



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85186>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Secure Multi-Agent Collaboration Using Large Language Models

Khaja Afrazuddin

Dept. of Information Technology, M.S. Ramaiah Institute of Technology, Bengaluru

Abstract: As Large Language Models have emerged and become increasingly significant, the way to develop smart multi-agent systems has been transformed, enabling them to communicate, collaborate and decision more efficiently. These systems rely on agents working independently which need to effectively and safely coordinate their actions, to maintain reliability in messy real-world environments. However, bringing LLMs into the mix introduces new risks like data leaks, trickery from bad actors, and trust issues. This paper breaks down a full plan for keeping multi-agent teamwork secure when using LLMs. We look at everything from how they are built and the rules they use to communicate to the safety tools used to stop attacks. We also talk about where this is being used now, the hurdles we still face, and where the tech is headed for building secure AI ecosystems.

Index Terms: Agentic AI, Collaboration, Large Language Models, Multi-Agent Systems, Security, Trust.

I. INTRODUCTION

Artificial Intelligence has come a long way, and we have systems now that are able to think and make choices on our own. A big part of that is the Large Language Models, which can comprehend context and model knowledge in ways we've never seen before. This has led to them being used in multi-agent systems, where groups of smart agents team up to tackle big problems. These systems live in spread-out environments where agents have to coordinate and cooperate to hit their goals. LLMs simplify these interactions by enabling agents to grasp intricate orders and adjust their approach to accommodate changes. This is monumental stuff for fields such as medicine, banking, and self-driving tech. Introducing LLM models into such environments also leaves significant security vulnerabilities. Sharing channels to exchange data means that the system is susceptible to hacks and theft of data. The only way to ensure the system remains honest and reliable is to keep the collaboration safe. In this paper, those are addressed with a safe structure that emphasizes growth and safety requirements.

II. BACKGROUND AND FUNDAMENTALS

A. Multi-Agent Systems

These are groups of independent entities that can see, think, and act. They live in a shared space and work toward their own or shared goals. Being decentralized makes them easy to scale, but it also makes keeping them safe and coordinated a real challenge.

B. Large Language Models

LLMs are deep learning models trained on massive amounts of data to talk and think like humans. They act as the "brain" for an agent, handling the planning and communication. Their ability to understand context makes them perfect for working in a team.

C. Need for Secure Collaboration

Agents in these environments are constantly talking. Without the right safety nets, these conversations can be hacked, leading to bad behaviour or total system failure. Secure teamwork keeps data private and ensures the system is always there when you need it.

