Reducing hullucination of Generative AI via Agentic AI and Edge Computing

Laha Ale[1]

[1]School of Computing and Artificial Intelligence, Southwest Jiaotong University

April 23, 2025

Abstract

Generative artificial intelligence (GenAI), particularly large language models (LLMs), has revolutionized various applications by producing coherent and contextually relevant text. However, despite their advancements, LLMs are prone to hallucinations-instances where the AI generates inaccurate or fabricated information. Retrieval-augmented generation (RAG) has emerged as a technique to enhance GenAI by integrating external knowledge sources beyond the model's training data. While RAG improves factual grounding, it alone cannot fully eliminate hallucinations. To address this limitation, agentic workflows that incorporate external tools such as APIs, search engines, and self-reflective mechanisms offer a promising solution. These workflows enable models to iteratively assess and refine their outputs, thereby reducing errors and enhancing factual accuracy. This paper presents a novel framework that combines agentic workflows with RAG within 6G networks to achieve more reliable generative AI by deploying autonomous agents that reflect on outputs and leverage real-time knowledge from external sources to improve response quality and accuracy. We explore the deployment of these workflows in 6G-enabled edge environments, facilitating scalable, real-time knowledge integration and model refinement. Our framework addresses current limitations in RAG-enhanced services by utilizing 6G edge intelligence for data fusion, dynamic knowledge base updates, and customizable AI service delivery. Through a multi-agent system comprising generator and critic agents, we effectively reduce hallucinations via iterative self-criticism, paving the way for more reliable and accurate generative AI services across diverse applications.

Reducing hallucination of Generative AI via Agentic AI and Edge Computing

Laha Ale[1 †]
[1] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China.
[†]e-mail: lala.ale@ieee.org

Abstract | Generative artificial intelligence (GenAI), particularly large language models (LLMs), has revolutionized various applications by producing coherent and contextually relevant text. However, despite their advancements, LLMs are prone to hallucinations—instances where the AI generates inaccurate or fabricated information. Retrieval-augmented generation (RAG) has emerged as a technique to enhance GenAI by integrating external knowledge sources beyond the model's training data. While RAG improves factual grounding, it alone cannot fully eliminate hallucinations. To address this limitation, agentic workflows that incorporate external tools such as APIs, search engines, and self-reflective mechanisms offer a promising solution. These workflows enable models to iteratively assess and refine their outputs, thereby reducing errors and enhancing factual accuracy. This paper presents a novel framework that combines agentic workflows with RAG within 6G networks to achieve more reliable generative AI by deploying autonomous agents that reflect on outputs and leverage real-time knowledge from external sources to improve response quality and accuracy. We explore the deployment of these workflows in 6G-enabled edge environments, facilitating scalable, real-time knowledge integration and model refinement. Our framework addresses current limitations in RAG-enhanced services by utilizing 6G edge intelligence for data fusion, dynamic knowledge base updates, and customizable AI service delivery. Through a multi-agent system comprising generator and critic agents, we effectively reduce hallucinations via iterative self-criticism, paving the way for more reliable and accurate generative AI services across diverse applications.

Introduction

Generative artificial intelligence (GenAI)[1] has transformed numerous fields, from healthcare and marketing to education and manufacturing, through its ability to generate human-like text and other content. However, one of the most pressing challenges in GenAI models including large language models (LLMs)[2], such as BERT[3], LLaMA[4], and ChatGPT[5], and large vision models (LVMs)[6], multimodality models such as ImageBind[7], Gimini[8], and Sora[9], are the tendency to hallucinate[10][11]—producing information that is either inaccurate or entirely fabricated[12]. While some hallucinations may be benign, such as incorrect trivia, others, like fabricated legal precedents or falsified medical advice, pose significant risks to users and systems alike.

Retrieval-augmented generation (RAG)[13] has emerged as a powerful tool to mitigate hallucinations by incorporating knowledge beyond a model's fixed training set. By allowing models to access external data sources during the generation process, RAG enhances factuality and improves the relevance of the output. However, RAG alone is insufficient to fully eliminate hallucinations, especially in complex or highly specialized tasks. To address this, agentic workflows—leveraging external tools such as APIs[14], search engines, and self-reflective mechanisms[15]—have shown promise. These workflows enable models to iteratively assess and refine their outputs, creating a feedback loop that significantly reduces errors and improves factual accuracy.

