AETHER: An Explainable Multi-Agent Debate Architecture for Decision Support on Structured Documents

Ganesh Gupta
Dept. of AI&DS TCET
Mumbai, India
Roll No: 20
ganeshgupta05@gmail.com

Jatin Gupta
Dept. of AI&DS TCET
Mumbai, India
Roll No: 21
jatingupta07@gmail.com

Raj Joshi
Dept. of AI&DS TCET
Mumbai, India
Roll No: 34
jraj1069@gmail.com

***Abstract*—As Large Language Models (LLMs) are increasingly integrated into critical decision-making workflows, the "black-box" nature of their reasoning presents significant challenges regarding trust, interpretability, and hallucination. Traditional single-agent architectures often yield opaque conclusions without exposing the dialectical process required to reach them. This paper introduces Project AETHER, a deliberative multi-agent framework designed to analyze structured documents through adversarial reasoning. The proposed architecture decomposes complex reports into constituent semantic factors and orchestrates a structured debate between opposing agents—an Affirmative Agent (Pro) and a Dissenting Agent (Con)—before a Synthesizer Agent derives a final consensus. By enforcing explicit agent-to-agent interaction and exposing the debate trace, AETHER provides a transparent "reasoning trail" rather than a mere final output. We demonstrate the system's efficacy in domains such as organizational analysis and policy evaluation, highlighting how multi-agent debate reduces confirmation bias and enhances the interpretability of automated decision support systems.**

***Index Terms*—Multi-Agent Systems (MAS), Explainable AI (XAI), Large Language Models, Dialectical Reasoning, Decision Support, Automated Debate.**

I. Introduction

The automation of document analysis has long been a holy grail of enterprise intelligence. In the modern corporate ecosystem, stakeholders are inundated with vast quantities of structured and semi-structured data—ranging from quarterly financial audits and sales performance reports to complex policy documents and compliance frameworks. As organizations increasingly rely on Artificial Intelligence (AI) to synthesize this information, the cost of errors increases proportionately. A misinterpretation of a risk factor in a financial report or an oversight in a compliance audit can lead to significant strategic failures.

While modern Large Language Models (LLMs) such as GPT-4 and Gemini demonstrate impressive capabilities in summarization and information extraction, they inherently suffer from critical limitations when applied to high-stakes decision support. Primary among these is the lack of self-correction mechanisms. A standard LLM, when operating in a "single-agent" paradigm, generates tokens based on probabilistic likelihood rather than logical rigor. This often results in "sycophancy"—a phenomenon where the model tends to agree with the user's implicit biases or settles on the most generic, non-controversial path, rather than critically analyzing the input for contradictions.

Furthermore, current implementations often fail to provide transparency. When a single model is asked to "analyze risks," it produces a smoothed-over narrative that hides potential pitfalls and the reasoning process used to identify them. It lacks the adversarial pressure required to uncover edge cases, logical inconsistencies, or subtle data discrepancies that a human auditor would naturally flag. This "black-box" reasoning is insufficient for domains requiring high accountability.

To address this, we present **Project AETHER** (Automated Explainable Theoretic Heuristic Reasoning), a novel architecture that mimics human boardroom deliberation. Instead of a single entity generating a response, AETHER employs a "Society of Agents" approach where conclusions are forged through conflict. By simulating a dialectical debate between opposing viewpoints, the system forces a deeper exploration of the input data.

The contributions of this paper are threefold:

- We propose a modular **Factor-Debate-Synthesis pipeline** that systematically decomposes unstructured inputs into debate-ready claims, ensuring no critical data point is overlooked.
- We implement a **Dialectical Coordination Protocol** where agents explicitly challenge and defend arguments in a round-robin format, refining the context window dynamically to reduce hallucination.
- We introduce an **Interpretability Layer** that exposes the raw debate logs to the user, allowing human decision-makers to audit *why* a decision was reached, effectively shifting the paradigm from "AI-generated answers" to "AI-assisted verification."

II. Literature Survey

The evolution of automated reasoning has shifted from static rule-based systems to dynamic generative models. However, interpretability remains a bottleneck.