III. SYSTEM ARCHITECTURE FOR SECURE MULTI-AGENT COLLABORATION

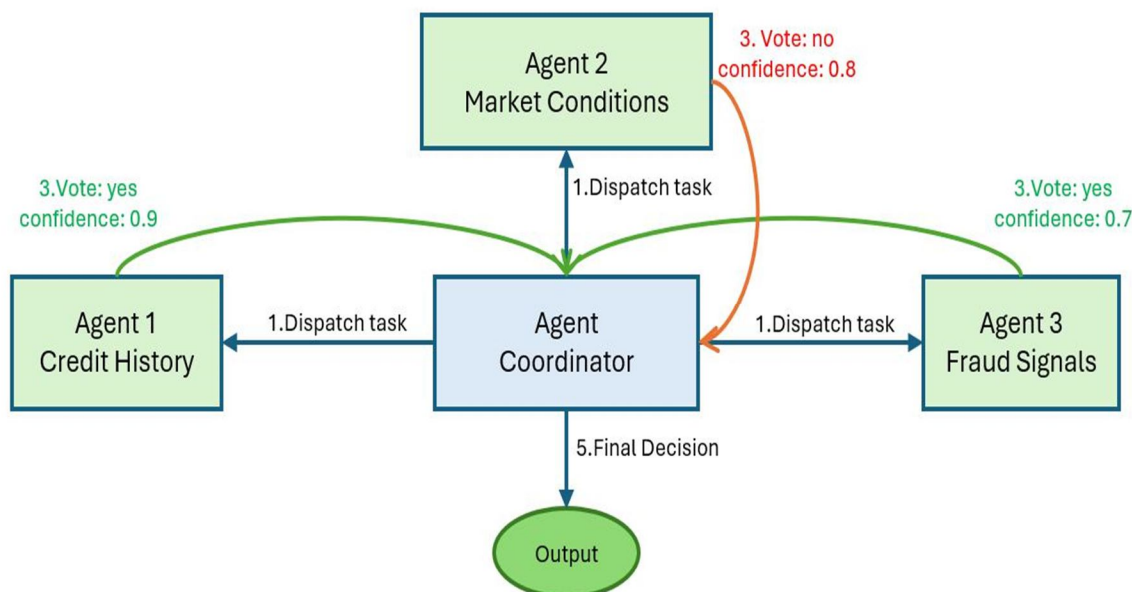
We build these secure systems such that they are robust and trustworthy even when they span multiple locations. The design is broken up into several layers, each one doing a different part of the job. At the core is the agent layer where the LLM-powered bots reside. These agents can understand natural language and make complex autonomous decisions. They work on their own, but they do follow some rules so that they are working as a team. The communication layer handles the actual talking, using encryption to keep messages private. We can use things like Transport Layer Security or even blockchain to make sure no one messes with the data. This layer also checks that only invited agents are part of the conversation.

Above that, the coordination layer divvies up the work, so tasks go to the agents best suited for them. To keep everyone honest, a trust layer constantly checks how each agent is acting. It gives out scores based on past behaviour, which helps the system avoid agents that are acting weird or malicious. Finally, the data layer keeps all the info locked down using encryption and privacy tools like federated learning.

IV. METHODOLOGY

Our approach focuses on making teamwork safe by mixing LLM thinking with trust and risk checks. First, every agent gets a specific role, a set of skills, and the right digital keys to get in. The LLMs inside are tuned to handle specific tasks so they stay accurate. When agents talk, they use secure lines and digital signatures to prove who they are and that the message has not been changed. Along with this, we use a dynamic trust system.

Agents get scores based on how well they have done their jobs and what other agents think of them. This lets the system pivot if an agent starts acting up. Risk checking is also a huge part of the process. Before a task is handed out, the system guesses what could go wrong. If an agent looks too risky, it gets blocked or watched more closely. Decisions are made through a mix of LLM reasoning and group agreement, using things like voting or negotiation to find the best path forward without risking security.

