International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4918-4926
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0646

Language Model Council: A Multi-Agent Framework using Explainable AI

Deekshitha M[1], Shwetha KR[2], Divya G S[3], Deepika M[4], Abhishek Somanath Mali[5], Abhishek Mudakayyan Math[6]

[1,4,5,6]Student, Department of Computer Science and Engineering, AMCEC, India.

[2,3]Assistant Prof, Department of Computer Science and Engineering, AMCEC, India.

Emails: *deekshum1309@gmail.com[1], kr.shwetha12@gmail.com[2], divyags.siddaraj@gmail.com[3], deepikam0531@gmail.com[4], abhiasm25@gmail.com[5], abhim0353@gmail.com[6]*

Abstract

The rapid advancement of Large Language Models (LLMs) has significantly expanded the capabilities of Artificial Intelligence in language understanding, reasoning, and automated decision support. Despite these achievements, systems built around a single language model remain vulnerable to problems such as factual inaccuracies, hallucinated information, inconsistent outputs, limited explainability, and unintended bias. These shortcomings restrict their use in applications where decisions must be accurate, transparent, and accountable. This work introduces the Language Model Council (LMC), a collaborative framework that combines the expertise of multiple specialized AI agents to evaluate a user query from different perspectives. Their independent analyses are consolidated through a consensus-driven mechanism that selects the most reliable response. To further improve transparency, the framework integrates Explainable Artificial Intelligence (XAI), providing confidence estimates together with concise reasoning summaries that clarify how the final decision was derived. Experimental evaluation indicates that the proposed approach outperforms traditional single-model systems by improving response quality, reducing hallucinations, and increasing user trust through enhanced explainability.

Keywords: *Large Language Models, Multi-Agent Systems, Explainable Artificial Intelligence, Consensus-Based Decision Making, Trustworthy AI, AI Governance, Decision Support.*

1. Introduction

Recent developments in Large Language Models (LLMs) have significantly expanded the capabilities of Artificial Intelligence in understanding and generating human language. Owing to advances in deep learning and large-scale training, these models are now widely used in applications such as question answering, document summarization, machine translation, conversational assistants, software development, and decision-support systems. Their ability to process complex language tasks has accelerated the adoption of AI across sectors including healthcare, finance, education, legal services, and customer support. Despite these advances, depending on a single language model presents several challenges. Individual models may occasionally generate inaccurate information, produce inconsistent responses, or exhibit unintended bias. Another important limitation is the lack of transparency in the reasoning process, making it difficult for users to understand how conclusions are reached. These issues reduce confidence in AI-generated outputs and limit the suitability of standalone models for applications where accuracy, accountability, and reliability are critical. A promising way to address these limitations is through collaborative intelligence. Similar to human expert panels, multiple AI agents can independently analyze the same problem, contribute different perspectives, and collectively determine the most reliable solution. Such cooperation helps identify weaknesses in individual analyses while improving the overall quality and consistency of the final response. Motivated by this concept, this paper presents the Language Model Council (LMC), a collaborative multi-agent framework designed to improve both the reliability and explainability of AI-generated decisions. Instead of assigning every responsibility to a single language model, the framework employs

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4918-4926
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0646

specialized agents responsible for logical reasoning, factual verification, safety assessment, and explanation generation. Their independent outputs are integrated using a consensus-based decision mechanism to produce responses that are more accurate, dependable, and transparent. To further enhance user trust, the proposed framework incorporates Explainable Artificial Intelligence (XAI) techniques. In addition to generating a final response, the system provides confidence estimates and concise reasoning summaries that help users understand the factors influencing the decision. By making the reasoning process more transparent, the framework enables users to better evaluate the reliability of the generated output. The major contributions of this research are summarized as follows:

- A collaborative multi-agent framework that enables specialized language model agents to analyze user queries independently and generate consensus-driven responses.
- A decision-making strategy that combines reasoning quality, factual verification, safety assessment, and confidence estimation to improve response reliability.
- An Explainable AI module that provides interpretable reasoning and transparent outputs to improve user understanding and trust.
- A comprehensive evaluation demonstrating improvements in accuracy, consistency, safety, hallucination reduction, and explainability when compared with conventional single-model language systems.

