Conversational Agent Debating Capability

Mohamad Faraj Makkawi
IMT Atlantique
Rennes, France
mohamad.makkawi@imt-atlantique.net

Hassan Khan
IMT Atlantique
Rennes, France
hassan.khan@imt-atlantique.net

***Abstract*—**

***Index Terms*—Debate Systems, Argumentation, Conversational Agents, Argumentation Service Platform with Integrated Components (ASPIC), Retrieval-Augmented Generation (RAG).**

I. Introduction

Dialogue systems capable of debating at a level of sophistication are a growing area of research in Artificial Intelligence. Not only do these systems need to understand and generate arguments, but they must stick to principles of logic and persuasion, the same as advanced human interactions. Creating arguing agents that can present justified positions, decide controversies, and rebuttal counterarguments in a logical manner is a significant technical challenge with tremendous potential for application ranging from educational software to sophisticated decision support systems. Several approaches have been formulated to meet this challenge, each with its own merits and weaknesses.

One of the earliest approaches is that of Rule-based agents. These systems function on the basis of pre-established templates and logical rules picked from expert knowledge. Although they provide predictability and control in precise, limited domains with the assurance of consistency and sound reasoning, their use of fixed structures tends to yield inflexible responses and makes them less adaptable in dynamic or unexpected debate situations [**b12**].

Alternatively, Retrieval-based agents, sometimes combined with techniques like Retrieval-Augmented Generation (RAG), directly retrieve arguments from massive databases or text corpora [3]. Using vast amounts of existing knowledge, this allows for fast generation of contextually diverse responses. However, ensuring the factual correctness, contextual appropriateness, and logical consistency of the generated or retrieved arguments remains a significant challenge for such systems.

Hybrid argumentation agents seek to combine clear logical rules with adaptable responses learned from data. By typically integrating retrieval mechanisms and explicit reasoning mechanisms (e.g. rule-based logic or knowledge graphs), these agents aim to produce more flexible, relevant, and logically coherent arguments than purely rule-based or retrieval-based systems alone can [3].

Further complexity is added by Hierarchical persuasion agents. These systems aim to model the interaction of argumentation itself, perhaps constructing arguments step-by-step based on dialogue context or feedback from the user. This is to maximize the overall quality of the interaction and the persuasive effect over time by dynamic adjustment of the argumentative strategy [5].

Finally, Explainable debate agents prioritize transparency in their argumentative process [**b13**]. By making their claims and reasoning steps explicitly justified, these agents aim to enhance user trust along with facilitating the ease of inspection of the agent's logical validity, which is crucial in critical applications.

Although many methodologies are facilitating AI debate capabilities, this study investigates the specific contributions of structured argumentation theory-based and symbolic reasoning-based methods to make the agent able to conduct a good debate. We placed a particular emphasis on evaluating how frameworks like Argument Graphs, used because of their ability in the intuitive visualization of complex inter-argument relationships (support/attack), the Argumentation Structure for Preference, Inference, and Choice (ASPIC) framework, used because of the comprehensiveness with which it defines formally structured and logically sound arguments, and Multi-Attribute Argumentation Frameworks, used because of their ability in enabling multi-perspective argument comparison based on different dimensions, can be used to improve agent-based argumentation's robustness and comprehensibility. We explain how these techniques facilitate structured reasoning and knowledge representation, highlighting in particular their advantages of explainability and logical consistency [**b11**]. Meanwhile, we acknowledge and resolve inherent challenges for such symbolically-biased systems, such as managing uncertainty and dynamically revising reasoning protocols in real-time debate contexts [**b11**].

This paper will initially develop a brief definition of debate and effective argumentation criteria. We then explore leading computational approaches (Argument Graphs, ASPIC, and Multi-Attribute frameworks) before comparatively evaluating a number of agent architectures for debate. The approach further demonstrates that symbolic reasoning is critical for achieving explainability and maintaining logical soundness. We subsequently present the general limitations and challenges for the development of proficient AI debaters, setting the stage for concluding observations.