Humans rarely tackle complex problems using only their bare hands; instead, we employ a variety of tools, deconstruct tasks into manageable components, and iteratively refine our solutions. Similarly, agentic workflows provide a framework for AI to adopt a more structured and reflective approach to problem-solving. By breaking down tasks into subtasks, utilizing prompt-engineering strategies, and incorporating self-reflection mechanisms, agentic systems can significantly enhance AI performance. Additionally, recent advancements such as memory-tuning—an innovative technique pioneered by Lamini[16] with a reported 95% success rate—have further augmented the ability of AI models to evolve and improve over time.

In this paper, we introduce a novel framework that integrates RAG with agentic workflows within the context of 6G networks[17][18] and edge computing[19] to bolster the reliability of generative AI. Our approach involves deploying autonomous agents, specifically a generator agent and a critic agent[20][21], which collaborate to perform iterative self-criticism and content refinement. This process effectively reduces hallucinations—instances where AI generates inaccurate or fabricated information—and enhances the overall quality and factual accuracy of the generated outputs. Furthermore, we explore the implementation of this framework in 6G-enabled mobile edge computing environments, which facilitate scalable and real-time integration of multimodal data sources, thereby supporting continuous improvement of AI models.

The remainder of this paper is organized as follows: we begin by outlining the current limitations of RAG and GenAI models. Next, we discuss the challenges associated with deploying these technologies in 6G network environments. Finally, we present the potential of agentic workflows to effectively address these issues, demonstrating how our proposed framework can achieve more reliable and accurate generative AI services.

Self-Reflection and Game Theory

In the context of enhancing GenAI accuracy, self-reflection[20] serves as a critical component of agentic workflows. This concept involves prompting models, particularly LLMs, to critique their own outputs iteratively. The mechanism allows for the identification and resolution of errors such as factual inconsistencies or inefficiencies. Self-reflection can be as simple as prompting the LLM to re-evaluate a previously generated response and assess its accuracy, efficiency, or appropriateness for the task. Through this iterative process, models generate more refined and accurate responses, whether in coding tasks, knowledge generation, or other AI applications.

While self-reflection allows for internal refinement, it can be significantly enhanced by introducing multi-agent frameworks, where a generator agent and a critic agent interact [21]. In fact, multiple critic agents can be employed, each designed to evaluate distinct aspects of the output. The generator agent creates content, and the critic agent provides constructive feedback to improve it. Game theory principles[22] can guide this interaction, framing the relationship as a cooperative game where both agents aim to maximize the accuracy and reliability of the final output. By iteratively improving through feedback and reflection, the two agents can drive the generative process toward optimal outcomes. The critic agent evaluates the output by incorporating external tools such as APIs and IoT-based factual information[23], further enhancing the GenAI's accuracy.