2. Related work

Du et al. [1] introduced the concept of Multi-Agent Debate, where several language models independently generate responses and then evaluate each other's reasoning before reaching a final conclusion. Their work showed that collaborative discussion significantly improves factual accuracy and logical reasoning while reducing hallucinated content. This research laid the foundation for many later multi-agent reasoning frameworks. Liang et al. [2] proposed the Multi-Agent Debate (MAD) framework to encourage diverse reasoning among AI agents. Instead of producing similar responses, the participating agents were encouraged to explore different perspectives, after which a judge agent selected the most convincing solution. Their experiments demonstrated noticeable improvements in arithmetic reasoning, commonsense understanding, and complex problem-solving, highlighting the value of structured collaboration among language models. Guo et al. [3] presented a comprehensive survey of LLM-based multi-agent systems, covering communication strategies, planning methods, memory sharing, coordination mechanisms, and cooperative intelligence. The survey also identified several open research challenges, including efficient coordination, scalability, trustworthiness, and evaluation methods. These findings emphasize the importance of designing robust collaboration mechanisms for future AI systems. The GPT-4 Technical Report published by OpenAI [4] demonstrated the remarkable capabilities of transformer-based language models across tasks such as reasoning, coding, summarization, and text generation. At the same time, the report acknowledged persistent limitations, including hallucinations, factual inconsistencies, and limited interpretability, reinforcing the need for additional validation and governance mechanisms. Similarly, the Gemini research by Google [5] introduced multimodal language models capable of understanding text, images, and other data types within a unified architecture. Although the system achieved strong benchmark performance, the authors noted that transparency, robustness, and reliable decision-making remain important areas for future improvement, particularly in high-risk applications.

3. Proposed Solution and System Architecture

The Language Model Council (LMC) is introduced as a collaborative multi-agent framework that improves the reliability, transparency, and overall quality of AI-assisted decision-making. Unlike conventional architectures that rely on a single language model to perform every stage of processing, the proposed framework distributes different responsibilities among specialized agents. Each agent focuses on a specific function, including logical reasoning, factual verification, safety evaluation, and explanation generation. Once all agents complete

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4918-4926
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0646

their independent analyses, their outputs are reviewed collectively through a consensus-driven decision process. Rather than accepting the result from a single model, the framework compares multiple viewpoints and selects the response that demonstrates the highest level of consistency, factual correctness, and confidence. This collaborative strategy reduces the likelihood of hallucinated information while improving the accuracy and interpretability of the final output. The design philosophy is inspired by the way multidisciplinary expert committees solve complex problems. Instead of relying on the judgment of a single individual, several specialists examine the same issue from different perspectives before reaching a common conclusion. Applying a similar principle to AI allows the framework to produce responses that are more dependable, balanced, and explainable than those generated by traditional single-model systems. The end-to-end workflow of the proposed architecture is shown in Figure 1.

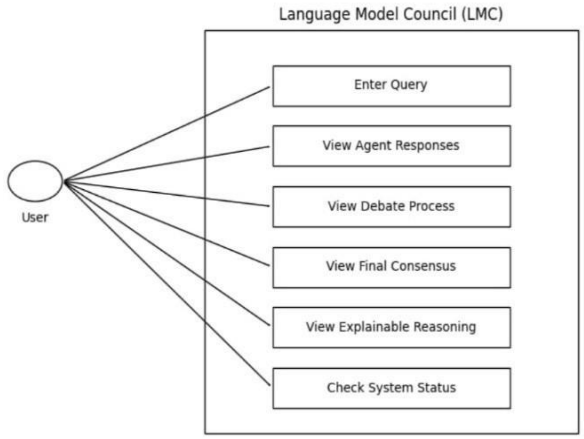

Figure 1 Proposed Architecture Diagram

3.1.User Interface Layer

The User Interface Layer serves as the communication point between users and the Language Model Council. It accepts queries through different platforms such as web applications, mobile applications, desktop software, or API-based services. Besides collecting the user's request, this layer also captures contextual information, user preferences, and session details whenever available. The objective is to ensure smooth interaction while providing the framework with sufficient context for accurate decision-making.