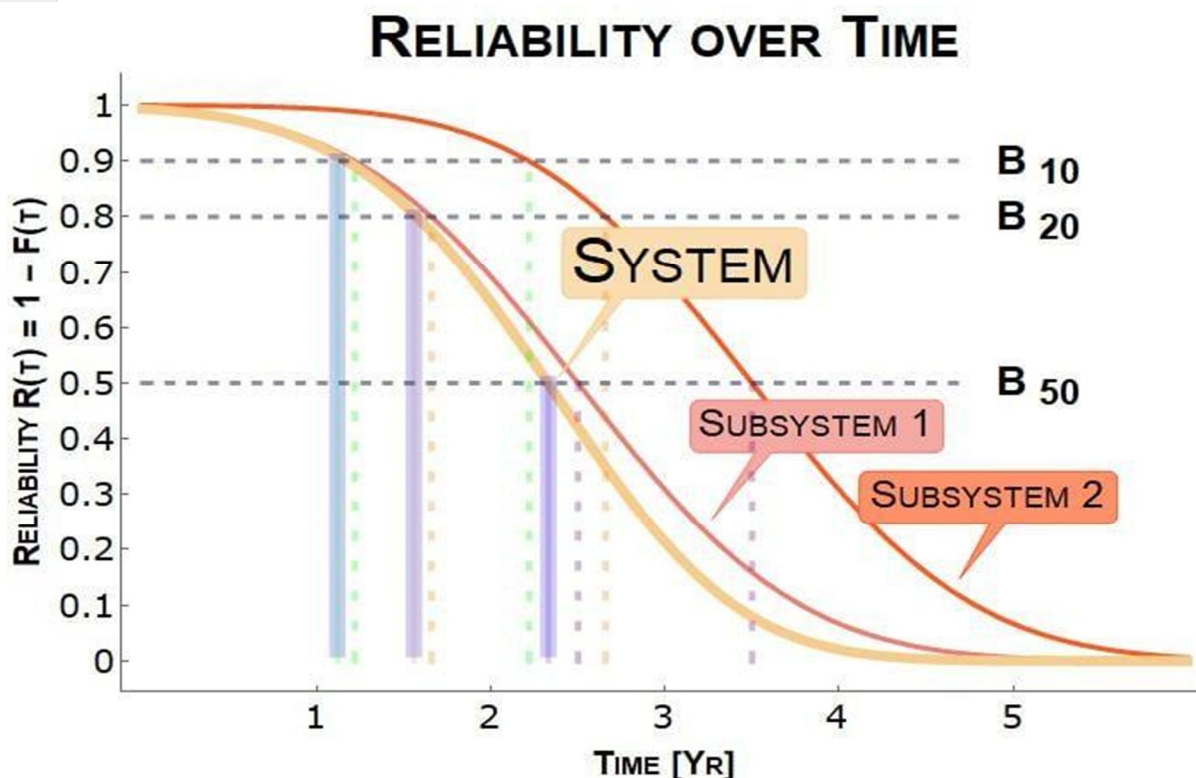


V. APPLICATIONS

Safe AI teamwork using LLMs is going to be huge in several areas. In self-driving cars and robotics, agents work together to navigate and avoid obstacles. Safety tools ensure they stay coordinated even when the environment gets messy. In cybersecurity, these systems can team up to find and stop hacks as they happen. LLMs help them understand complex attacks, while secure lines let them fight back as a unit. Hospitals use them for diagnosing patients and planning treatments while keeping sensitive records private. Smart cities also win here, using agents to run things like traffic lights and power grids safely. In the financial world, agents catch fraud and trade stocks, where trust tools are vital for stopping big money losses.

VI. CHALLENGES

Even with the wins, there are some big hurdles. Managing trust in a fast-moving environment is hard because both behaviour changes and info is not always perfect. Security is a constant concern as LLMs can be tricked by malicious prompts or leak data accidentally. Scaling up is also an issue: the more agents you add, the more computer power you need just to keep them talking safely. Already LLMs are resource intensive and hard to deploy in small devices or locations with limited power. We have also got to think about the ethics. The only way people will trust these bots in the real world is if they are fair and transparent.



VII. FUTURE SCOPE

Future work will likely look at using quantum computing to make encryption unhackable and speed up big decisions. We also expect trust models to get better by using blockchain to keep records that cannot be faked. Another big step is building agents that can learn and change on their own using reinforcement learning. "Explainable AI" will also be important, allowing us to see why a bot made a specific choice. Combining the smarts of LLMs with traditional logic could lead to even more reliable systems. Finally, the only way we can ensure these AI ecosystems are safe and work together across different companies and countries is to establish global rules and standards.

VIII. CHALLENGES

Despite all the progress, some big issues remain. Scaling is a major headache because tracking trust across thousands of agents is incredibly complex. Keeping security consistent across a huge system is just as hard. Plus, LLMs need a ton of power, which can limit where you can use them. Privacy and ethics are still big questions, especially for sensitive work. And without standard rules, it is hard for different AI systems to "talk" to each other properly.