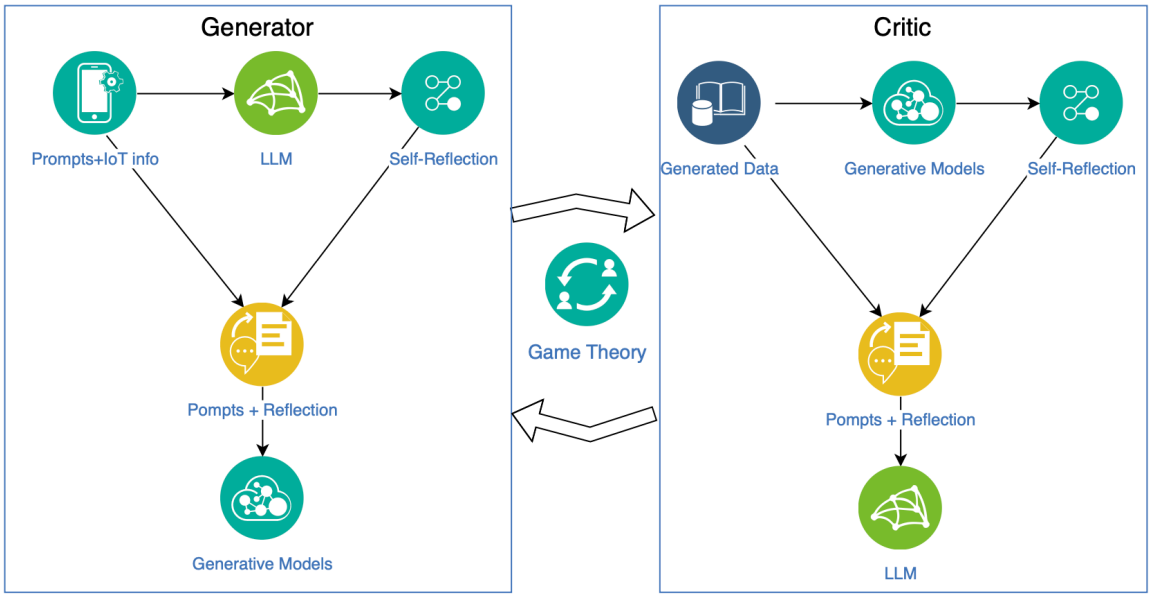

Figure 1 ***IoT-Enhanced Prompts and Self-Reflection, and Game-Theoretic Critique in a 6G Network.*** *In this framework, IoT devices in a 6G network supply additional context to the generator agent, enabling more accurate outputs from Large Language Models or other generative models. The generator agent employs internal self-reflection to refine its content, while a separate critic agent—using a game-theoretic approach—evaluates and critiques the generated results, further improving data quality and reliability.*

The integration of self-reflective agentic workflows within 6G networks offers a robust solution for reducing hallucinations in generative AI. 6G's high-speed, ultra-reliable, and low-latency characteristics enable real-time collaboration between the generator and critic agents. By leveraging mobile edge computing, these agents can access large volumes of real-time data from multiple IoT devices, improving the factual accuracy and relevance of generated content. The workflow also supports the continuous updating of knowledge bases through real-time data fusion making it possible for generative models to integrate current, contextually relevant information, into their outputs.

Game theory facilitates the effective collaboration of the generator and critic agents, optimizing the trade-offs between performance metrics like response latency and energy consumption. This approach results in a scalable, robust system capable of providing accurate, real-time generative services across diverse applications in the 6G landscape.

External Resource Access in 6G Agentic Frameworks

Tool integration[14] represents a foundational design pattern in the development of AI agentic workflows, enhancing the capabilities of LLMs by enabling interaction with external systems, data sources, and computational functions[24]. While LLMs possess significant potential through their pre-trained knowledge, their utility in real-world applications is often constrained by the static nature of this information. Tool integration addresses this limitation by allowing LLMs to dynamically access external resources, including APIs, databases, and programming environments, thereby expanding their functional reach[25].

In the context of 6G networks, tool integration is poised to become a crucial component of real-time AI applications. The high-speed, ultra-reliable, and low-latency characteristics of 6G provide an ideal infrastructure for LLMs to interact seamlessly with diverse external resources, including the Internet of Things (IoT)[26], remote databases, and real-time sensory data. This ability to integrate and process external information in real time greatly enhances the relevance, accuracy, and efficiency of LLM outputs, especially in complex, fast-paced environments.

