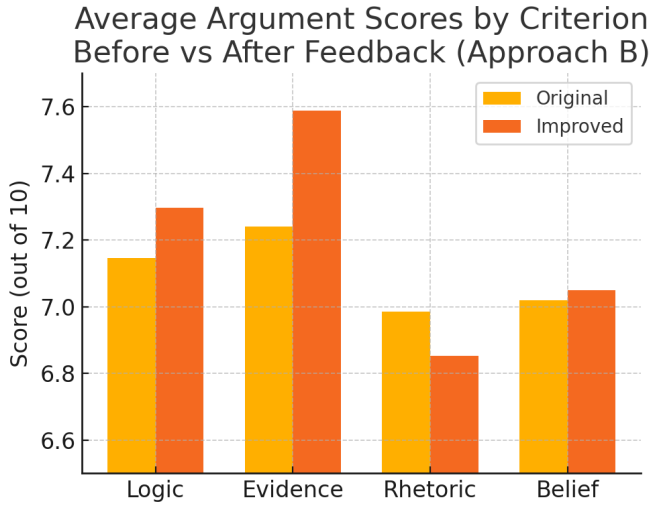

Fig. 5. Average argument scores by criterion before vs. after judge feedback in Approach B. The orange bars (improved arguments) show higher Logic, Evidence, and Belief scores than the initial arguments (yellow bars), while Rhetoric saw a slight decline on average.

V. Analysis

Our multi-agent debate framework produced several interesting findings regarding argument quality, evaluation patterns, and model capabilities. This section analyzes some key results.

A. Scoring Patterns and Evaluation Consistency

Our experimental conditions revealed consistent and meaningful patterns in debate evaluation scores. We observed several key dynamics that validate our evaluation approach:

Model capability significantly influenced argument quality and subsequent scores. Higher-capacity models (DeepSeek-R1 14B, Mixtral) consistently earned stronger evaluations (7.4–7.9/10) compared to smaller models (6.1–7.1/10). This pattern emerged regardless of stance, suggesting our judges detected genuine differences in reasoning capability rather than arbitrary variations.

Judge impartiality proved remarkably strong across our experiments. Even when evaluating its own arguments, models demonstrated fairness—Phi4 scored its own arguments in the 5–6.0 range while assigning 7–8.0 scores to more sophisticated responses from Mixtral and DeepSeek. This suggests LLM judges can maintain evaluative consistency even when assessing competitors.

We experimented with various judging configurations (single judge, 4–6 specialized judges, and dual participants) and found consistent scoring trends across these arrangements.

Capability thresholds became apparent during testing, as extremely weak models occasionally produced entirely irrelevant, broken, or null responses, earning scores of zero. This performance cliff demonstrates that while multi-agent debate frameworks can enhance reasoning, they require sufficiently capable base models.

Topic design proved critical to meaningful evaluation. **Asymmetrically winnable topics**—where one position is overwhelmingly stronger (e.g., "Is slavery justifiable today?")—skewed results dramatically, with winning positions earning 7.3–8.4 scores while opposing positions scored extremely low. Such topics violate the principles of clash and balance essential for productive debate, reducing evaluation to virtue signaling rather than logical reasoning.

B. Argument Quality Analysis

Our judge evaluations revealed several patterns in argument construction and persuasiveness:

1) Assumption Identification and Framing: Sophisticated debate responses demonstrated the ability **to identify unstated assumptions** in opposing arguments. For example, in a debate on humanities versus STEM funding, one model noted:

"The argument champions 'equal funding' without justifying why equal funding is the appropriate metric. It implicitly assumes that both fields generate comparable societal and economic returns..."

This critique demonstrates advanced meta–reasoning, identifying not just what an opponent argued, but the underlying premises upon which their argument depends. Such framing challenges proved to be to be more effective than surface-level rebuttals.

2) Evidence Quality and Causal Attribution: Judges consistently **rewarded specificity and causal reasoning** over correlation claims. A notable critique highlighted:

"The argument repeatedly associates humanities skills with positive outcomes (critical thinking, adaptability, job success) without demonstrating a causal link or explaining how those skills translate into concrete economic benefits."