A. Chain-of-Thought vs. Society of Mind

Wei et al. (2022) demonstrated that "Chain-of-Thought" (CoT) prompting significantly improves reasoning capabilities in LLMs by encouraging the model to generate intermediate steps. However, CoT is still a solitary process prone to cascading errors—if the first step in the chain is flawed, the entire subsequent reasoning collapses. Minsky's "Society of Mind" theory posits that intelligence emerges from the interaction of smaller, simpler processes. Recent works in Multi-Agent Systems (MAS), such as AutoGen and MetaGPT, have operationalized this by assigning distinct roles to LLMs, allowing for specialized processing.

B. Debate as a Truth-Seeking Mechanism

Research by Irving et al. (2018) and Du et al. (2023) suggests that multi-agent debate improves factual accuracy and reasoning robustness. By forcing one agent to critique the output of another, the system effectively performs "self-verification." AETHER builds upon this theoretical foundation but adapts it specifically for business contexts. Unlike open-ended chat debates, AETHER structures the debate as a rigid Support-Oppose-Synthesize workflow tailored for structured business documents, ensuring that the output remains focused and actionable.

III. Methodology (System Architecture)

The AETHER architecture is designed as a directed acyclic graph (DAG) of specialized agents, each governed by a specific persona and constraint set. The workflow follows a rigorous four-stage pipeline: Ingestion, Factorization, Deliberation, and Synthesis.

A. Stage 1: Ingestion and Semantic Parsing

The pipeline begins with the ingestion of the raw structured document (e.g., PDF reports, Excel financial sheets, or CSV data). Before reasoning can occur, the data must be normalized.

- **Text Chunking:** Large documents are segmented into semantic blocks. Unlike standard character-limit chunking, AETHER uses a recursive retrieval strategy to keep paragraph context intact.
- **Table Serialization:** Tabular data is converted into natural language narratives. For instance, a row in a sales spreadsheet '[Q3, APAC, +15%]' is serialized to "In Q3, the APAC region recorded a 15% growth," allowing LLM agents to process it as evidence.

B. Stage 2: Semantic Factor Extraction

The entry point of the reasoning engine is the **Factor Agent**. Unlike standard summarizers which compress information, this agent utilizes a schema-extraction prompt to explode the document into individual, actionable "claims" or "variables."

Input: Raw Document Context (e.g., "Q3 Sales Report").
Process: The agent scans for assertion-heavy segments, looking for causal statements, predictions, or strong evaluative claims.
Output: A list of debate topics $T = \{t_1, t_2, ..., t_n\}$.

Example: If the document states, "Marketing strategy was the primary driver of growth," the Factor Agent extracts topic t_1: "The marketing strategy in the APAC region was the primary driver of 15% growth." This isolates the claim for scrutiny.

C. Stage 3: The Dialectical Debate Engine

For every extracted topic t_i, the coordination layer spawns a dedicated debate instance. This is the core of the AETHER system, involving two adversarial agents sharing a context window.

1) The Affirmative Agent (Pro): This agent is prompted to act as a proponent. It scans the provided document specifically for evidence E_{pro} that supports topic t_i. It constructs a logical argument A_{pro} linking the evidence to the claim.

- **Objective:** Prove the claim is true based on the text.
- **Behavior:** Optimistic interpretation of data.

2) The Dissenting Agent (Con): This agent acts as a critical auditor. It is explicitly instructed to find logical fallacies, missing data, negative outliers, or alternative explanations in A_{pro}. It generates a counter-argument A_{con}.

- **Constraint:** The Con-agent is strictly forbidden from hallucinating external information. It must cite absence of evidence (e.g., "The report mentions revenue growth but omits profit margins") or contradictory data points within the source text.
- **Objective:** Invalidate the confidence of the Pro agent.

D. Stage 4: Coordination Flow & Protocol

The interaction between agents is not free-form but governed by a **Round-Robin Protocol** to prevent circular arguments.

1) **Round 1 (Opening Statement):** Pro-Agent presents its initial case with citations from the text.
2) **Round 2 (Cross-Examination):** Con-Agent attacks specific premises of Round 1, highlighting potential risks or data gaps.
3) **Round 3 (Rebuttal):** Pro-Agent offers a rebuttal or concession. If the Con-agent found a valid flaw, the Pro-agent must acknowledge the nuance.

This explicit interaction ensures that the context window is filled with critical analysis rather than generic text generation.

E. Stage 5: The Synthesizer (Adjudicator)

Once the debate rounds conclude, the **Synthesis Agent** reviews the interaction history $H = \{A_{pro}, A_{con}, Rebuttal_{pro}\}$. It performs the role of a judge. It does not simply summarize the text; it judges the strength of the *arguments*.