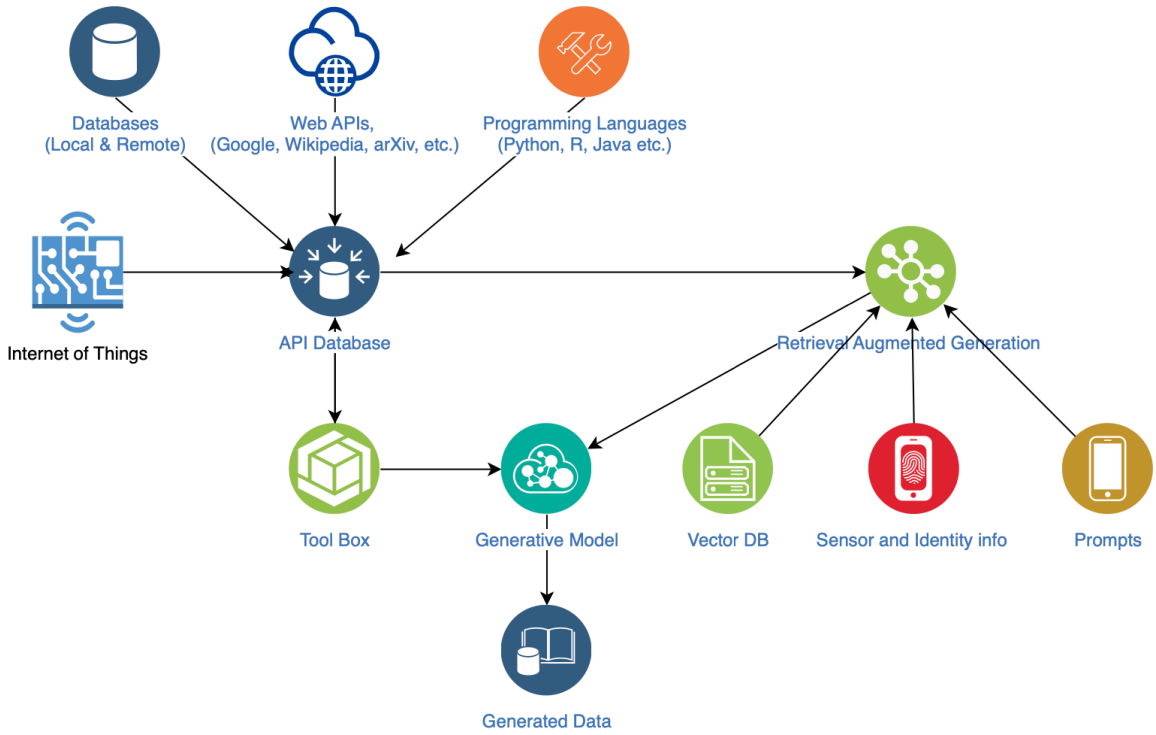

Figure 2 ***Seamless Integration of External Resources in 6G Networks for Enhanced Generative AI****. Illustrates how 6G's high-speed, ultra-reliable, and low-latency infrastructure enables large language models to interact with IoT devices, remote databases, and real-time sensory data. This integration enhances the relevance, accuracy, and efficiency of AI-generated outputs in complex and dynamic environments.*

Tool Integration in LLM-based systems typically involves the integration of functions that enable the retrieval and processing of real-time information. For instance, when an LLM encounters a query requiring knowledge beyond its training set, it may call upon a web search API or a specialized database to retrieve relevant information. This approach not only augments the generative capabilities of the LLM but also ensures that outputs remain contextually appropriate and factually accurate. Furthermore, tools can be employed for more computationally intensive tasks, such as executing code or performing mathematical operations, where the LLM generates the instructions for an external system to process and return results.

Within the 6G framework[27], such interactions are amplified by the network's capacity to support real-time data retrieval and processing. The API Database, as depicted in the workflow diagram, serves as a central hub where LLMs can request access to external functions and information sources, ranging from web APIs and databases to IoT sensors. This flexibility allows LLMs to act as dynamic agents capable of addressing a wide range of user queries and tasks with up-to-date, context-sensitive information.

The 6G network ecosystem, characterized by its massive bandwidth and ultra-low latency, is essential for realizing the full potential of tool use in generative AI systems. One of the key benefits of 6G is its ability to integrate and fuse data from multiple sources in real time. For example, LLMs operating in a smart city environment can interface with IoT devices, such as traffic sensors or weather stations, to provide accurate and real-time updates on urban conditions. This interaction relies on the RAG system, which ensures that the LLM's output is both timely and contextually relevant.

Moreover, the availability of edge computing[28] in 6G networks allows for the distributed processing of data, enabling LLMs to interact with local data sources directly at the edge of the network. This reduces the latency involved in accessing remote databases or cloud-based services and enhances the overall responsiveness of the generative AI system. Tool use in such a scenario allows LLMs to access vector databases, sensor information, and other local resources, ensuring that outputs are not only accurate but also grounded in the most recent data available.

Implications for AI-Driven Applications

Tool integration within agentic workflows, especially when deployed in 6G networks, has far-reaching implications for various AI-driven applications. By enabling LLMs to interact dynamically with real-time data and external tools, generative AI systems can provide more reliable, contextually accurate outputs across a wide range of domains, including healthcare, autonomous vehicles, and smart infrastructure. This capability reduces the risk of hallucinations—instances where the AI generates incorrect or seemingly fabricated information—by grounding responses in external, verifiable sources. While all AI-generated results are approximations, the term 'fabricated' here refers to outputs that diverge significantly from the input data or extrapolate beyond it, rather than interpolating within known information.

In addition to improving the accuracy of LLM-generated outputs, tool use also supports more complex workflows by enabling LLMs to automate processes such as querying databases, executing code, or interacting with multimodal data. This multi-functionality is essential for next-generation AI systems operating in environments where the real-time processing of data from various modalities (text, image, sensor data) is critical for decision-making.

