Uma Revisão Sistemática Sobre o Uso de LLMs Multiagentes Colaborativos Como Uma Estratégia Para Reduzir a Taxa de Alucinações

Augusto V. Diogenes[1], **Jordan P. Mesquita**[1], **Gilardo B. da Silva**[1]

[1]Câmpus de Quixadá – Universidade Federal do Ceará (UFC)
Av. José de Freitas Queiroz, 5003 – CEP 63902-580 – Quixadá – CE – Brasil

augustovnd@alu.ufc.br, gilardobento@alu.ufc.br, jordanpinheiro@alu.ufc.br

Resumo. *Large Language Models representam o estado da arte em inteligência artificial. No entanto, sua utilidade é limitada pela tendência de alucinar. A colaboração multiagente é proposta como um mecanismo para reduzir as taxas de alucinação, e este artigo busca validar essa proposta por meio de uma revisão dos estudos que desenvolveram LLMs multiagentes e registraram suas taxas de alucinação. Identificamos que o uso de LLMs multiagentes reduziu a taxa de alucinação por no mínimo 43% e no máximo 90%.*

Palavras-chave. *Large Language Models; Inteligência Artificial; Sistemas Multiagentes.*

1. Introdução

Large Language Models (LLMs) se destacam como um dos maiores avanços recentes no campo de inteligência artificial generativa e de agentes inteligentes. No entanto, a utilidade de LLMs é limitada devido a sua incidência a alucinações – isto é, a geração de informações falsas apresentadas ao usuário ou a outros agentes como se fossem verdadeiras [Alansari and Luqman 2026].

O artigo ”When One LLM Drools, Multi-LLM Collaboration Rules”[Feng et al. 2025] propõe que o uso de LLMs multiagentes colaborativos, ao invés de LLMs puros, pode reduzir a incidência alucinações e melhorar desempenho em sistemas de inteligência artificial. Este artigo também propôs múltiplas diferentes protocolos de colaboração de LLMs, incluindo colaboração a nível de API, a nível de texto, a nível logit, e a nível de pesos.

No interesse de contribuir a futuros desenvolvimentos na área de sistemas inteligentes, essa pesquisa busca analisar e discutir a literatura científica para determinar se os dados de pesquisas reais corroboram a proposta de Feng et al. 2025. Tal análise será feita através de uma revisão sistemática da literatura, focando em trabalhos que discutem a implementação prática de sistemas de LLMs multiagentes e trazem dados sobre a taxa de alucinação desses sistemas.

1.1. Objetivos Específicos

Em função de analisar o desempenho prático de sistemas multiagentes de LLMs, essa pesquisa também busca:

- Identificar os diferentes protocolos de colaboração de LLMs utilizados nas pesquisas relevantes ao tema.
- Comparar as taxas de alucinações desses diferentes protocolos.
- Discutir quais protocolos são promissores para futuras pesquisas.

2. Metodologia

A revisão sistemática desse estudo inicia-se com a formulação de uma estratégia de pesquisa. Determina-se que as bases de dados IEEE e ACM, acessadas através do sistema CAPES, serão usadas como fonte de dados. Seleciona-se a ferramenta online Parsif.al para organizar os artigos levantados pela busca, permitindo a seleção manual de artigos relevantes e a remoção de duplicatas. A string de busca ”(”LLM”OR ”Large Language Model”) AND (”Multi-Agent”OR ”Multiagent”) AND (”Hallucination”)”é usada para buscar artigos nas bases. Finalmente, estabelece-se os critérios da pesquisa. Esses são:

- O artigo deve estar dentro do escopo da implementação prática de sistemas com múltiplos agentes LLM colaborativos.
- O artigo deve trazer dados sobre a incidência de alucinações de seus sistemas.
- O artigo não pode ser literatura secundária.

Uma vez que a estratégia e os critérios da pesquisa estão estabelecidos, prossegue-se à busca e seleção de artigos, que é feita seguindo os passos de identificação, triagem, elegibilidade, e inclusão. Esses passos são descritos na figura 1.