II. Debate Definition and What Makes a Good Debate

Debate is a formal process in which two or more debaters take turns presenting systematic arguments and counterarguments on some prescribed subject, under established rules and time constraints, with the objective of advancing, defending, and challenging positions for judgment by an adjudicator or audience members [4][6]. It can also be viewed as a game of arguments where the exchanges are governed by pre-established rules, such as the number and function of the participants, strict alternation, timing and sequence of the speeches, and the requirement to rebut opposite arguments in a pre-established format [2]. A good debate consists of empirical evidence-based arguments that are well-supported, clear explanations that spell out reasoning using simple language, and an awareness of counterarguments to back persuasive attempts. Moreover, logical coherence between premises and the categorization of argument quality - that is, how well an argument is supported, clear, relevant, and logically sound - are elements that make responses effective as a whole [2].

In order to evaluate debate performance well, judges look for ordered argumentation with an organized structure, logical flow, and successful refutations. The quality of evidence is extremely pertinent and should be based on actual facts, credible sources, and firm resolutions. Teams must demonstrate prioritization of argumentation and insight into the most significant "point of clash" [**b8**]. Judges are also responsible for recognizing fabricated evidence, ad hominem attacks, or rule violations. Where qualitative evaluation focuses on the strength of the argument [**b7**][**b8**].

III. Conversational Agent Debating Techniques and Applications

- **Argument graphs:** A visual representations of interrelations between arguments, which can be formally represented as a directed graph G=(V,E), where V is the vertices (arguments) and E is the directed edges representing support and attack relations. The visualization through such graphical representations can make complicated interrelations between arguments easy to understand, making them understandable by audience members in a more readable way. Though argument graphs efficiently promote understanding and involvement, their persuasiveness relies on the quality and strength of the arguments. Therefore, they get a positive assessment for their ability to provide clarity in persuasive conversations [1][2].
- **ASPIC framework:** The ASPIC+ framework provides a formal and structured format for argument presentation and evaluation, particularly valuable in contexts like debate where logical rigor is essential. At its core, ASPIC+ specifies how arguments can be systematically constructed by applying inference rules to a base of premises or accepted facts to derive claims. These rules are typically divided into two types: strict rules, which represent deductive, logically necessary inferences (e.g., 'if X is a square, then X is a rectangle'), ensuring the conclusion must hold if the premises do; and defeasible rules, which capture plausible, common-sense, or presumptive reasoning where the conclusion typically holds but may be defeated by exceptions or contrary evidence (e.g., 'birds typically fly'). Arguments are thus chains linking premises to a final claim via sequences of these strict and/or defeasible rules. ASPIC+ then explicitly defines various ways arguments can conflict through different types of 'attack' relations, such as challenging a premise used in an argument or rebutting a conclusion derived via a defeasible rule. Crucially, preferences between arguments are often used to determine which attacks are successful, leading to a formal assessment of which arguments are ultimately justified. With this formal setup, revealing the step-by-step application of specific rules, there comes the potential for unambiguous logical demonstration. While the full network of interacting arguments and attacks typically forms a complex argument graph, the structured nature allows specific lines of reasoning supporting a claim to be clearly traced. This explicit structure, revealing the logical connections and dependencies, helps 'fortify' an argument for persuasion by making its reasoning transparent to an audience for evaluation. Evaluative perspective shows that ASPIC+'s strong logical structure is well-suited to enable sound arguments, thereby aiding in the promotion of reasoned persuasion [2].
- **Multi-Attribute Argumentation Framework:** Ranks arguments on multiple criteria, such criteria being measurable through the formula:

$$S(A) = \sum_{i=1}^{n} w_i \cdot v_i$$

where w_i are weights for each criterion and v_i are the respective scores for each criterion. The multi-aspect analysis process allows for sensitive rankings that are capable of communicating directly with the values of the audience, optimally maximizing the persuasive power of the framework. This approach is widely valued for its ability to accommodate heterogeneous preferences, enhancing the persuasiveness of arguments. [2].

IV. Comparative Analysis and Evaluation of Debate Agents

To facilitate an easy understanding of the landscape of debate agents, this paper has two significant tables. Table I has a simplified overview of different types of conversational agents utilized in debate use cases, including their approaches, strengths, and weaknesses. Table II takes it further by listing a comparison of evaluation metrics for these agents in tabular form. These tables are designed to be an easy reader reference, enabling us to make useful comparisons and contrasts between different agent architectures and assessment criteria in the field.