Tool integration in agentic workflows represents a significant advancement in the development of AI systems, particularly within the context of 6G networks. The ability of LLMs to interact with real-time data sources and computational tools enables a level of flexibility and accuracy that is essential for addressing the dynamic needs of various real-world applications. As 6G networks continue to develop, the role of Tool Use in enhancing the capabilities of generative AI will only expand, making it a key component in the deployment of intelligent systems across diverse sectors.

Planning in Agentic AI Workflows

Planning[29] like Chain-of-Thought[30] is a key agentic AI design pattern in which an LLM autonomously decides the sequence of steps required to complete a larger task. This ability to dynamically decompose tasks is essential when predefined workflows are impractical or insufficient. For

example, an agent tasked with conducting online research might use planning to break the objective into subtasks[31] such as identifying relevant subtopics, retrieving information, synthesizing findings, and compiling a report.

A notable demonstration of planning occurs when an agent adapts to unexpected scenarios. For instance, during a live demo, an LLM-based agent pivoted from a web search API (which encountered a rate-limiting error) to a Wikipedia tool to complete a task—something not anticipated by its developer. This highlights the agent's ability to autonomously adjust its strategy and achieve its goal.

Planning becomes particularly useful for multi-step tasks, such as generating an image from a complex set of inputs. In one illustrative example, an LLM might be tasked with detecting a pose in an image and then rendering a new image based on that pose. The model could dynamically organize the sequence of tools needed to complete this task, demonstrating how planning supports complex workflows.

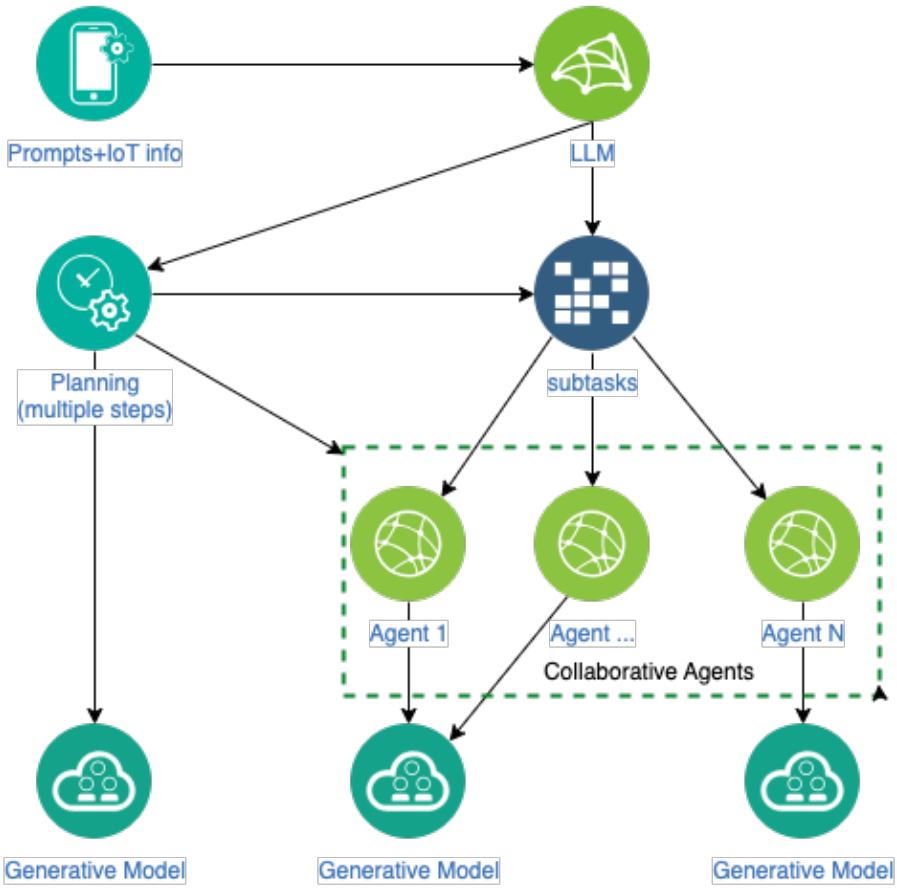

Figure 3 ***Task Decomposition and Cooperative Multi-Agent Workflows in 6G-Enabled Agentic AI****. Illustrates how planning breaks down complex tasks into subtasks, which are then addressed step-by-step by specialized AI agents. The diagram highlights cooperative agents operating within a 6G network, leveraging its high-speed, low-latency infrastructure to interact with IoT devices and external resources. This collaborative multi-agent approach enhances the execution, accuracy, and efficiency of generative AI tasks*

However, planning introduces challenges in terms of predictability. Unlike the more deterministic agentic patterns such as tool use[14] or reflection[15], planning outcomes can vary, making it harder to anticipate the exact steps the agent will choose. While the technology is powerful, it remains less mature and can lead to unpredictable results. Despite these challenges, advancements in AI research are likely to enhance planning capabilities in the near future, making it more reliable.