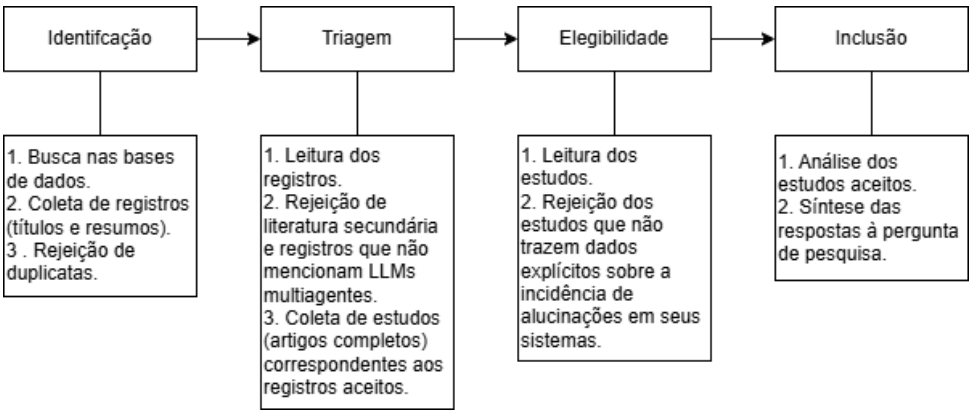

Figura 1. Os passos da seleção de artigos

3. Resultados e Discussão

182 artigos foram levantados pela busca inicial nas bases selecionada para essa pesquisa, e ao final do processo de seleção, 10 artigos foram aceitos para a análise e discussão.

Uma grande quantidade de artigos foi eliminada durante a avaliação de estudos completos, pois eles não traziam dados explícitos sobre a taxa de halucinações. Ao invés disso, os estudos avaliados traziam predominantemente dados sobre a acurácia dos modelos – uma métrica que, embora correlata à taxa de alucinações, não necessariamente corresponde a esta.

Por exemplo, é possível que um modelo com alta taxa de alucinação possua uma acurácia maior do que um modelo com baixa taxa de alucinação, pois o modelo que não alucina admite quando não é capaz de responder uma pergunta, enquanto o modelo que alucina tenta ”adivinhar”uma resposta certa [Kalai et al. 2025]. Por causa disso, somente dados explícitos sobre a taxa de alucinação foram considerados elegíveis para essa pesquisa.

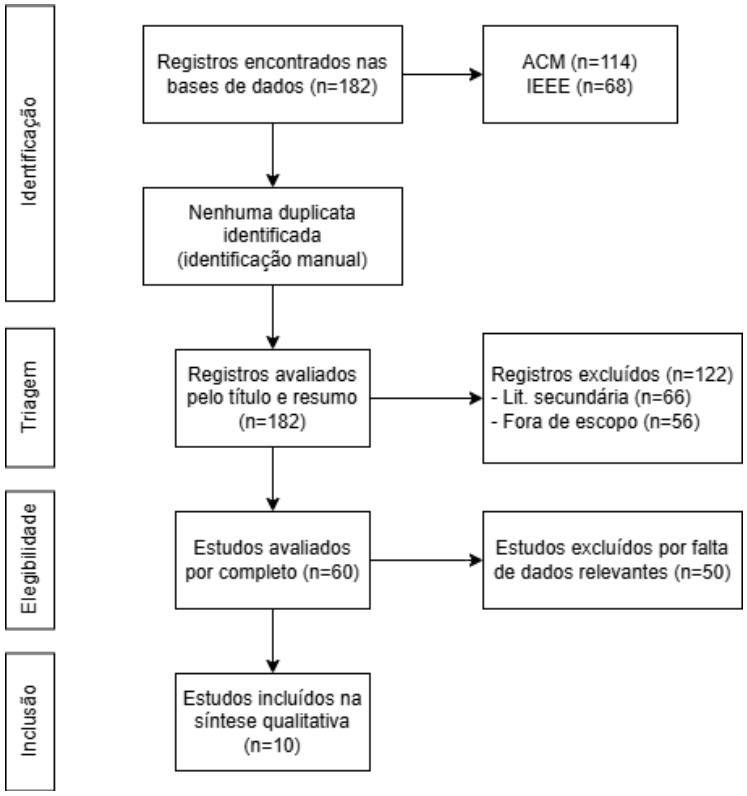

Figura 2. A seleção de artigos para essa pesquisa

O processo de busca e seleção de artigos para esse estudo é ilustrado pelo diagrama na figura 2.