3.2.Input Processing Layer

After receiving a query, the framework performs several preprocessing operations before passing the request to the council agents. The input text is first cleaned by removing unnecessary symbols and correcting formatting inconsistencies. It is then normalized into a consistent representation suitable for processing. The system next identifies the user's intent, determining whether the request involves explanation, prediction, recommendation, question answering, or another task. Important entities, keywords, and contextual information are also extracted to help the downstream agents better understand the problem.

3.3.Language Model Council Layer

The Language Model Council forms the core of the proposed framework. Rather than relying on a single language model, the architecture employs several specialized agents that independently analyze the same query from different perspectives. Each agent focuses on a dedicated responsibility, allowing the framework to evaluate both the quality and reliability of the generated information before producing a final response. The Reasoning Agent performs logical analysis of the user's query. It breaks complex problems into smaller reasoning steps, identifies relationships between relevant information, and constructs a structured solution. For analytical or multi-step problems, this agent ensures that the reasoning process remains coherent and logically consistent. The Fact Checker Agent verifies the correctness of the information produced during reasoning. It compares generated statements with trusted knowledge sources and identifies unsupported claims, contradictions, or factual inconsistencies. By validating information before it reaches the final decision stage, this agent helps reduce hallucinations and improves the overall reliability of the framework. The Safety and Risk Agent evaluates whether the generated response complies with ethical and security guidelines. It identifies potentially harmful, biased, offensive, or unsafe content that could negatively affect users. The Explainability Agent increases transparency by describing how the final decision was reached. Instead of presenting only the final answer, it summarizes the major reasoning steps,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4918-4926
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0646

highlights the evidence considered during decision-making, and provides concise explanations that users can easily understand. These explanations improve user confidence and make the framework more trustworthy.

3.4.Consensus and Decision Layer

After each specialized agent completes its independent analysis, their outputs are forwarded to the Consensus and Decision Layer. At this stage, the framework compares the responses, evaluates their confidence levels, and identifies areas of agreement and disagreement. Rather than selecting the response of a single agent, the framework combines multiple viewpoints using a weighted consensus strategy. Agents that produce more reliable and consistent outputs receive greater influence during decision-making. This collaborative process helps eliminate isolated errors and produces a final response that is more accurate, balanced, and dependable than any individual agent could generate on its own.

3.5.Mathematical Reliability Model

$R = \alpha Q + \beta F + \gamma S + \delta E$ Where R represents the overall reliability score, Q denotes reasoning quality, F indicates factual accuracy, S represents safety compliance, and E corresponds to explainability score. The coefficients α, β, γ, and δ are adjustable system weights.

3.6.Explainability and Output Layer

The final stage of the framework focuses on presenting the selected response in a way that is both informative and easy to understand. In addition to the final answer, users receive a confidence score, supporting evidence, and a brief explanation describing the reasoning behind the decision. Depending on the intended application, the framework can also generate detailed reports, maintain decision histories, and use user feedback to improve future system performance. Providing transparent explanations alongside reliable responses helps users better understand the decision-making process, increases confidence in the system, and encourages the responsible use of AI in real-world applications.