In the context of 6G networks, planning takes on additional importance. The high-speed, low-latency infrastructure of 6G allows agents to interact with real-time data sources, such as IoT devices, to dynamically adjust their actions in response to changing environments. For example, in a smart city, an agent might plan a series of actions involving traffic sensors and weather data to optimize traffic flow.

In summary, while planning is a less predictable but highly promising design pattern in agentic AI workflows, it holds great potential, especially when combined with the capabilities of 6G networks. Its ability to dynamically decide on sequences of actions makes it invaluable for handling complex, multi-step tasks that require flexibility and adaptability.

Multi-Agent Collaboration in Agentic AI Workflows

Multi-agent collaboration[32] is the final key design pattern in the agentic AI framework, focusing on decomposing complex tasks into subtasks, each handled by specialized agents. This mirrors human team structures, where individuals with distinct expertise collaborate[33] on different aspects of a project, such as software development or research. In AI, this involves prompting one or more LLMs to simulate different agents, each tasked with a specific role, such as coding, project management, or quality assurance.

Though counterintuitive—since all agents may stem from the same LLM—multi-agent collaboration offers several advantages:

- *Superior Performance*: Studies, such as those in the AutoGen[34] framework, show that multi-agent systems outperform single agents by dividing tasks into manageable subtasks, improving task-specific focus.
- *Enhanced Focus*: By allocating specific roles to agents, such as a software engineer or security expert, the LLM can concentrate on distinct subtasks, optimizing each for clarity, efficiency, or security as needed.
- *Task Decomposition:* From a development perspective, multi-agent systems provide a framework for breaking down large tasks, similar to splitting a software program into processes or threads, making complex tasks easier to manage and execute.

In multi-agent workflows, agents operate independently, often invoking other agents for assistance or combining planning, tool use, and reflection within their roles. These dynamic interactions create complex, layered workflows where agents collaborate, refine outputs, and delegate tasks. As multi-agent systems advance, their ability to handle intricate tasks will improve, taking on functions traditionally performed by human teams.

Like planning, multi-agent collaboration can yield unpredictable results due to the autonomy of each agent. This unpredictability arises from decentralized decision-making, which can make outcomes difficult to foresee. However, emerging frameworks like AutoGen[34], Crew AI, and LangGraph[35] provide structured environments to manage these systems, enabling more reliable multi-agent collaboration. While early systems like ChatDev[32] may not always produce perfect results, they represent significant progress in AI collaboration.

Discussion

Agentic AI workflows offer a transformative approach to enhancing the capabilities of large language models (LLMs), enabling them to operate with greater autonomy, flexibility, and precision. By integrating design patterns such as reflection, tool use, planning, and multi-agent collaboration, we can address some of the core limitations of LLMs, including their tendency to generate hallucinations and the difficulty they face in handling complex, multi-step tasks.

Reflection and tool use have proven to be reliable and effective, significantly improving the quality and accuracy of LLM outputs. Reflection allows models to iteratively critique and refine their results, while tool use empowers models to access real-time data and external resources beyond their training set. These patterns are crucial for ensuring that LLMs can perform well in real-world applications that demand precision, such as coding, research, and decision-making.

However, planning and multi-agent collaboration present a more complex challenge. While these patterns provide a powerful framework for solving intricate problems by allowing LLMs to sequence tasks autonomously or collaborate with specialized agents, their unpredictability limits their current utility. Planning often results in unexpected steps, and managing the interactions between multiple agents can lead to unforeseen complications. Nevertheless, as AI systems and frameworks mature, such as AutoGen[34], LangGraph[35], and Crew AI[36], these limitations are expected to diminish, unlocking new potential for multi-agent systems and dynamic task management.

The integration of these agentic workflows into 6G networks[37] opens up further possibilities. With high-speed, low-latency infrastructure, 6G networks enable real-time collaboration[38], data access, and tool integration, all of which are vital for multi-agent systems to operate effectively. Use cases, such as smart healthcare, autonomous driving, and supply chain optimization, can benefit significantly from these advancements, as agents can collaborate across distributed environments, accessing real-time IoT data and databases to make timely, data-driven decisions.