Na literatura aceita para a análise, foram identificados 4 sistemas que utilizam os protocolos de colaboração a nível de API, e 6 sistemas que utilizam colaboração a nível de texto, assim como definidos em Feng et al. 2025. Esses sistemas são descritos nas subsessões 3.1 e 3.2. Não foram encontrados sistemas de colaboração a nível de peso ou logits.

3.1. Sistemas de Colaboração a Nivel de API

O estudo ”Dynamic Synthetic Data Generation through Human-Guided Multi-Agent Collaboration”[Prakash et al. 2025] criou um sistema com o framework LangGraph para a criação de dados sintéticos. Esse sistema possui um agente orquestrador que roteia chamadas a 5 agentes especializados: um que recebe entradas de usuário, um que organiza entrada em tópicos e palavras-chave, um que define os atributos dos dados sintéticos, um que define as variações nos dados sintéticos, e um que gera os dados sintéticos. O agente orquestrador chama vários desses agentes em sequência para realizar a tarefa do sistema. Todos os agentes possuem memória compartilhada. O sistema como um todo atingiu uma taxa de alucinação de 1%.

O estudo ”CowNet-AI: A Multi-Agent Decision Support Framework for Social Network–Driven Welfare Insights in Dairy Cattle”[Shanteer et al. 2026] criou um sistema – também com o framework LangGraph – que cria relatórios sobre o bem-estar de animais em fazendas pecuárias a partir de dados de sensores de monitoramento. Esse sistema

possui um agente supervisor que recebe entradas de usuário e chama 6 outros agentes especializados com memória compartilhada. Esses são: um agente que recupera dados de sensores, um que organiza dados em grafos e computa métricas de bem-estar, um que manipula a estrutura de grafos para criar cenários hipotéticos, um que pesquisa repositórios externos para formular explicações sobre os dados, grafos, e métricas de bem-estar, e um que condensa todas essas informações em uma resposta para o usuário. A sequência de chamada desses agentes é determinada pelo orquestrador. O sistema completo apresenta uma taxa de alucinação de 5%, enquanto um LLM simples apresenta uma taxa de 40% na mesma tarefa.

O estudo "Multi-Agent Brainstorming for Interpreting and Mitigating Hallucination in Multimodal-LLM"[Lin et al. 2026] criou um sistema de propósito geral que usa três agentes: dois jogadores e um moderador. Os dois jogadores tentam primeiro responder uma pergunta do usuário, e então o moderador roteia chamadas entre esses dois agentes para que eles possam discutir e ajustar suas respostas até que eles entrem em consenso. O sistema possui uma taxa de alucinação 43% menor do que LLMs simples.

O estudo "Explainable AI for Marketing Intelligence: Interpretable Multi-Agent Systems for Data-Driven Decision Support"[Rudra et al. 2026] criou um sistema com seis agentes especialistas com memória compartilhada. Esses agentes são: um que recebe e valida dados de entrada, um que idenfica atributos em dados na memória, um que gera hipóteses e previsões, um que gera relatórios de eventos e ações, um que valida ações e detecta alucinações, e um que chama intervenção humana para ações importantes. Esse sistema possui uma taxa de alucinação de 3,2%, 78% menos do que LLMs simples.

3.2. Sistemas de Colaboração a Nivel de Texto

O estudo "TriGen: A Semantic-Feedback Collaborative LLM Test Case Generation Model"[Han et al. 2025] criou um sistema para a geração de casos de teste utilizando agentes baseados em LLM. O primeiro agente recebe a entrada do usuário e identifica as entidades, ações, e resultados esperados dos casos de teste. O segundo agente recebe a saída do primeiro e acrescenta a essa informação os tipos de teste a serem realizados. O terceiro agente recebe a saída do segundo e a gera os passos de teste a serem implementados. A saída do terceiro agente é verificada por ferramentas text2vec e se ela é rejeitada, ela é retornada ao início. Se ela for aceita, ela é formatada e retornada ao usuário. A taxa de alucinação desse sistema é de 0,183%.

