Alucinações em Modelos de Linguagem de Grande Escala: Uma Revisão Sistemática da Literatura

Andressa Silva Pereira[1], Simone Azevedo Bandeira de Melo Aquino[1], Emmanuel Silva Xavier[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Maranhão (IFMA) – Campus Imperatriz

andressasp68@gmail.com, simonebandeira@ifma.edu.br, emmanuel.xavier@ifma.edu.br

Abstract. Large Language Models (LLMs) generate fluent, coherent text but frequently produce factually incorrect or fabricated content, a phenomenon known as hallucination. This paper presents a systematic literature review of 32 studies published between 2020 and 2025, guided by four research questions covering hallucination types, mitigation strategies, structural gaps, and coverage of low-resource languages. Results are organized into five categories: retrieval-augmented generation (RAG), reinforcement learning from human feedback (RLHF), self-verification mechanisms, architectural interventions, and evaluation benchmarks. Factual hallucinations account for approximately 78% of the reviewed research effort, while semantic and linguistic types remain largely unaddressed. Three critical gaps are identified: absence of standardized benchmarks for cross-study comparison, lack of multi-strategy integration, and scarcity of studies targeting low-resource languages such as Portuguese. These gaps represent concrete research opportunities for the Brazilian NLP community.

Resumo. Modelos de Linguagem de Grande Escala (LLMs) geram texto fluente e coerente, mas frequentemente produzem conteúdo factualmente incorreto ou fabricado, fenômeno denominado alucinação. Este artigo apresenta uma revisão sistemática da literatura abrangendo 32 estudos publicados entre 2020 e 2025, orientada por quatro questões de pesquisa que cobrem tipos de alucinação, estratégias de mitigação, lacunas estruturais e cobertura de línguas de menor recurso. Os resultados são organizados em cinco categorias: geração aumentada por recuperação (RAG), aprendizado por reforço com feedback humano (RLHF), mecanismos de auto-verificação, intervenções arquiteturais e benchmarks de avaliação. Alucinações factuais concentram aproximadamente 78% do esforço de pesquisa, enquanto os tipos semântico e linguístico permanecem subexplorados. Três lacunas críticas são identificadas: ausência de benchmarks padronizados para comparação entre estudos, carência de integrações multi-estratégia e escassez de trabalhos voltados a línguas de menor recurso, como o português, representando oportunidades concretas para a comunidade brasileira de PLN.

1. Introdução

LLMs como ChatGPT, Gemini e Llama 3 estão integrados a fluxos de trabalho em saúde, direito, educação e jornalismo, processando bilhões de consultas diárias. Esse

alcance torna crítica uma limitação bem documentada: a tendência dos modelos de gerar conteúdo factualmente incorreto, incoerente ou simplesmente inventado. Ji et al. (2023) denominam esse comportamento de alucinação e atribuem sua origem ao processo autorregressivo de geração, no qual o modelo seleciona tokens com base em probabilidade condicional aprendida, sem mecanismo nativo de verificação factual.

As consequências práticas são documentadas e graves. No campo jurídico, casos de advogados citando jurisprudências inexistentes geradas por LLMs foram registrados em tribunais norte-americanos em 2023, resultando em sanções processuais e danos à credibilidade profissional. Na área da saúde, pesquisas identificaram alucinações em instruções de medicamentos geradas por LLMs mesmo com modelos ajustados para o domínio clínico, evidenciando o risco em contextos de alta responsabilidade. No campo educacional, Bang et al. (2023) identificaram erros sistemáticos de raciocínio causal e factual em avaliações do ChatGPT em 23 tarefas distintas, incluindo resolução de problemas matemáticos e perguntas históricas. Tonmoy et al. (2024) posicionam a mitigação de alucinações como um dos problemas abertos mais urgentes em IA generativa, destacando que a adoção em larga escala de LLMs sem controle adequado de factualidade representa um risco sistêmico para a qualidade da informação circulante na sociedade.

O fenômeno é agravado pela característica de fluência dos LLMs: respostas alucinatórias são tipicamente bem construídas gramaticalmente e apresentadas com confiança aparente, tornando difícil para usuários não especialistas identificar erros sem verificação independente. Esse aspecto distingue as alucinações de LLMs de outros tipos de erros computacionais e justifica a preocupação crescente com a confiabilidade desses sistemas em contextos críticos.