V. Symbolic Methods: Explainability and Validity in Logical Reasoning

Symbolic techniques emphasize representation of knowledge as ordered, legible symbols derived from sound logical

TABLE I
OVERVIEW OF DIFFERENT CONVERSATIONAL AGENT TYPES FOR DEBATE APPLICATIONS

Agent Type	**Approach**	**Use Cases**	**Strengths**	**Weaknesses**
Rule-Based	Predefined logical templates	Compliance systems [**b12**]	Predictable outcomes	Rigid structure
Retrieval-Based	DB queries + RAG	Customer support chatbots [4]	Dynamic data handling	Retrieval dependency
Hybrid (Rule-Retrieval Based)	Rule-retrieval integration	Legal argumentation [3]	Context-aware reasoning	Complex setup
Hierarchical	Feedback-driven adaptation [5]	Policy negotiation	Personalized arguments	Feedback latency
Explainable	Justification chains [**b13**]	Diagnostics, Legal systems	Transparent explanations	Depth limitations

TABLE II
EVALUATION METRICS FOR CONVERSATIONAL AGENTS IN DEBATE APPLICATIONS

Evaluation criteria	**Rule-Based**	**Retrieval**	**Hybrid**[**b19**]	**Hierarchical** [**b18**]	**Explainable** [**b14**] [**b17**]
Accuracy	High in narrow domains [**b14**]	Moderate (depends on data quality) [**b15**]	Balanced across tasks	Varies by hierarchy level	Moderate (prioritizes transparency over optimization)
Response Time [**b16**]	Fast (deterministic)	Moderate (retrieval latency)	Slow (multi-layer processing)	Slowest (sequential decision-making)	Moderate (additional explainability overhead)
Cost Efficiency	Low (static rules)	Moderate (retrieval infrastructure)	High (compute-intensive)	High (maintenance complexity)	Moderate (explanation generation costs)
Scalability	Poor (rule complexity)	High (modular data updates)	Moderate (resource-intensive)	Best (layered abstraction)	Limited (Model Size Complexity)
Stability	High (deterministic)	Moderate (data drift sensitivity)	High (adaptive tuning)	High (isolated failures)	Moderate (explanation consistency)
User Satisfaction	Low (inflexible)	Moderate (context-aware)	High (balanced UX)	Moderate (learning curve)	Highest (trust through clarity)

reasoning to argue about a problem. They leverage formal logic, systems of rules, and graphs of knowledge. Their greatest strength is explainable reasoning, in order to substantiate an argument or decision [**b11**].

To evaluate reasoning in symbolic methods, we ensure that each step logically follows earlier premises without contradiction, usually through formal entailment tests or truth tables of propositional logic. A formal test of entailment works by checking if a conclusion is formally derivable from an argument set of premises based on formal laws of logic. This is most commonly done by means of the employment of application of truth table or by examination of the rules of inference in order to determine if in all cases where premises are valid, the conclusion as well shall have to be true. When this is to be true, entailment is true, making the argument valid from being contradictory in nature but, alternatively, being valid. In richer systems, soundness is achieved by analyzing streams of reasoning which is a sequence of logically connected steps or conclusions that follow each other to reach a conclusion. It is defined formally as a sequence of statements or conclusions, each supported by antecedent ones and founded on accepted rules, forming an intelligible and traceable chain from premises to conclusion., comparing them with symbolic ground truths (for example, mathematical theorems or logical derivations), and cross-verifying against recognized rules [**b20**]. This explainability enhances user trust in the system and its usability [**b11**].

VI. LIMITATIONS OF CONVERSATIONAL DEBATING AGENTS

- **Uncertainty and Big Data Management:** AI Debate Systems may need to handle vast amounts of data from multiple sources, often containing noisy or inconsistent information [**b11**].
- **Knowledge Acquisition Bottleneck:** Symbolic AI programs typically require hand-coded rules and knowledge, leading to high human effort and cost in translating real-world problems into system inputs [**b11**].
- **Real-time Dynamic Assessments:** AI debate systems face challenges in reacting dynamically to emerging arguments or rapidly changing contexts in real-time scenarios [**b11**].
- **Logical Inconsistencies:** Internal logical inconsistency is a major problem in designing secure AI systems, such as debate systems. The problem highlights the need to incorporate constrained logical verification methods—methods that prove AI reasoning against predetermined logical constraints or rules—or consistency inference tools as part of the AI system architecture [**b10**].