O estudo "STaRQ-Agent-Investigating Generalization in Knowledge Graph Question Answering via a Multi-Agent Collaborative Framework"[Jiang et al. 2026] criou um sistema que responde perguntas utilizando um grafo de conhecimento e quarto agentes LLM. O primeiro agente recebe a pergunta do usuário e identifica as entidades e relações relevantes a pergunta. O segundo agente recebe a saída do primeiro e gera um pseudocódigo para a busca no grafo de conhecimento. O terceiro agente recebe a saída do segundo e gera a consulta em SPARQL. O quarto agente inspeciona a consulta para detectar e remover alucinações. A consulta final é feita sobre o grafo e o resultado é retornado ao usuário. O sistema possui uma taxa de alucinação de 22%. A omissão do agente inspetor de consultas aumenta essa taxa para 34%.

O estudo "Self-Auditing LLMs for Bias and Hallucination Reduction Through Multi-Agent Frameworks"[Anointina et al. 2026] criou um sistema que avalia as respos-

tas de um agente LLM simples com 4 agentes LLM especializados. O primeiro desses avalia a resposta para detectar erros factuais. O segundo avalia a resposta para detectar viéses demográficos e linguagem discriminatória. O terceiro avalia a resposta para detectar inconsistências lógicas. Finalmente, o quarto estima o nível de confiança da resposta. Esses quatro agentes votam para aceitar a resposta original – retornando-a ao usuário – ou rejeitar a resposta, retornando-a para o agente original junto com sugestões de correção. Esse sistema possui uma taxa de alucinação de 6,1%. Sem os quatro agentes avaliadores, foi medida uma taxa de alucinação de 23,4%.

O estudo "MACV: A Specialized Multi-Agent and Consensus Framework for Reliable LLM Outputs"[More et al. 2026] criou um sistema modular chamado Multi-Agent Cross-Verification (MACV), projetado especificamente para mitigar alucinações em grandes modelos de linguagem (LLMs). Este framework coordena agentes heterogêneos com papéis especializados, começando pelo Primary Response Generator (PRG), que gera a resposta inicial e extrai alegações estruturadas para verificação. Em seguida, o Fact-Checking Agent (FCA) valida as informações em fontes externas como Wikipedia e PubMed, enquanto o Domain-Specific Validator (DSV) inspeciona conteúdos técnicos em áreas como finanças e medicina usando bases de dados estruturadas. O sistema conta ainda com um Adversarial Tester (AT), que analisa a resposta em busca de contradições internas e falhas de raciocínio lógico. Todas essas análises são integradas por um Mecanismo de Consenso que agrega os outputs dos agentes para emitir uma decisão final de aceitação ou rejeição da resposta, acompanhada de uma etiqueta de confiança. Nos testes realizados com os benchmarks SQuAD, SciQ e FinQA, o sistema demonstrou resultados significativos, reduzindo a taxa de alucinação de 50,0% (no modelo baseline) para 27,3%. Além disso, a acurácia das respostas aceitas subiu de 50,0% para 72,7%, representando uma melhoria relativa de 45% na confiabilidade do output. O estudo também revelou que 50% das rejeições foram causadas por alegações refutadas pelos agentes de checagem de fatos. Contudo, o sistema apresenta um trade-off de eficiência, exigindo um tempo de processamento aproximadamente 23 vezes maior que o modelo baseline devido ao rigor das verificações cruzadas.