No contexto brasileiro, a relevância do tema é reforçada pela rápida adoção de assistentes baseados em LLMs em serviços públicos, plataformas educacionais e aplicações de saúde. Medeiros e Oliveira (2025) investigaram especificamente o comportamento de modelos RAG em português no SEMISH 2025, identificando variações significativas de desempenho entre modelos multilíngues e modelos especializados, o que indica que os padrões de alucinação documentados em inglês não se transferem diretamente para o português. Essa lacuna reforça a necessidade de investigações sistemáticas voltadas ao contexto linguístico brasileiro.

Este artigo é orientado por quatro questões de pesquisa (QP): (QP1) Quais tipos de alucinações em LLMs são identificados na literatura e como se distribuem no esforço de pesquisa? (QP2) Quais são as principais estratégias de mitigação propostas e quais suas contribuições e limitações? (QP3) Que lacunas estruturais existem na literatura? (QP4) Como a literatura cobre línguas de menor recurso, especialmente o português? A contribuição principal é uma revisão sistemática de 32 estudos (2020–2025) que categoriza os tipos de alucinação, mapeia suas causas e sintetiza criticamente as estratégias de mitigação propostas. A análise comparativa das abordagens identifica convergências, divergências e lacunas estruturais, com ênfase nas oportunidades de pesquisa para línguas de menor recurso.

2. Fundamentação Teórica

2.1. Large Language Models

LLMs são redes neurais de grande escala baseadas na arquitetura Transformer [Vaswani et al. 2017], treinadas em corpora de centenas de bilhões de tokens por objetivos autorregressivos ou de mascaramento. Wolf et al. (2020) documentaram que bibliotecas como Hugging Face Transformers tornaram modelos pré-treinados amplamente acessíveis, acelerando sua adoção em pesquisa e indústria. O processo de pré-treinamento, embora responsável pela ampla capacidade de generalização dos LLMs, é também a raiz do problema das alucinações: ao aprender associações estatísticas entre tokens, o modelo pode gerar afirmações plausíveis que não correspondem a nenhuma fonte factual verificável.

Modelos com centenas de bilhões de parâmetros, como o GPT-4 e o Llama 3 70B, demonstram capacidades emergentes notáveis em raciocínio e seguimento de instruções. Entretanto, Ouyang et al. (2022) mostraram que o tamanho do modelo não é o fator mais determinante na qualidade das respostas: um modelo de 1,3B de parâmetros alinhado via RLHF superou o GPT-3 (175B) em avaliações humanas, evidenciando que alinhamento e factualidade são desafios ortogonais ao simples escalonamento. Essa evidência tem implicações práticas relevantes: sugere que investir em técnicas de alinhamento e verificação pode ser mais eficiente do que simplesmente aumentar a escala do modelo.

2.2. Tipologia das Alucinações

Ji et al. (2023) propõem uma taxonomia tripartite amplamente adotada na literatura. Alucinações factuais referem-se à geração de afirmações contraditas por fatos verificáveis, como datas incorretas, nomes fabricados ou eventos inexistentes. Alucinações semânticas envolvem inconsistências de sentido intra-resposta: o modelo produz afirmações individualmente plausíveis, mas mutuamente contraditórias ou incoerentes com o contexto fornecido [Mundler et al. 2023]. Alucinações linguísticas correspondem a construções gramaticalmente corretas, porém semanticamente vazias ou circulares. Huang et al. (2023) acrescentam uma dimensão de fidelidade à fonte, relevante em tarefas de sumarização e RAG, na qual o modelo gera conteúdo não suportado pelo documento de entrada, categoria operacionalizada por Maynez et al. (2020) e Min et al. (2023).

A distinção entre esses tipos tem implicações diretas para a escolha da estratégia de mitigação mais adequada. Alucinações factuais são as mais diretamente endereçadas por técnicas de RAG e verificação externa, que podem confrontar as afirmações do modelo com bases de conhecimento verificáveis. Alucinações semânticas, por sua vez, exigem mecanismos de consistência interna, como os propostos por Mundler et al. (2023) e Du et al. (2023). Alucinações linguísticas são as menos estudadas e as mais difíceis de operacionalizar, pois sua detecção frequentemente requer avaliação humana especializada.