IX. FUTURE DIRECTIONS

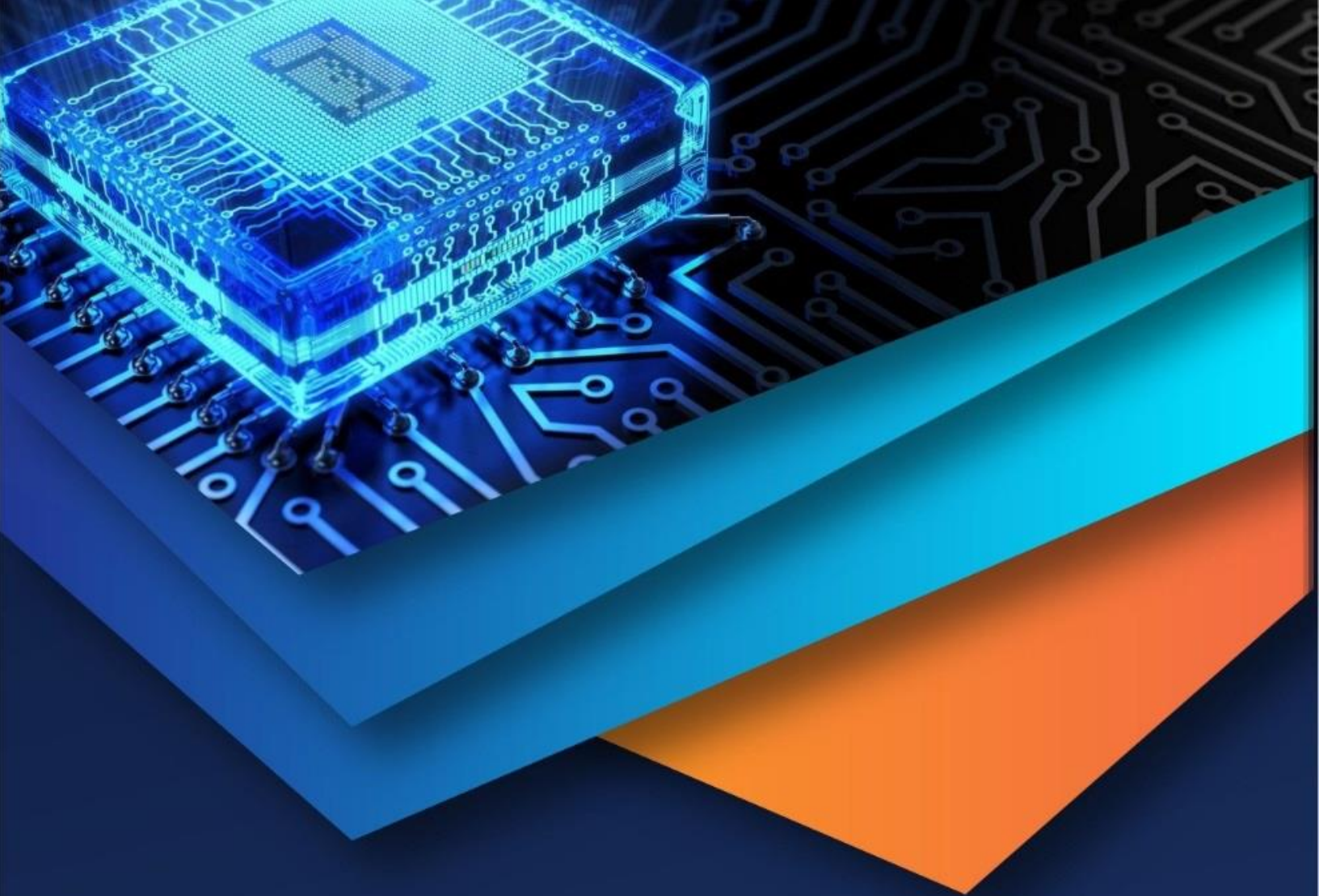
Down the road, we want to bring in quantum computing to make things more efficient. Quantum security could offer better locks and faster ways to spot threats. Agentic AI will likely move toward systems that can govern themselves and manage their own security. We might also see more blockchain use for trust that nobody can tamper with. Making AI more "explainable" will help humans understand what is going on under the hood, which builds more trust in the long run.