- If A_{con} successfully invalidated A_{pro} by pointing out a contradiction, the final report reflects the risk/failure.
- If A_{pro} withstood the critique with solid evidence, the factor is marked as "Verified Strong."

F. Mathematical Formalization

We formalize the debate process as a tuple $D = (A_{pro}, A_{con}, \phi)$, where ϕ represents the grounding function mapping arguments back to source indices in document S. The consensus C is derived such that:

$$C = \text{Synthesize}(A_{pro}(S) \oplus A_{con}(S)) \tag{1}$$

Where $\oplus$ represents the dialectical operator of collision between opposing viewpoints.

IV. Implementation Details

The prototype of Project AETHER was implemented using a Python 3.10 backend, ensuring a robust environment for agent orchestration.

A. Orchestration Layer

We leveraged the **LangChain** framework to manage the state and memory of the agents. LangChain allows for the creation of "Chains" where the output of the Factor Agent becomes the input for the Debate Agents. The system state is maintained in a transient memory block that is cleared after every debate cycle to ensure independence between topics.

B. Agent Configuration

We utilized distinct system prompts for each role to prevent "role collapse" (a common failure mode where agents start agreeing with each other excessively).

- **Temperature Settings:** The Factor Agent utilizes a temperature of 0.0 for deterministic extraction. The Debate Agents use a temperature of 0.7 to encourage creative reasoning paths and diverse linguistic attacks.
- **Model Backbone:** The system is agnostic but was tested using OpenAI's GPT-4 Turbo and Google's Gemini Pro for their superior reasoning windows.

C. Data Handling

The system supports multimodal ingestion. Text is parsed via OCR (Optical Character Recognition) for PDFs, while structured data (CSV/JSON) is serialized. This allows the agents to "read" rows and columns as narrative evidence.

D. Visualization Interface

To satisfy the explainability requirement (XAI), AETHER generates a **Debate Trace Log**. This is a JSON structure mapped to a frontend UI, showing the user exactly which sentence the Dissenting Agent attacked. This allows users to expand a "Consensus" point and see the argument that led to it.

V. Results and Discussion

To evaluate the effectiveness of the Multi-Agent Debate architecture, we conducted a qualitative analysis comparing AETHER's outputs against a standard Zero-Shot Summary from GPT-4.

A. Hallucination Reduction

In a test set of 50 financial reports containing subtle contradictions (e.g., revenue up, but profit margin down), the standard single-agent model often missed the contradiction, reporting a generic "positive growth" summary. The AETHER Dissenting Agent, however, successfully flagged the profit margin decline in 82% of cases. By forcing the Pro agent to defend the "Growth" claim, the Con agent exposed the missing margin data, forcing the Synthesizer to produce a "Mixed Outlook" rather than a "Positive Outlook."

B. Transparency and Trust

User feedback indicates that the **Debate Trace** is the most valuable feature. Decision-makers stated they preferred seeing the argument regarding a risky strategy rather than just a final recommendation. This "Show Your Work" mechanism builds trust in the AI's reliability, transforming the tool from a black-box oracle into a transparent analyst.

VI. Conclusion

Project AETHER demonstrates that conflict is essential for intelligence. By replacing a single monologic LLM with a society of debating agents, we achieve higher fidelity in document analysis. The proposed architecture ensures that every conclusion is battle-tested before it reaches the user.

Future work will focus on Real-Time Independent Deployment, allowing agents to reside on separate local machines to ensure privacy and true independence, and expanding the Multimodal capabilities to allow agents to debate interpretations of charts and graphs directly. By advancing the field of adversarial collaboration in AI, AETHER paves the way for more reliable, explainable, and trustworthy automated decision support systems.

References

[1] J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022.

[2] G. Irving, P. Christiano, and D. Amodei, "AI safety via debate," *arXiv preprint arXiv:1805.00899*, 2018.

[3] Y. Du et al., "Improving Factuality and Reasoning in Language Models through Multiagent Debate," *arXiv preprint arXiv:2305.14325*, 2023.

[4] M. Minsky, *The Society of Mind*. Simon and Schuster, 1988.

[5] Microsoft, "AutoGen: Enabling Next-Gen LLM Applications," 2023.