3. Metodologia

Esta revisão segue o protocolo de Revisão Sistemática da Literatura (RSL) proposto por Kitchenham e Charters (2007), sendo orientada pelas quatro questões de pesquisa enunciadas na Seção 1. As buscas foram conduzidas nas bases Google Scholar, Semantic Scholar, ACL Anthology e ACM Digital Library com as strings: "LLM hallucination", "language model factuality", "hallucination mitigation" e "AI

hallucination detection". O escopo temporal abrangeu publicações de 2020 a 2025, cobrindo o período desde a popularização do GPT-3 até os modelos de última geração.

Foram aplicados critérios de inclusão: (i) foco explícito em alucinações de LLMs ou em sua mitigação; (ii) metodologia experimental ou de revisão claramente descrita; (iii) publicação em venues de reconhecimento científico (ACL, EMNLP, NeurIPS, AAAI, ICML ou ACM Computing Surveys) ou, no caso de preprints arXiv, presença de citação substancial na literatura primária dos venues listados — critério adotado para refletir a dinâmica de publicação da área de NLP, em que avanços relevantes frequentemente circulam primeiramente como preprints antes da revisão formal. Critérios de exclusão eliminaram trabalhos sem texto completo disponível, publicações de divulgação sem base experimental e estudos anteriores a 2020.

A triagem foi conduzida de forma independente pelos três coautores. Divergências na classificação foram resolvidas por rodadas de consenso, nas quais os avaliadores justificaram suas decisões; o acordo final foi verificado antes da inclusão definitiva de cada estudo. Não foi utilizado software dedicado de RSL, mas os registros de triagem foram sistematizados em planilha compartilhada. A Tabela 2 apresenta o fluxo PRISMA adaptado do processo de seleção.

Tabela 2. Fluxo de seleção de estudos (adaptado de PRISMA)

Etapa	Descrição	Registros
Identificação	Buscas nas bases Google Scholar, Semantic Scholar, ACL Anthology e ACM Digital Library com strings: "LLM hallucination", "language model factuality", "hallucination mitigation", "AI hallucination detection" (2020–2025)	218 registros identificados
Triagem	Remoção de duplicatas e triagem por título e resumo pelos três avaliadores independentes; divergências resolvidas por consenso	147 candidatos; 71 excluídos por duplicidade ou falta de relevância
Elegibilidade	Leitura do texto completo; aplicação de critérios de inclusão/exclusão; acordo interavaliadores verificado	76 avaliados integralmente; 44 excluídos
Inclusão	Estudos com foco explícito em alucinações de LLMs, metodologia experimental ou de revisão descrita e publicação em ACL, EMNLP, NeurIPS, AAAI, ICML, ACM Computing Surveys ou arXiv com citação substancial na literatura primária	32 estudos incluídos na RSL

Fonte: Elaborada pelos autores (2025).

Cada estudo foi categorizado segundo quatro dimensões: (i) tipo de alucinação abordada, seguindo a taxonomia de Ji et al. (2023); (ii) técnica de mitigação proposta, organizada nas cinco categorias identificadas na análise; (iii) metodologia experimental utilizada, incluindo datasets, métricas e modelos avaliados; e (iv) resultados principais reportados. Essa matriz de categorização subsidiou a construção da tabela comparativa e as análises qualitativas das seções subsequentes.

Os datasets mais recorrentes nos estudos analisados foram TruthfulQA [Lin et al. 2022], utilizado em experimentos de ITI e DoLa para medir respostas verídicas; FActScore [Min et al. 2023], para avaliação de precisão factual em geração longa; e Natural Questions, amplamente usado em experimentos de RAG. As métricas predominantes incluem precisão factual, taxa de alucinação por amostragem

(SelfCheckGPT), ROUGE/BLEU (com ressalvas quanto à sensibilidade a erros pontuais) e avaliação humana. Os modelos mais avaliados são GPT-3/GPT-4 (OpenAI), Llama 2/3 (Meta) e PaLM 2 (Google), com menor cobertura de modelos em português.