VII. DISCUSSION

The development of conversational agents capable of conducting sophisticated debate is one of the major areas of AI research, involving not only language understanding and production but also the adherence to logic and principles of persuasion [4]. This paper explored a number of approaches, highlighting the relative advantages and disadvantages of rule-based, retrieval-based, hybrid, hierarchical, and explainable agents (Table I). Hybrid and retrieval-based systems, although offering flexibility and access to large data sources [3], it is challenging to ensure logical correctness and factual accuracy. Rule-based systems are predictable and ensure logical consistency for a small domain but are inflexible [**b12**]. Our work placed particular emphasis on the function of formal argumentation structures and symbolic reasoning. Techniques like Argument Graphs offer good visualization alternatives, enhancing users' comprehensibility of complex argument relations [1][2]. ASPIC enables the formal definition of logically valid arguments necessary to build convincing and verifiable chains of arguments [2]. Furthermore, Multi-Attribute Argumentation Frameworks facilitate advanced assessment of arguments against multiple attributes and can augment persuasive power by appealing to values in the audience [6]. Such systematic approaches address important ingredients of a "good debate," such as organization, logical consistency, and adequately supported reasoning [**b7**][**b8**]. The comparison analysis (Table II) also brings out the inherent trade-offs. For instance, formal structure-based methods (e.g., rule-based or augmented hybrid/explainable systems) can be very precise in usage and very explainable, and thus achieve user trust, but very poor in response time, scalability, or adapt-ability compared to pure retrieval-based systems [**b15**]. Symbolic methods, in particular, are the foundation of explainability and logical soundness necessary for user trust and system verification. The capacity to trace out steps of reasoning, as represented by symbolic methods and explicit chains of justification, is to be compared with the typically black-box character of data models alone. However, there remain important weaknesses in AI debate agents. Dealing with uncertainty and large amounts of potentially noisy data is a major issue. Knowledge acquisition bottleneck, especially for symbol-based systems and rule-based programs hand-coded, remains costly and time-consuming. Additionally, real-time debate's dynamic nature demands agents with the capacity to judge and respond in real time, which is inappropriate for many modern architectures [**b11**]. Internal logical consistency is required for system integrity and to avoid irrelevant argumentation in the integration of heterogeneous information or reasoning protocols [**b10**]. Eliminating these weaknesses is essential to future research on strong and effective AI debaters.

VIII. Conclusion

In this paper, we went through the skills required for conversational agents to engage in a good debate, pointing to the potential of structured theories of argumentation and symbolic reasoning. We covered the foundational concepts defining a good debate [4] [6] [**b8**] and outlined the key computational approaches, including Argument Graphs [1], the ASPIC framework, and Multi-Attribute Argumentation Frameworks [2]. These approaches are of significant advantages in promoting logical consistency, presenting arguments in an orderly fashion, enabling explainability, and enabling multi-faceted argument evaluation, thereby giving rise to more robust and persuasive AI debaters. Comparative analysis highlighted the trade-offs between different agent architectures, emphasizing the strengths of symbolic approaches in ensuring logical validity and transparency [**b11**]. However, there remain significant challenges, in particular regarding data management, knowledge acquisition, real-time responsiveness, and logical consistency maintenance [**b10**] [**b11**]. Lastly, the integration of the strengths of symbolic argumentation and structured reasoning can potentially produce AI agents that can not only participate in debates but do so in a manner that is logical, coherent, and consistent. Research in the future must tackle the limitations recognized, perhaps in new hybrid architectures compromising formal rigour with data-driven plasticity, in order to be in a position to maximize the strengths of AI for high-level argumentative discourse.

Acknowledgment

We would like to thank John Puentes and Mihai Andries, Christophe Lohr for their invaluable supervision and guidance throughout this research.

References

[1] Lisa A Chalaguine et al. "A persuasive chatbot using a crowd-sourced argument graph and concerns". *Computational Models of Argument*. IOS Press, 2020, pp. 9–20.

[2] Débora Engelmann et al. "Argumentation as a method for explainable AI: A systematic literature review". *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE. 2022, pp. 1–6.