4. Methodology

The proposed Language Model Council (LMC) follows a collaborative multi-agent methodology in which several specialized AI agents work together to process a user's query. Instead of depending on a single language model, the framework distributes the task among agents with different responsibilities, allowing each one to evaluate the problem independently. Their individual findings are then combined through a consensus mechanism to generate a reliable and explainable final response. The complete methodology consists of six major stages: problem analysis, input preprocessing, multi-agent reasoning, consensus-based decision making, explanation generation, and performance evaluation. Each stage contributes to improving the overall accuracy, transparency, and reliability of the framework.

4.1.Problem Analysis

Large Language Models (LLMs) have made significant advances in recent years and now perform exceptionally well across a wide range of language-based tasks. However, relying on a single model can still present several challenges. Responses may occasionally contain factual errors, inconsistent reasoning, or unintended bias, particularly when handling complex or ambiguous queries. In addition, many models provide answers without clearly explaining the reasoning behind their conclusions, making it difficult for users to assess the reliability of the information. These limitations are especially critical in domains such as healthcare, finance, legal services, and public administration, where incorrect or non-transparent decisions may lead to serious consequences. Therefore, a framework capable of validating information, improving reasoning quality, and explaining its decisions is essential for developing trustworthy AI systems.

4.2.Framework Development

To improve the reliability and transparency of AI-generated responses, the proposed Language Model Council was designed as a collaborative architecture consisting of multiple specialized AI agents. Rather than assigning every task to a single language model, the framework distributes responsibilities across agents dedicated to reasoning, factual verification, safety assessment, and explanation generation. Each agent independently evaluates the same user query according to its assigned role. The resulting outputs are collected and compared through a consensus mechanism that determines the most reliable

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4918-4926
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0646

response. This collaborative strategy allows the framework to identify potential errors made by individual agents while combining their strengths to produce a more balanced and dependable outcome.

4.3. Data Collection and Input Processing

The framework begins its execution when a user submits a query through the interface. Before the query is analyzed, it undergoes a preprocessing stage to ensure that the input is suitable for subsequent reasoning. Initially, unnecessary symbols, formatting inconsistencies, and irrelevant characters are removed from the text. The cleaned input is then normalized into a consistent representation that can be processed efficiently by all participating agents. Next, the system identifies important keywords, extracts contextual information, and determines the user's intent. After identifying the user's intent, the framework categorizes the request into tasks such as question answering, recommendation, explanation, prediction, summarization, or decision support. The preprocessing phase organizes and refines the input, ensuring that the Language Model Council receives accurate contextual information. As a result, each specialized agent can focus on the aspects of the query that are most relevant to its specific responsibility, improving the overall quality of the final response.

4.4.Multi-Agent Query Distribution

After preprocessing is complete, the refined query is distributed simultaneously to all participating agents within the Language Model Council. Since each agent specializes in a different aspect of decision-making, they analyze the same problem independently while working in parallel. The Reasoning Agent performs logical analysis and constructs a structured solution based on the available information. The Fact Checker Agent verifies factual correctness and identifies unsupported or contradictory statements. The Safety Agent examines the generated content to detect harmful, biased, or ethically inappropriate information before it reaches the final stage. The Explainability Agent is responsible for summarizing the reasoning process and presenting clear, concise explanations that describe the factors influencing the final decision. These explanations improve transparency and make it easier for users to understand why a particular response was selected. 4.5 Independent Response Generation Each participating agent produces an independent assessment of the user's query based on its assigned responsibility. Along with the generated response, every agent assigns a confidence score that reflects its certainty regarding the correctness of its analysis. These individual responses are temporarily stored within the council for comparison. During the consensus stage, the framework evaluates areas of agreement, identifies conflicting opinions, and considers the confidence assigned by each agent. Responses supported by stronger evidence and higher confidence receive greater influence in the final decision.

4.5.Explainability Generation

Once the consensus mechanism selects the final response, the Explainability Agent generates a clear description of how the decision was reached. Rather than presenting only the final answer, the framework summarizes the key reasoning steps, identifies the evidence considered by different agents, and explains why the selected response was preferred over alternative interpretations Rather than presenting only the final answer, the framework also explains the reasoning behind it. This additional context helps users understand the basis of the recommendation, evaluate its reliability, and develop greater confidence in the overall decision-making process.