As we move forward, a key area of focus will be developing robust agent memory systems, allowing agents to retain context from previous interactions and decisions. Memory plays a critical role in enhancing multi-agent collaboration, ensuring that agents can reference past outcomes and optimize their workflows based on prior experiences. The evolution of memory architectures will be essential to making agentic workflows more reliable and adaptable in real-world scenarios.

Conclusion

Agentic AI workflows present a promising future for the development of large language models, offering a structured approach to improve their efficiency, autonomy, and ability to handle complex tasks. Reflection and tool use have already demonstrated their effectiveness in enhancing the performance of LLMs in real-world applications while planning and multi-agent collaboration offer significant potential, though they require further refinement to overcome current challenges.

The upcoming era of 6G networks provides an ideal infrastructure for deploying these agentic workflows, supporting high-speed, real-time collaboration and access to distributed data sources. This advancement, combined with ongoing improvements in agent memory and multi-agent frameworks, will enable more sophisticated AI systems capable of tackling increasingly complex, dynamic environments.

In summary, while agentic design patterns are still evolving, their integration into AI systems represents a key step toward more autonomous, flexible, and intelligent generative models. As these workflows mature, we expect them to play a pivotal role in revolutionizing industries from healthcare to autonomous systems, ultimately driving AI capabilities to new heights.

References

[1] E. Brynjolfsson, D. Li, and L. R. Raymond, "Generative AI at Work," *Stanford Grad. Sch. Bus. Work. Pap.*, vol. No. 4141, 2023.

[2] S. Minaee *et al.*, "Large Language Models: A Survey".

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, {NAACL-HLT} 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, 2018, pp. 4171–4186. doi: 10.18653/V1/N19-1423.

[4] H. Touvron *et al.*, "LLaMA: Open and Efficient Foundation Language Models," *CoRR*, 2023, [Online]. Available: http://arxiv.org/abs/2302.13971

[5] S. Bubeck *et al.*, "Sparks of Artificial General Intelligence: Early experiments with {GPT-4}," 2023.

[6] Y. Bai *et al.*, "Sequential Modeling Enables Scalable Learning for Large Vision Models," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 22861–22872. doi: 10.1109/CVPR52733.2024.02157.

[7] R. Girdhar *et al.*, "ImageBind One Embedding Space to Bind Them All," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15180–15190. doi: 10.1109/CVPR52729.2023.01457.

[8] S. Pichar and D. Hassabis, "Introducing Gemini: our largest and most capable AI model," google. [Online]. Available: https://deepmind.google/technologies/gemini/#introduction

[9] OpenAI, "Creating video from text." [Online]. Available: https://openai.com/sora

[10] Z. Xu, S. Jain, and M. S. Kankanhalli, "Larger and more instructable language models become less reliabl," *ArXiv*, vol. abs/2401.1, 2024, [Online]. Available:

https://api.semanticscholar.org/CorpusID:267069207
[11] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, “Detecting hallucinations in large language models using semantic entropy,” *Nature*, vol. 630, no. 8017, pp. 625–630, 2024, doi: 10.1038/s41586-024-07421-0.
[12] L. Zhou, W. Schellaert, F. Martínez-Plumed, Y. Moros-Daval, C. Ferri, and J. Hernández-Orallo, “Larger and more instructable language models become less reliable,” *Nature*, vol. 634, no. October, 2024, doi: 10.1038/s41586-024-07930-y.
[13] P. Lewis *et al.*, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, no. NeurIPS, 2020.
[14] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, “Gorilla: Large Language Model Connected with Massive APIs,” *ArXiv*, pp. 1–18, 2023, [Online]. Available: http://arxiv.org/abs/2305.15334
[15] A. Madaan *et al.*, “SELF-REFINE: iterative refinement with self-feedback,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, in NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., 2024.
[16] J. Li *et al.*, “Banishing LLM Hallucinations Requires Rethinking Generalization,” no. 1972, pp. 1–14, 2024, [Online]. Available: http://arxiv.org/abs/2406.17642
[17] K. Trichias, A. Kaloxylos, and C. Willcock, “6G Global Landscape: A Comparative Analysis of 6G Targets and Technological Trends,” in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2024, pp. 1–6. doi: 10.1109/EuCNC/6GSummit60053.2024.10597064.
[18] X. Huang, Y. Tang, J. Li, N. Zhang, and X. Shen, “Toward Effective Retrieval Augmented Generative Services in 6G Networks,” *IEEE Netw.*, vol. 38, no. 6, pp. 459–467, 2024, doi: 10.1109/MNET.2024.3436670.
[19] L. Ale, N. Zhang, S. A. King, and D. Chen, “Empowering generative AI through mobile edge computing,” *Nat. Rev. Electr. Eng.*, vol. 1, no. 7, pp. 478–486, 2024, doi: 10.1038/s44287-024-00053-6.
[20] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: language agents with verbal reinforcement learning,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, in NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., 2024.
[21] Z. Gou *et al.*, “CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing,” 2024. [Online]. Available: https://arxiv.org/abs/2305.11738
[22] J. Nash, “Non-Cooperative Games,” *Ann. Math.*, vol. 54, no. 2, pp. 286–295, 1951, Accessed: Dec. 29, 2024. [Online]. Available: http://www.jstor.org/stable/1969529
[23] L. Ale, N. Zhang, S. A. King, and J. Guardiola, “Spatio-Temporal Bayesian Learning for Mobile Edge Computing Resource Planning in Smart Cities,” *ACM Trans. Internet Technol.*, vol. 21, no. 3, Jun. 2021, doi: 10.1145/3448613.
[24] Z. Yang *et al.*, “MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action,” 2023. [Online]. Available: https://arxiv.org/abs/2303.11381
[25] S. Gao *et al.*, “Efficient Tool Use with Chain-of-Abstraction Reasoning,” 2024. [Online]. Available: https://arxiv.org/abs/2401.17464
[26] L. Ale, N. Zhang, H. Wu, D. Chen, and T. Han, “Online Proactive Caching in Mobile Edge Computing Using Bidirectional Deep Recurrent Neural Network,” *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5520–5530, Jun. 2019.
[27] M.-I. Corici, F. Eichhorn, V. Gowtham, T. Magedanz, E.-R. Modroiu, and F. Schreiner, “How Organic Networking meets 6G Campus Network Management Challenges,” in *2023 26th Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*,

2023, pp. 169–173. doi: 10.1109/ICIN56760.2023.10073499.
[28] L. Ale, N. Zhang, X. Fang, X. Chen, S. Wu, and L. Li, “Delay-Aware and Energy-Efficient Computation Offloading in Mobile Edge Computing Using Deep Reinforcement Learning,” *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 3, pp. 881–892, Sep. 2021.
[29] X. Huang *et al.*, “Understanding the planning of LLM agents: A survey,” 2024. [Online]. Available: https://arxiv.org/abs/2402.02716
[30] J. Wei *et al.*, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” *Adv. Neural Inf. Process. Syst.*, vol. 35, no. NeurIPS, pp. 1–14, 2022.
[31] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, “HuggingGPT: solving AI tasks with chatgpt and its friends in hugging face,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, in NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., 2024.
[32] C. Qian *et al.*, “ChatDev: Communicative Agents for Software Development,” 2024. [Online]. Available: https://arxiv.org/abs/2307.07924
[33] S. Hong *et al.*, “MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework,” 2024. [Online]. Available: https://arxiv.org/abs/2308.00352
[34] Q. Wu *et al.*, “AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation,” 2023. [Online]. Available: https://arxiv.org/abs/2308.08155
[35] J. Wang and Z. Duan, “Agent AI with LangGraph: A Modular Framework for Enhancing Machine Translation Using Large Language Models,” 2024. [Online]. Available: https://arxiv.org/abs/2412.03801
[36] “crewAI: Framework for orchestrating role-playing, autonomous AI agents. By fostering collaborative intelligence, CrewAI empowers agents to work together seamlessly, tackling complex tasks.” Accessed: Dec. 29, 2024. [Online]. Available: https://github.com/crewAIInc/crewAI
[37] M. Corici, F. Eichhorn, and T. Magedanz, “Organic 6G Continuum Architecture: A Uniform Control Plane Across Devices, Radio, and Core,” *IEEE Netw. Lett.*, vol. 6, no. 1, pp. 11–15, 2024, doi: 10.1109/LNET.2023.3338363.
[38] C.-X. Wang *et al.*, “On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds,” *IEEE Commun. Surv. Tutorials*, vol. 25, no. 2, pp. 905–974, 2023, doi: 10.1109/COMST.2023.3249835.

Author contributions

All authors contributed equally to the preparation of this manuscript.

Competing interests statement

The authors declare no competing interests.