[3] Iryna Kulatska. "Arguebot: Enabling debates through a hybrid retrieval-generation-based chatbot". MA thesis. University of Twente, 2019.

[4] Geetanjali Rakshit et al. "Debbie, the debate bot of the future". *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*. Springer. 2019, pp. 45–52.

[5] Kazuki Sakai et al. "Hierarchical argumentation structure for persuasive argumentative dialogue generation". *IEICE TRANSACTIONS on Information and Systems* 103.2 (2020), pp. 424–434.

[6] Chenhao Tan et al. "Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions". *Proceedings of the 25th international conference on world wide web*. 2016, pp. 613–624.

References

[1] Debating for Everyone. "How to Judge a Debate." Available: https://www.debatingforeveryone.com/resources/how-debating-works/how-to-judge-a-debate. [Accessed: Apr. 3, 2025].

[2] Shuster, Kate, and Meany, John. "Judging Debates: The Middle School Public Debate Program Judge Certification Manual." Available: https://esu.fcny.org/esu/programs/middle_school_debate/educators/lesson_plans_teaching_materials/MSPDP_judging_manual:en-us.pdf. [Accessed: Apr. 3, 2025].

[3] Vassiliades, A., Bassiliades, N., and Patkos, T. "Argumentation and explainable artificial intelligence: A survey," *Knowledge Engineering Review*, vol. 36, e5, 1-35, 2021. doi:10.1017/S0269888921000011. [Online]. Available: https://doi.org/10.1017/S0269888921000011. [Accessed: Mar. 23, 2025].

[4] Ilkou, E., and Koutraki, M. "Symbolic vs Sub-symbolic AI Methods: Friends or Enemies?" In *Proceedings of the CIKM 2020 Workshops*, CEUR Workshop Proceedings, ISSN 1613-0073, October 19-20, 2020, Galway, Ireland. [Online]. Available: http://ceur-ws.org. [Accessed: Mar. 23, 2025].

[5] Kasif, Simon. "A Trilogy of AI Safety Frameworks: Paths from Facts and Knowledge Gaps to Reliable Predictions and New Knowledge." *Department of Biomedical Engineering, Program in Bioinformatics, Department of Computer Science, Boston University, 2024.*

[6] Monte-Alto, H. H. L. C., Possebom, A. T., Morveli-Espinoza, M. M. M., and Tacla, C. A. "A rule-based argumentation framework for distributed contextual reasoning in dynamic environments." *DYNA*, vol. 88, no. 217, pp. 120-130, 2021. doi:10.15446/dyna.v88n217.90858.

[7] Ali, B., Pawar, S., Palshikar, G. K., & Singh, R. (2022). Constructing a dataset of support and attack relations in legal arguments in court judgements using linguistic rules. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pp. 491-500. Marseille, 20-25 June 2022. European Language Resources Association (ELRA). Retrieved from: https://www.elra.info.

[8] IBM. "AI Agent Evaluation." Available: https://www.ibm.com/think/topics/ai-agent-evaluation. [Accessed: Apr. 3, 2025].

[9] Smythos. ”AI Agent Performance Measurement.” Available: https://smythos.com/ai-agents/agent-architectures/ai-agent-performance-measurement/. [Accessed: Apr. 3, 2025].

[10] Galileo. ”AI Agent Metrics.” Available: https://www.galileo.ai/blog/ai-agent-metrics. [Accessed: Apr. 3, 2025].

[11] SuperAnnotate. ”AI Agent Evaluation.” Available: https://www.superannotate.com/blog/ai-agent-evaluation. [Accessed: Apr. 3, 2025].

[12] Galileo. ”Evaluating AI Agent Performance: Benchmarks for Real-World Tasks.” Available: https://www.galileo.ai/blog/evaluating-ai-agent-performance-benchmarks-real-world-tasks. [Accessed: Apr. 3, 2025].

[13] Aisera. ”AI Agent Evaluation.” Available: https://aisera.com/blog/ai-agent-evaluation/. [Accessed: Apr. 3, 2025].

[14] Lee, Jinu, and Hockenmaier, Julia. ”Evaluating Step-by-step Reasoning Traces: A Survey.” Available: https://arxiv.org/html/2502.12289v1. [Accessed: Apr. 3, 2025].