This suggests that **LLMs can identify and evaluate the strength of causal chains** in arguments, a crucial capability for effective reasoning.

3) Counterargument Anticipation and Response: Higher scoring arguments demonstrated **anticipation of counterarguments**. Rather than merely defending their position, these responses preemptively addressed likely objections.

"The assertion that STEM education only drives economic growth is a demonstrably simplistic view... attributing all economic success to them ignores the significant contributions of the creative industries, which... contributed $2.25 trillion to the global economy in 2022."

This approach, accepting partial truths in opposing views while reframing with specific data, consistently earned higher evaluations from our judging system.

4) Citation and Disciplinary Context: Stronger models **cited named studies, meta-analyses, and publication years**, while weaker ones referenced vague or generic sources. Advanced models also contextualized arguments within disciplines (e.g., linking design thinking to anthropology), unlike smaller models, which lacked interdisciplinary framing.

C. Quantitative Results - Iterative Feedback Variants

Our experiments comparing the delayed feedback (Approach A) and integrated feedback (Approach B) debate frameworks yielded several key insights into argument quality and the impact of refinement mechanisms.

Initial experiments focused on establishing baseline performance using Approach A. Across the four debate topics, average combined scores for Approach A generally fell within the 6.7 to 6.9 range on a 10-point scale, indicating moderately competent argumentation. Some variance was observed depending on the topic; for instance, the climate regulation debate yielded slightly higher average scores ($\tilde{6}.94$), potentially due to the availability of concrete scientific and economic data for arguments. By contrast, the AI ethics and the funding priority debate had lower averages ($\tilde{6}.72$–6.82), possibly reflecting the broader scope and moral nuance that made it harder for the agents to provide very specific, evidence-rich arguments.

The introduction of the self-improvement mechanism in Approach B led to noticeable performance increases. As illustrated in Figure 4, Approach B consistently achieved higher overall average scores than Approach A on the same topics. The most significant improvement was observed in the AI adoption debate, where the average score increased from approximately 6.72 in Approach A to 7.68 in Approach B (a $\tilde{1}4\%$ relative gain). This suggests that the breadth of that topic benefited from iterative refinement, as the pro-AI debater was able to address concerns (like job displacement and ethical safeguards) that it initially glossed over. Other topics showed more modest, yet consistent, positive gains reflecting that some value-laden or less evidence-driven disputes may not benefit as much from the feedback loop

Isolating just the evaluation criterion, the improvements in Approach B (comparing initial vs. improved arguments) were most pronounced in the **Evidence** and **Logic** scores as shown in Fig. 5. Evidence score increased by about +0.35 after the agent incorporated the judge's feedback that the debater misses adding factual support or examples. In the social media debate, the Policy Advocate's improved opening argument included a reference to the WHO dubbing misinformation an "**infodemic**," lending authority to its claims about public health risks. Logical consistency scores improved by +0.15, suggesting that some internal contradictions were corrected. In the climate debate, after feedback, the Agent against strict regulations tempered an overly broad claim that regulations *always* harm the economy by acknowledging short-term costs but pointing to long-term innovation benefits, which the judge noted as a fix to a potential fallacy.

D. Qualitative Analysis - Iterative Feedback Variants

The judge feedback provides insight into how the arguments were perceived. Common strengths noted for high-scoring arguments included a clear logical structure and the use of specific evidence or examples. On the rhetoric side, agents often used emotional appeals or vivid imagery (e.g. describing a "future brimming with possibility" in the AI debate) – sometimes this was praised by the judge, but in a few cases overly grandiose language without substantiated facts led to a lower Evidence score.

Interestingly, the **Rhetoric** scores in Approach B showed a slight *decrease* on average after improvement (a drop of about 0.1 points). A possible explanation is that when agents focused on addressing logical and factual critiques, their language became more precise and technical, sometimes at the expense of emotional appeal or stylistic flair. We observed a few cases of this trade-off, an original argument might use a catchy metaphor or impassioned tone that scored well in rhetoric, but the judge's feedback pushed the agent to add data and qualifiers, making the improved argument more dry.