A Tabela 3 sintetiza as questões de pesquisa e as seções do artigo em que são respondidas:

Tabela 3. Questões de pesquisa e mapeamento para seções do artigo

ID	Questão de Pesquisa	Seção de Resposta
QP1	Quais tipos de alucinações em LLMs são identificados na literatura e como se distribuem no esforço de pesquisa?	Seções 2.2, 4 e 7
QP2	Quais são as principais estratégias de mitigação propostas e quais suas contribuições e limitações?	Seções 4 e 5
QP3	Que lacunas estruturais existem na literatura e quais direções de pesquisa se mostram mais promissoras?	Seções 5.6 e 7
QP4	Como a literatura cobre línguas de menor recurso, em especial o português, e que oportunidades existem para a comunidade brasileira de PLN?	Seções 1, 7 e 8

Fonte: Elaborada pelos autores (2025).

4. Resultados: Mapeamento das Abordagens

Os 32 estudos analisados distribuem-se entre cinco categorias de mitigação e uma categoria dedicada a avaliação, respondendo à QP2. Alucinações factuais (QP1) são o alvo em aproximadamente 78% dos trabalhos. Técnicas baseadas em recuperação externa e RLHF concentram o maior número de estudos, enquanto intervenções arquiteturais representam a linha mais recente e de maior crescimento desde 2022. A Tabela 1 sintetiza as categorias com trabalhos representativos, contribuições e limitações identificadas.

Tabela 1. Categorias de abordagens de mitigação de alucinações em LLMs: síntese comparativa

Categoria	Trabalhos Representativos	Principais Contribuições	Limitações Identificadas
RAG / Recuperação externa	Borgeaud et al. (2022); Nakano et al. (2021); Shuster et al. (2021); Lazaridou et al. (2022)	Alta redução de alucinações factuais; grounding verificável	Depende de bases atualizadas; latência adicional
RLHF / Feedback humano	Stiennon et al. (2020); Ouyang et al. (2022)	Reduz respostas incoerentes; melhora percepção de qualidade	Alto custo de anotação; difícil escalar
Auto-verificação (CoVe / SelfCheck)	Weng et al. (2023); Dhuliawala et al. (2023); Manakul et al. (2023)	Detecta inconsistências sem fonte externa; aplicável em caixa-preta	Pode propagar erros do modelo base
Debate multi-agente	Du et al. (2023)	Ganhos de precisão superiores a modelos isolados	Custo computacional multiplicado pelo número de agentes

Intervenção em ativações (ITI / DoLa)	Li et al. (2023); Chuang et al. (2023)	Melhora factualidade sem re-treinamento	Requer acesso às representações internas do modelo
Chain-of-Thought / Tree of Thoughts	Wei et al. (2022); Yao et al. (2023)	Reduz erros em raciocínio e planejamento	Aumenta comprimento das saídas; eficácia variável por tarefa
Calibração de confiança	Guo et al. (2022); Kadavath et al. (2022)	Reduz assertividade em respostas incorretas; sinaliza incerteza	Não elimina alucinações; dependente da calibração do modelo
Fine-tuning especializado	Lewkowycz et al. (2022)	Alta precisão em domínios específicos	Não generaliza; requer dados especializados

Fonte: Elaborada pelos autores (2025).

Os dados revelam que aproximadamente 78% dos estudos revisados focam em alucinações do tipo factual (QP1). As categorias semântica e linguística, embora definidas em surveys como Ji et al. (2023) e Huang et al. (2023), carecem de metodologias experimentais dedicadas. A predominância de abordagens na fase de inferência ou pós-geração, em contraposição a intervenções no pré-treinamento, é expressiva, sugerindo preferência por soluções compatíveis com modelos já implantados.