4.6.Experimental Evaluation Method

The performance of the proposed Language Model Council was examined using benchmark datasets designed to represent [6] a variety of AI tasks, including logical reasoning, factual question answering, decision support, and safety-critical scenarios. To provide a reliable comparison, the proposed framework and a standalone language model were tested with identical datasets, prompts, and evaluation parameters. Using the same experimental setup ensured that any observed differences in performance resulted from the framework itself rather than variations in the testing conditions. The experimental study focused on determining whether collaborative multi-agent reasoning could improve response quality without sacrificing consistency or transparency [7].

4.7.Performance Metrics

System performance was measured using several

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4918-4926
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0646

evaluation parameters. Accuracy was used to measure correctness of answers. Consistency was used to determine whether repeated queries produced stable responses. Hallucination rate measured unsupported or false outputs. Safety compliance evaluated rejection of risky content. Explainability score measured clarity and usefulness of generated reasoning [8].

5. Results And Discussions

5.1.Experimental Setup

To validate the proposed Language Model Council, experiments were carried out using benchmark datasets that represent a wide range of real-world AI tasks. The evaluation covered logical reasoning, factual question answering, ambiguous decision-support situations, and safety-sensitive queries to examine the framework under different operating conditions [9].The same benchmark queries were provided to both the proposed framework and a standalone language model to ensure an unbiased comparison. Unlike the baseline model, which generated responses independently, the Language Model Council processed each query through multiple specialized agents responsible for reasoning, factual validation, safety evaluation, and explainability before combining their outputs through a consensus mechanism. System performance was assessed using five complementary metrics: accuracy, consistency, hallucination rate, safety compliance, and explainability. Collectively, these measures provide a comprehensive evaluation of response quality by considering not only correctness but also reliability, transparency, and practical usefulness in real-world decision-making scenarios.

5.2.Performance Comparison

Performance Metric	Standalone LLM	Proposed LMC
Accuracy	84%	93%
Consistency	78%	95%
Hallucination Rate	18%	6%
Safety Compliance	82%	94%
Explainability Score	71%	96%

Figure 2 Performance Comparison

Table 1 shows that the proposed Language Model Council outperformed the standalone model across all evaluation parameters.

5.3.Accuracy Analysis

The proposed framework achieved an accuracy of 93%, compared with 84% for the standalone language model. This improvement can largely be attributed to the collaborative reasoning strategy adopted by the Language Model Council. Since multiple agents independently evaluated each query, errors introduced by one agent were often identified and corrected during the consensus stage. The additional fact-verification process further improved the correctness of the final responses, resulting in higher overall accuracy [10].

5.4.Consistency Analysis

Consistency measures how reliably a system produces similar responses when presented with identical or closely related queries. In the experimental evaluation, the Language Model Council achieved a consistency score of 95%, whereas the standalone model recorded 78%. The improvement demonstrates the effectiveness of collaborative reasoning. Because the final response is generated only after comparing multiple independent analyses, random variations and contradictory outputs are significantly reduced. As a result, users receive responses that are more stable and dependable across repeated interactions [11].

5.5.Explainability Performance

Explainability is one of the core strengths of the proposed framework. Rather than limiting the output to a single response, the Language Model Council supplements each recommendation with a concise description of the reasoning process. Providing this additional context helps users understand the basis of the decision and makes the system more transparent than traditional single-model approaches. Experimental results showed that the explainability score increased from 71% in the standalone model to 96% in the proposed framework. By presenting confidence scores and concise reasoning summaries, the Explainability Agent enables users to better understand the factors that influenced the final recommendation. This additional transparency strengthens user confidence and supports informed decision-making, particularly in high-risk application

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4918-4926
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0646

domains.