O estudo "Cognitive SOC: Evidence-Backed Narrative Generation for Security Operations with Multi-Agent LLM Architecture"[Sheikhi et al. 2025] criou o sistema Cognitive SOC, um framework multiagente projetado para transformar alertas brutos de segurança em narrativas fundamentadas por evidências com cadeias de citação completas. A arquitetura do sistema baseia-se em um pipeline de três agentes especializados que colaboram entre si: o TriageAgent, responsável pela análise inicial dos alertas, classificação de severidade e geração de hipóteses testáveis; o EnrichmentAgent, que aumenta os alertas com inteligência contextual integrando fontes como o framework MITRE ATT&CK, regras Sigma e bancos de dados CVE; e o StoryWeaverAgent, encarregado de sintetizar as informações em sumários executivos e relatórios técnicos detalhados. Todo o fluxo de dados é gerenciado através de um Mapa de Evidências compartilhado que mantém a proveniência estrita de cada alegação, sendo validado por um componente de verificação de citações que calcula métricas de responsabilidade. Quanto aos resultados encontrados, os testes realizados com 1.000 alertas do dataset UNSW-NB15 demonstraram que o sistema alcançou uma taxa de alucinação excepcionalmente baixa de 0,1%, mantendo uma taxa de vinculação de evidências de 99,6% e cobertura de citações de 100%. Em comparação, uma abordagem de RAG de agente único operando sob as mesmas condições apresen-

tou uma taxa de alucinação de 11,8% e apenas 59.8% de vinculação de evidências. O Cognitive SOC gerou uma média de 3,8 artefatos de evidência e 5,4 citações por alerta, superando largamente os baselines estáticos e de agente único. O tempo médio para a geração de recomendações foi de 1,860 segundos, o que prova a viabilidade do sistema para uso prático em centros de operações de segurança (SOC). Uma análise qualitativa revelou que os raros casos flagged como alucinação eram, na verdade, inferências contextuais válidas baseadas em reconhecimento de padrões e conhecimento de domínio, o que reforça a confiabilidade do framework para tomada de decisões estratégicas e suporte a auditorias de conformidade.

O estudo "Hallucination-Free Automatic Question & Answer Generation for Intuitive Learning"[Wang and Katsaggelos 2025] criou um framework de geração multiagente projetado especificamente para mitigar alucinações na criação automática de questões de múltipla escolha (MCQs) para fins educacionais. O sistema baseia-se em um pipeline iterativo de dois agentes principais: um Gerador, encarregado de produzir a pergunta, as alternativas e a respectiva explicação, e um Detector, que avalia o conteúdo gerado contra uma tipologia de quatro tipos de alucinação (inconsistências de raciocínio, questões insolúveis, erros factuais e erros matemáticos). Este detector fornece feedback direcionado ao gerador, que revisa a questão em um ciclo iterativo que utiliza raciocínio contrafactual e Chain-of-Thought (CoT) até que a pontuação de alucinação caia abaixo de um limite aceitável. O sistema redefine a geração de questões como uma tarefa de otimização, buscando minimizar o risco de erros enquanto maximiza a validade pedagógica e a eficiência de custos. Quanto ao resultado do que ele encontrou, os experimentos realizados com questões de STEM alinhadas ao nível AP demonstraram que o sistema reduziu as taxas de alucinação em mais de 90% em comparação com a geração baseline. O estudo revelou que as taxas de erro exibem uma tendência de queda acentuada conforme o número de iterações aumenta, convergindo para quase 0% após sete rodadas de refinamento. Observou-se que mesmo uma única iteração já é capaz de reduzir a alucinação em aproximadamente 50%. Além disso, o framework provou ser altamente econômico ao utilizar o modelo GPT-4.1-nano, que apresenta um custo 11 vezes menor que modelos de raciocínio de ponta (como o GPT-o3-mini), mantendo um desempenho superior na eliminação de falhas lógicas e factuais sem sacrificar o valor educacional ou a criatividade das questões geradas.

4. Conclusão

Uma vez que o objetivo do estudo foi revisar e analizar as principais maneiras de reduzir a taxa de alucinação com o uso de LLMs Multiagentes Colaborativos, os resultados foram satisfatorios na medida em que revelaram uma melhoria significativa nas taxas de alucinação. Entre os estudos seus sistemas com LLMs simples realizando as mesmas tarefas, foi observada uma redução de alucinação de 43% [Lin et al. 2026] até 90% [Wang and Katsaggelos 2025]. No entanto, mesmo reduzidas, as taxas de alucinação ainda permanecem significativas em vários estudos, com alguns [Jiang et al. 2026, More et al. 2026] apresentando taxas na ordem de dois dígitos de porcentagem. No geral, ao comparar os dois níveis de colaboração, nota-se que, em média, a colaboraçao a nivel de API obtem uma melhora na performance melhor e mais consistente na comparação dos estudos analisados. A analise feita nos permite concluir que a área de LLMs ainda está amadurecendo, e que ainda ha ampla oportunidade para melhoria e optimizaçao,