Em termos de distribuição temporal, observa-se um crescimento acentuado nas publicações sobre auto-verificação e intervenções arquiteturais a partir de 2022, período que coincide com a adoção em massa do ChatGPT e com o aumento da pressão pública por sistemas mais confiáveis. As técnicas de RAG, embora mais antigas em sua concepção, também experimentaram ressurgimento de interesse com a disponibilização de frameworks de implementação acessíveis como LangChain e LlamaIndex. Por outro lado, o RLHF mostra relativa estabilização no número de publicações após 2022, possivelmente refletindo a consolidação do paradigma e a transição do interesse da comunidade para técnicas de menor custo operacional.

5. Discussão

5.1. Recuperação Aumentada e Grounding Externo

RAG é a abordagem mais consolidada para mitigação de alucinações factuais (QP2). Borgeaud et al. (2022) demonstraram com o modelo RETRO que sistemas com recuperação de trilhões de tokens superam modelos paramétricos muito maiores em precisão factual, resultado que desafia a premissa de que o escalonamento de parâmetros é a principal via para melhorar factualidade. Nakano et al. (2021) estenderam esse princípio ao WebGPT, permitindo buscas reais durante a inferência; além de reduzir alucinações, a abordagem aumenta auditabilidade das respostas ao citar fontes.

Shuster et al. (2021) validaram RAG em sistemas de diálogo, reduzindo afirmações não verificáveis. Lazaridou et al. (2022) mostraram que busca por similaridade semântica permite atualizar o conhecimento do modelo sem re-treinamento, o que representa uma vantagem operacional relevante em domínios de evolução rápida. A principal limitação das abordagens RAG é a dependência de bases

de conhecimento atualizadas: em línguas com menor cobertura textual, o grounding pode ser insuficiente ou introduzir novos vieses (QP4).

5.2. Alinhamento por Feedback Humano

RLHF estabeleceu-se como abordagem fundamental para reduzir respostas incoerentes e melhorar a qualidade percebida. Stiennon et al. (2020) demonstraram que modelos menores treinados com feedback humano superam modelos maiores sem esse ajuste em sumarização. Ouyang et al. (2022) generalizaram o resultado ao InstructGPT: um modelo de 1,3B parâmetros com RLHF foi preferido por avaliadores humanos ao GPT-3 (175B), estabelecendo que alinhamento por preferência humana é mais determinante que tamanho para qualidade percebida.

As limitações do RLHF são práticas e estruturais: custo de anotação elevado, escalabilidade restrita e sensibilidade aos vieses dos anotadores. Yu et al. (2023) propuseram plug-and-play retrieval feedback como alternativa sem re-treinamento, com resultados parcialmente competitivos. A combinação de RLHF com RAG permanece pouco explorada empiricamente e representa uma lacuna de pesquisa relevante (QP3).

5.3. Auto-verificação e Verificação Cruzada

Manakul et al. (2023) propuseram SelfCheckGPT, que detecta alucinações por consistência entre múltiplas amostras da mesma pergunta, sem acesso a fontes externas, sendo aplicável em configurações de caixa-preta. A premissa é que informações factuais tendem a ser reproduzidas de forma estável, enquanto conteúdo alucinatório varia entre amostras. Weng et al. (2023) e Dhuliawala et al. (2023) avançaram com Chain-of-Verification (CoVe), que decompõe a resposta em afirmações atômicas, gera perguntas de verificação e revisa a saída com base nas respostas, reduzindo erros em listas biográficas e perguntas factuais.

Du et al. (2023) exploraram debate multi-agente: múltiplos LLMs questionam mutuamente suas respostas até um consenso, com ganhos de precisão superiores a modelos individuais. Kadavath et al. (2022) mostraram que LLMs calibrados estimam sua própria incerteza com razoável acurácia, apontando para sistemas que sinalizem proativamente ao usuário quando a confiança na resposta é baixa.

5.4. Intervenções Arquiteturais e de Decodificação

Li et al. (2023) propuseram ITI (Inference-Time Intervention), que identifica direções de ativação associadas a afirmações verídicas e desloca as representações do modelo nessas direções durante a inferência, sem re-treinamento. Experimentos no TruthfulQA demonstraram aumentos significativos na taxa de respostas factuais. Chuang et al. (2023) desenvolveram DoLa (Decoding by Contrasting Layers), amplificando informação factual ao contrastar distribuições de probabilidade entre camadas iniciais e superiores do transformer, partindo da premissa de que as camadas superiores codificam conhecimento factual mais refinado.