5.6.Comparative Graph Representation

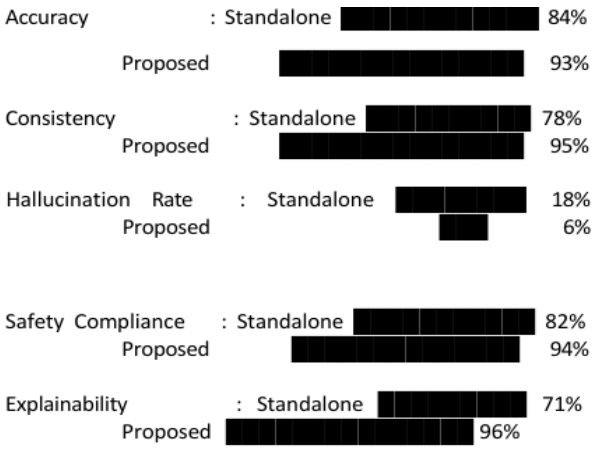

Figure 3 Graph Representation

6. Limitations and Future Work

Although the proposed Language Model Council (LMC) demonstrates clear improvements in reliability, consistency, safety, and explainability compared with conventional single-model approaches, several challenges remain that should be addressed in future research. The collaborative nature of the proposed framework introduces additional computational requirements. Since every user query is processed by multiple specialized agents, the system requires greater computing resources, including increased processing capacity and memory, before producing the final response. Although this design contributes to higher-quality and more dependable outputs, it also presents challenges for deployment in resource-constrained environments. Another challenge is the additional response time introduced by the consensus process. Before generating the final output, the framework must collect, compare, and evaluate responses from multiple agents. Although this improves the reliability of the final decision, it may increase latency when handling complex queries or operating in environments with limited computational resources. Consequently, the current implementation may be less suitable for applications that require immediate real-time responses. Although the consensus mechanism helps improve decision quality, the framework ultimately relies on the performance of its individual agents. Weak or biased outputs from any participating agent can influence the collective result and reduce the reliability of the final response. Enhancing the robustness, accuracy, and adaptability of these specialized agents is therefore a key area for future investigation. Another practical challenge arises when agents reach substantially different conclusions for ambiguous or complex problems. Resolving these disagreements may require additional rounds of discussion or more sophisticated consensus strategies, which can further increase computational overhead. Developing adaptive mechanisms capable of efficiently handling conflicting opinions without sacrificing response quality represents an important research opportunity. At present, the proposed framework has been assessed mainly using standardized benchmark datasets. While the obtained results demonstrate its potential, evaluating the framework in real-world settings is essential to understand its practical applicability and long-term reliability. Future investigations should examine its deployment across diverse sectors, including healthcare, finance, education, legal services, and public administration, where accurate, transparent, and accountable decision support plays a crucial role. Evaluating explain ability also remains a challenging task because users with different backgrounds may interpret explanations differently. Future research should therefore consider user-centered evaluation methods that measure not only technical performance but also how effectively explanations improve user understanding, confidence, and trust. As collaborative AI frameworks become more widely deployed across industries and government services, ensuring secure, responsible, and compliant operation will become increasingly critical. Future developments should emphasize stronger data protection strategies, reliable communication mechanisms for inter-agent collaboration, and governance frameworks that support compliance with evolving AI regulations while promoting ethical and trustworthy system behavior. The Language Model Council represents an important step toward the development of reliable and explainable collaborative AI. Moving forward, research efforts will aim to improve the framework by reducing computational complexity, introducing adaptive

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4918-4926
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0646

strategies for agent selection, enhancing consensus algorithms, and enabling more flexible coordination among participating agents. In addition, extensive validation in real-world environments will be essential for assessing its practical performance. Collectively, these improvements will strengthen the framework's scalability, operational efficiency, and readiness for deployment in real-world applications.