X. CONCLUSION

Incorporating LLMs into multi-agent systems is a huge step forward, providing us with improved ways to collaborate and solve problems. But we need to take the security risks seriously. This paper gave a safe teamwork plan based on architecture, trust and learning by doing. "With encryption and risk tools, we can make these systems much more resilient. The key to the next generation of smart computing will be to build secure and honest AI ecosystems as the tech continues to move.

REFERENCES

- [1] M. Wooldridge, *An Introduction to MultiAgent Systems*, 2nd ed. Chichester, U.K.: Wiley, 2009.
- [2] C. Thumma et al., "Dynamic retrieval augmented generation for enterprise knowledge intelligence systems using large language models," in *Proc. 2026 Int. Conf. Intelligent and Sustainable AI Systems (ICOSAAS)*, IEEE, Jul. 2026, doi: 10.1109/ICOSAAS68663.2026.11648540.
- [3] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Hoboken, NJ, USA: Pearson, 2020.
- [4] T. B. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.
- [5] V. K. Kasula et al., "Agentic AI-driven CBDC: A privacy-preserving and regulatory-compliant digital payment framework," in *Proc. Int. Conf. Applied Machine Learning and Data Analytics*, Springer Nature Switzerland, Dec. 2024, pp. 74–88, doi: 10.1007/978-3-032-18293-7_6.
- [6] N. R. Jennings, "An agent-based approach for building complex systems," *Commun. ACM*, vol. 44, no. 4, pp. 35–41, 2001.
- [7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [8] Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
- [9] A. S. Sribhashyam et al., "Agentic AI framework for autonomous decision making in enterprise-scale distributed systems," *J. Intelligent Decision Making and Information Science (JIDMIS)*, vol. 3, no. 11s, pp. 2767–2784, Sep. 2026.
- [10] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.
- [11] V. Ravuri et al., "Cooperative spectrum sensing: A review," in *Proc. Int. Conf. Smart Computing and Informatics*, Springer Nature Switzerland, Apr. 2025, pp. 479–489, doi: 10.1007/978-3-032-08240-4_40.
- [12] L. Kagal et al., "Trust-based security in distributed systems," *Computer*, 2001.
- [13] S. Ayyamgari et al., "Authentication mechanisms in the Internet of Things: A comparative analysis across RFID, smart grids, vehicular networks, and smart homes," *Sensors*, MDPI, Sep. 2026, doi: 10.3390/s26185893.
- [14] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [15] G. Tushar et al., "Enhancing digital payment security: A blockchain-based framework with multi-factor authentication and quantum-resistant cryptography," in *Proc. 38th Conf. Open Innovations Association (FRUCT)*, Nov. 2025, pp. 364–369.
- [16] P. W. Shor, "Algorithms for quantum computation: Discrete logarithms and factoring," in *Proc. 35th Annu. Symp. Foundations of Computer Science (FOCS)*, 1994, pp. 124–134.
- [17] N. Suresh et al., "Graph-aware hierarchical federated intelligence with adaptive differential privacy for robust non-IID spectrum sensing in next-generation wireless networks," *GSC Advanced Research and Reviews (GSCARR)*, vol. 20, no. 3, Sep. 2024.
- [18] Q. Wu et al., "AutoGen: Enabling next-gen LLM applications via multi-agent conversation," *arXiv:2308.08155*, 2023.
- [19] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proc. 36th Annu. ACM Symp. User Interface Software and Technology (UIST)*, 2023.
- [20] K. Saicharan et al., "Enhancing healthcare data security using PQ-ABE and verifiable secret sharing schemes," Nov. 2025, 38th Conference of Open Innovations Association (FRUCT), pp. 377–382.
- [21] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," in *Proc. 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)