Wei et al. (2022) demonstraram que Chain-of-Thought (CoT) reduz erros em raciocínio aritmético e lógico ao induzir passos intermediários explícitos. Yao et al. (2023) generalizaram essa ideia com Tree of Thoughts, explorando múltiplos caminhos de raciocínio em paralelo. Ambas as abordagens reduzem alucinações emergentes de

raciocínio superficial, mas aumentam o custo de inferência e têm eficácia variável por domínio.

5.5. Avaliação e Benchmarks

Min et al. (2023) propuseram FActScore: textos longos são decompostos em fatos atômicos verificados individualmente contra uma base de conhecimento, superando BLEU e ROUGE em sensibilidade a erros pontuais. Maynez et al. (2020) documentaram por anotação humana que modelos de sumarização abstrata geram regularmente conteúdo não suportado pelo texto-fonte, invisível às métricas automáticas tradicionais. Guo et al. (2022) mostraram que modelos mal calibrados atribuem alta confiança a respostas incorretas, agravando o impacto das alucinações por torná-las mais convincentes.

Bang et al. (2023) avaliaram o ChatGPT em 23 tarefas e 10 línguas, documentando degradação sistemática de desempenho em raciocínio causal e idiomas sub-representados. Agrawal et al. (2023) mostraram que LLMs calibrados reconhecem, em alguma medida, quando estão propensos a errar, resultado que abre caminho para sistemas com sinalização proativa de incerteza. O problema estrutural central identificado por esses estudos é a ausência de benchmarks padronizados: cada trabalho avalia sua proposta em configurações incomparáveis, impedindo estimativas confiáveis de ganho relativo entre abordagens (QP3).

5.6. Análise Comparativa entre Categorias

Uma leitura transversal das cinco categorias revela padrões importantes que não emergem da análise isolada de cada abordagem, respondendo de forma integrada à QP2 e à QP3. RAG e RLHF são as técnicas com maior maturidade empírica e mais ampla adoção, mas apresentam requisitos de infraestrutura elevados: RAG exige bases de conhecimento atualizadas e sistemas de recuperação eficientes; RLHF demanda anotadores humanos especializados e processos de coleta de preferência custosos. Essas barreiras limitam sua aplicabilidade em contextos de recursos restritos, como pesquisa acadêmica independente ou sistemas de menor escala.

As técnicas de auto-verificação, como SelfCheckGPT, CoVe e debate multi-agente, oferecem uma alternativa mais acessível por não dependerem de infraestrutura externa, mas apresentam uma limitação fundamental: quando o modelo base é consistentemente incorreto sobre um fato, a verificação interna reproduz o mesmo erro de forma estável, mascarando a alucinação em vez de detectá-la. Essa fragilidade é especialmente crítica em domínios especializados onde o modelo foi pouco exposto durante o treinamento.

As intervenções arquiteturais como ITI e DoLa representam a fronteira mais recente e tecnicamente sofisticada, com resultados promissores em benchmarks controlados. Entretanto, sua generalização para cenários de produção ainda é limitada: ITI requer acesso às representações internas do modelo, indisponível em APIs comerciais; DoLa foi validado principalmente em tarefas de resposta curta, com comportamento menos documentado em geração longa. A ausência de comparações diretas entre essas técnicas em condições experimentais padronizadas, consequência da heterogeneidade metodológica identificada na Seção 5.5, representa a principal barreira para conclusões definitivas sobre qual abordagem é mais eficaz em cada tipo de tarefa.

Este trabalho se diferencia de surveys recentes como Ji et al. (2023), Huang et al. (2023) e Tonmoy et al. (2024) pela ênfase explícita em duas contribuições complementares: (i) a análise transversal das categorias na Seção 5.6, que identifica trade-offs e limitações cruzadas raramente discutidos nos surveys de escopo mais amplo; e (ii) o foco aplicado no contexto brasileiro de PLN, com agenda de pesquisa concreta para o português, dimensão ausente nos surveys internacionais.