1) Real-time feedback loop: The real-time feedback loop produced more robust arguments by the end of each round. Many judge comments that were given as feedback were directly traceable in the revised arguments. In the AI debate, the judge told the **Pro-AI debater** that while their vision was optimistic, they had not addressed the audience's fears about AI. The improved argument was changed to: *"I understand the apprehension; the unknown often breeds fear. Yet, I urge you to look beyond the dystopian narratives…"*, explicitly acknowledging fears. This kind of adjustment made the argument more rhetorically balanced and likely boosted its persuasive impact.

2) Across-Round Trends: With Approach B, as the debate progressed, the compounding improvements yielded stronger and stronger performances. In Round 1, the average scores of

arguments with Approach B were only slightly higher than Approach A's, as the debaters had just begun refining their points. By Rounds 3 and 4, Approach B arguments were substantially ahead. The closing statements in Approach B were some of the highest-scoring of all – agents by then had eliminated nearly all factual errors or logical gaps. In one of the debates in Approach A, the closing statements scores dipped, possibly because of agent stagnation. Another interesting dynamic was that the feedback loop sometimes altered the **balance of the debate**. In Approach A, if one side missed a crucial rebuttal, that lapse could swing the momentum to the opponent. In Approach B, the judge feedback would immediately flag such a lapse, giving the side a chance to recover in the improved argument.

3) Performance Variations by Topic: The efficacy of the self-improvement approach did vary with the nature of the topic. Topics rich in empirical data (AI's impact, climate policy) saw larger benefits, as there was clear guidance on what facts or examples to add. More philosophical debates (like funding priorities) saw smaller gains, as improvements related to acknowledging nuances rather than adding new evidence. This suggests that the utility of this self-improvement method may be context-dependent. We also note that in Approach B, the relative "winner" of the debate (higher final scores) sometimes flipped compared to Approach A. For instance, in the AI adoption debate under Approach A, the cautious (con) side had higher scores, but under Approach B the pro-AI side's improvements caused it to score higher overall.

E. Patterns in Evaluation Outcomes (Overall)

Multi-agent judging revealed consistent scoring advantages when debates were structured as adversarial engagements. Models subjected to these constraints produced more evidence-driven arguments with explicit citations and clearer logical progression.

VI. Conclusion

Our multi-agent debate framework demonstrates that structured adversarial interactions significantly enhance LLM reasoning capabilities. Across multiple experimental conditions, debates with evidence verification and critique mechanisms produced arguments with measurably higher logical consistency, factual accuracy, and persuasive strength compared to single-agent approaches. The most significant improvements appeared in evidence quality (+0.35 points) and logical consistency (+0.15 points), with real-time feedback systems outperforming delayed feedback.

Perhaps most striking was our discovery that debate quality depends not merely on model scale, but on the architecture of interaction. Even smaller models produced remarkably strong arguments when placed in environments that encouraged iterative improvement.

These findings suggest that debate frameworks represent **a viable alternative to parameter scaling** for improving LLM reasoning. The SocrAItic Circle work we did, not only advances AI capabilities but opens new possibilities for educational applications where students might engage with AI-mediated debates to develop their own critical thinking skills.

AI can move beyond superficial language generation toward more structured forms of reasoning—not by mimicking human thought, but by engaging in the dialectical processes that have driven human intellectual progress for millennia.

VII. Acknowledgment

We used **Ollama**, a local AI model management system, for orchestrating and evaluating smaller open-source models on local hardware.

For experiments requiring 70B or larger models, we utilized hosted inference through paid API credits provided via **Perplexity** and **Google**, enabling access to high-capacity commercial models.

Implementation of nested data handling and model output management in Python leveraged publicly available guides on JSON serialization [10], binary encoding [11], and modular object-oriented client-server architecture [12], [13].

VIII. Repository Links

A. Iterative Feedback Variants

You can find the implementation for Iterative Feedback (Approach A & B) variants and the raw results in
https://github.com/khgaurav/SocrAItic_Circle

Repository owned & maintained by Gaurav Kothamachu Harish.