bem como novas descobertas nos metodos de planejamento e execução de projetos. Esperamos com o nosso estudo contribuir para o cenario atual do trabalho mais efetivo e automatizado com o uso de IA. Como limitação, a falta de artigos usando os outros níveis de colaboração deixou lacunas que esperamos poder corrigir em trabalhos futuros, ou mesmo contribuir para a sua resolução. Mais detalhadamente, observou-se que os sistemas operando a nível de API fundamentam sua eficácia no uso de agentes orquestradores ou supervisores que coordenam o fluxo de trabalho entre múltiplos especialistas, todos compartilhando uma memória comum para manter a coerência. Essa estrutura permitiu, por exemplo, que sistemas de suporte à decisão reduzissem a taxa de alucinação de 40% para apenas 5%. Por outro lado, a colaboração a nível de texto demonstrou ser particularmente robusta ao empregar ciclos iterativos de refinamento e mecanismos de consenso ou votação entre os agentes. Estudos mostraram que o erro factual pode convergir para quase 0% após rodadas sucessivas de verificação, embora isso traga um trade-off de eficiência, podendo elevar o tempo de processamento em até 23 vezes em comparação com modelos isolados. Essa distinção técnica reforça que, enquanto o nível de API oferece uma orquestração consistente, o nível de texto provê um rigor verificador superior, essencial para tarefas onde a precisão absoluta é mandatória.

Referências

Alansari, A. and Luqman, H. (2026). Large language models hallucination: A comprehensive survey.

Anointina, Bamini, A., and Jegan, C. (2026). Self-auditing llms for bias and hallucination reduction through multi-agent frameworks. pages 1321–1328.

Feng, S., Ding, W., Liu, A., Wang, Z., Shi, W., Wang, Y., Shen, Z., Han, X., Lang, H., Lee, C.-Y., Pfister, T., Choi, Y., and Tsvetkov, Y. (2025). When one llm drools, multi-llm collaboration rules.

Han, P., Chen, Y., and Wang, J. (2025). Trigen: A semantic-feedback collaborative llm test case generation model. *IEEE Access*, 13:209556–209574.

Jiang, L., Banerjee, D., and Usbeck, R. (2026). Starq-agent-investigating generalization in knowledge graph question answering via a multi-agent collaborative framework. *IEEE Access*, 14:59158–59170.

Kalai, A. T., Nachum, O., Vempala, S. S., and Zhang, E. (2025). Why language models hallucinate. *ArXiv*, abs/2509.04664.

Lin, Z., Niu, Z., Wang, Z., Xu, X., Liu, H., and Gao, H. (2026). Multi-agent brainstorming for interpreting and mitigating hallucination in multimodal-llm. pages 3006–3010.

More, R., Varma, S., and Varma, N. (2026). Macv: A specialized multi-agent and consensus framework for reliable llm outputs. In *2026 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, pages 1326–1331.

Prakash, C., Sinha, A., and Selvakumar, P. (2025). Dynamic synthetic data generation through human-guided multi-agent collaboration. In *2025 5th International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–6.

Rudra, A., Karumuri, M. S., and Agrawal, M. (2026). Explainable ai for marketing intelligence: Interpretable multi-agent systems for data-driven decision support. In *SoutheastCon 2026*, pages 1–5.

Shanteer, T., Sailunaz, K., and Neethirajan, S. R. (2026). Cownet-ai: A multi-agent decision support framework for social network–driven welfare insights in dairy cattle. *IEEE Access*, 14:64882–64897.

Sheikhi, S., Kostakos, P., and Loven, L. (2025). Cognitive soc: Evidence-backed narrative generation for security operations with multi-agent llm architecture. In *2025 IEEE International Conference on Big Data (BigData)*, pages 7027–7036.

Wang, N. X. and Katsaggelos, A. K. (2025). Hallucination-free automatic question answer generation for intuitive learning. In *2025 IEEE International Conference on Image Processing Workshops (ICIPW)*, pages 464–468.