Conclusion

The proposed Language Model Council (LMC) demonstrates how collaboration among specialized AI agents can improve the quality of automated decision-making. By dividing complex tasks across agents dedicated to reasoning, factual validation, safety assessment, and explainability, the framework produces responses that are more reliable than those generated by a single language model. The consensus-based integration of individual agent outputs further enhances the consistency and accuracy of the final response. Another important feature of the framework is its emphasis on explainability. Along with each generated response, the system provides confidence estimates and concise reasoning summaries that help users understand the basis of its decisions. Rather than functioning as an opaque AI system, the framework promotes transparency by making the decision-making process easier to interpret. As a result, it is well suited for applications where reliable, explainable, and accountable AI is required. Experimental evaluation demonstrated that the proposed framework consistently outperformed a standalone language model across several performance metrics, including accuracy, consistency, safety compliance, hallucination reduction, and explainability. The results indicate that collaborative reasoning enables the framework to identify and correct individual agent errors before generating the final response, thereby improving the overall quality of AI-generated outputs. Although the framework introduces additional computational overhead and increased response time because of its multi-agent architecture, the improvements in reliability and interpretability make these trade-offs worthwhile for many decision-support applications. With continued optimization of consensus strategies, adaptive agent selection, and scalable deployment techniques, the framework can become more efficient while maintaining its high level of performance. The findings of this study indicate that collaborative multi-agent intelligence provides an effective strategy for overcoming many of the limitations of traditional standalone language models. By coordinating specialized agents responsible for reasoning, factual validation, safety assessment, and explainability, the proposed framework delivers reliable, transparent, and trustworthy decision support within a unified architecture. This approach has strong potential for deployment in critical sectors such as healthcare, finance, education, legal services, and public administration. Future investigations will emphasize large-scale real-world validation, improved computational efficiency, and the development of more adaptive and scalable multi-agent collaboration mechanisms.

References

[1]. Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, “Improving Factuality and Reasoning in Language Models through Multiagent Debate,” arXiv preprint arXiv:2305.14325, 2023.
[2]. T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, and S. Shi, “Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate,” arXiv preprint arXiv:2305.19118, 2023.
[3]. T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. Chawla, and H. Wi, “Large Language Model based Multi-Agents: A Survey of Progress and Challenges,” arXiv preprint arXiv:2402.01680, 2024. [4] OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.
[4]. Google, “Gemini: A Family of Highly Capable Multimodal Models,” arXiv:2312.11805, 2023. arXiv preprint
[5]. Meta AI, “Llama 2: Open Foundation and FineTuned Chat Models,” arXiv:2307.09288, 2023. arXiv preprint
[6]. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You? Explaining the Predictions of Any Classifier,” in Proceedings of the 22nd ACM SIGKDD

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 07 July 2026
Page No: 4918-4926
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0646

International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[7]. F. Doshi Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine arXiv:1702.08608, 2017. Learning," arXiv preprint

[8]. D. Bahri, J. Tay, and D. Metzler, "On the Reliability of Large Language Models in Decision Making Tasks," Proceedings of AAAI Workshop on Trustworthy AI, 2024.

[9]. Park, J. S., O'Brien, J., Cai, C., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative Agents: Interactive Simulacra of Human Behavior. Link: https://arxiv.org/abs/2304.03442

[10]. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. Link: https://arxiv.org/abs/2210.03629

[11]. Madaan, A., Tandon, N., Clark, P., & Yang, Y. (2023). Self-Refine: Iterative Refinement with Self-Feedback. Link: https://arxiv.org/abs/2303.17651 [13] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving Factuality and Reasoning in Language Models through MultiAgentDebate.Link:https://arxiv.org/abs/2305.14325

[12]. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., & others. (2023). Holistic Evaluation of Language Models. Link: https://arxiv.org/abs/2211.09110

[13]. Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. Link: https://arxiv.org/abs/2303.11366