B. Adversarial Critique & Self-Evidence Check Variant

The full implementation of the Critique & Self-Evidence variant—including source code, supplementary materials (e.g., README, environment files), and evaluation resources—is publicly available at:
https://github.com/mjsushanth/Multi_Agent_LLM_Debater

Repository owned & maintained by Joel Markapudi.

IX. Credit Assessment

- **Joel Markapudi**: Architecture design, Python implementation of multi-agent debate framework, YAML-based prompt management system, Debater Agent Prompts (structured adversarial critique, evidence self-verification mechanisms, and development of all seven specialized judge types with corresponding evaluation criteria), single/multi-judge evaluation pipeline, cognitive psychology and argumentation theory research, experimental testing and results analysis, authorship of relevant report sections
- **Gaurav Kothamachu Harish**: Design and Python implementation of Iterative Feedback Variants (Approach A & B); Agent principles design and evaluation; integration of real-time judge feedback loop; implementation and logging of scoring and evaluation pipelines; comparative performance analysis across varying agent and judge strengths; qualitative and quantitative analysis of judge feedback and argument improvements
- **Tanay Grover**: Explored integration of human feedback using preference modeling and basic reward signal tuning in LLM-driven multi-agent systems. Investigated alignment strategies such as supervised fine-tuning and prompt-based conditioning to guide agent responses.
- **Pengkun Ma**: Design and Coding Alternative Evaluations (Semantic Change, Feedback Alignment). Researching Sentence-BERT, Readability Analysis. Investigating NLP lexical techniques.
- **Easha Meher Koppisetty**: Implemented a 4-round World Schools Debate format using role-specific prompt templates for Proposition, Opposition, Rebuttal, and Summary rounds. Simulated debates across diverse topics and evaluated agent and judge outputs using metrics.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] L. Zhang, Y. Zhu, G. Wang, L. Zheng, X. Lin, Y. Lin, S.-W. Liu, and Y. Zhang, "CoEvol: Constructing Better Responses for Instruction Finetuning through Multi-Agent Cooperation," *arXiv preprint arXiv:2305.12524*, 2023.

[3] T. Liang *et al.*, "Multi-Agent Debate," GitHub, 2023. Available: https://github.com/Skytliang/Multi-Agents-Debate

[4] R. Orji, R. L. Mandryk, and J. Vassileva, "Improving the Efficacy of Games for Change Using Personalization Models," *ACM Transactions on Computer-Human Interaction*, vol. 22, no. 1, pp. 1–29, 2015.

[5] T. J. M. Bench-Capon and P. E. Dunne, "Argumentation in Artificial Intelligence," *Artificial Intelligence*, vol. 171, no. 10-15, pp. 619–641, 2007.

[6] R. Duthie and K. Budzynska, "A Deep Modular RNN Approach for Ethos Mining," *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4041–4047, 2018.

[7] H. Oinas-Kukkonen and M. Harjumaa, "Persuasive Systems Design: Key Issues, Process Model, and System Features," *Communications of the Association for Information Systems*, vol. 24, article 28, pp. 485–500, 2009.

[8] S. E. Toulmin, "The Uses of Argument," *Cambridge University Press*, Updated edition, 2003.

[9] R. E. Petty and J. T. Cacioppo, "The Elaboration Likelihood Model of Persuasion," *Advances in Experimental Social Psychology*, vol. 19, pp. 123–205, 1986.

[10] Bradley, J. (2020). *Working With JSON Data in Python*. Real Python. Available at: https://realpython.com/python-json/

[11] Kenneth Reitz et al. *Data Serialization*. The Hitchhiker's Guide to Python. Available at: https://docs.python-guide.org/scenarios/serialization/

[12] DigitalOcean Community. (2021). *Python Socket Programming: Server and Client Example Guide*. Available at: https://www.digitalocean.com/community/tutorials/python-socket-programming-server-client

[13] Sturtz, G. (2022). *Socket Programming in Python (Guide)*. Real Python. Available at: https://realpython.com/python-sockets/