6. Ameaças à Validade

Como toda revisão sistemática, este estudo está sujeito a ameaças à validade que devem ser consideradas na interpretação dos resultados. Em relação à validade de construto, a definição operacional de alucinação varia entre os estudos revisados: alguns focam em erros factualmente verificáveis, outros incluem inconsistências contextuais ou geração de conteúdo não suportado pela fonte. Essa heterogeneidade conceitual dificulta comparações diretas e pode ter influenciado a categorização das abordagens.

Em relação à validade interna, o processo de seleção dos 32 estudos, embora baseado em critérios explícitos e conduzido com triagem independente por três avaliadores, envolveu julgamentos interpretativos na fase de triagem. Estudos publicados em venues de menor visibilidade ou em idiomas diferentes do inglês podem ter sido sub-representados, introduzindo viés de publicação. A limitação ao período 2020–2025, embora justificada pela relevância do período pós-GPT-3, exclui trabalhos anteriores que podem ter contribuições relevantes para a compreensão de fenômenos relacionados. O uso extensivo de preprints arXiv como fonte — embora necessário para capturar o estado da arte em uma área de evolução rápida — constitui uma limitação adicional, pois preprints não passaram por revisão por pares formal; esse risco foi mitigado pela exigência de citação substancial na literatura primária como critério de inclusão.

Em relação à validade externa, os resultados refletem o estado da literatura em um período de rápida evolução tecnológica. Novos modelos como GPT-4o, Claude 3.5 e Llama 3.1 foram lançados durante ou após o período de coleta, e suas características de alucinação podem diferir substancialmente dos modelos avaliados nos estudos revisados. As conclusões sobre eficácia relativa das técnicas de mitigação devem ser interpretadas como válidas para os modelos e tarefas avaliados nos estudos originais, não necessariamente generalizáveis aos sistemas mais recentes. Essas limitações reforçam a necessidade de revisões periódicas e de benchmarks padronizados que permitam acompanhar a evolução do campo de forma sistemática.

7. Lacunas e Direções Futuras

Quatro lacunas estruturais emergem da análise, respondendo à QP3 e à QP4. A primeira é o desequilíbrio no esforço de pesquisa: aproximadamente 78% dos estudos tratam exclusivamente de alucinações factuais, deixando as variantes semântica e linguística sem definições operacionais precisas, datasets dedicados ou métricas consolidadas. Essa assimetria limita a compreensão holística do fenômeno e a capacidade de selecionar estratégias de mitigação por tipo de alucinação.

A segunda lacuna é a ausência de integrações multi-estratégia. RAG, RLHF, CoVe e ITI são avaliados de forma isolada na quase totalidade dos estudos. A combinação modular de recuperação externa com auto-verificação, por exemplo, poderia gerar ganhos sinérgicos não documentados. Tonmoy et al. (2024) identificam essa integração como a direção mais promissora para a próxima geração de sistemas de mitigação.

A terceira lacuna é a cobertura de línguas de menor recurso (QP4). Bang et al. (2023) quantificaram que LLMs apresentam taxas mais altas de alucinação em idiomas sub-representados nos dados de treinamento. Para o português, língua falada por mais de 260 milhões de pessoas e com LLMs amplamente adotados no Brasil, não há benchmark dedicado de alucinações nem avaliação sistemática das técnicas de mitigação disponíveis. Tapajós et al. (2025) demonstraram que modelos pré-treinados em português apresentam comportamentos distintos dos modelos multilíngues gerais, reforçando a necessidade de avaliações específicas.

A quarta lacuna diz respeito a contextos multimodais. Com a adoção de modelos como GPT-4V e Gemini em tarefas que integram texto e imagem, as taxonomias e métricas de alucinação desenvolvidas para texto puro não se aplicam diretamente. A extensão dessas abordagens ao regime multimodal representa uma fronteira de pesquisa incipiente com alta relevância aplicada, incluindo alucinações em sistemas agênticos baseados em LLMs que combinam planejamento, uso de ferramentas e geração de texto.

8. Conclusão

Este artigo apresentou uma revisão sistemática de 32 estudos (2020–2025) sobre alucinações em LLMs, orientada por quatro questões de pesquisa. As abordagens de mitigação foram organizadas em cinco categorias e suas contribuições, limitações e lacunas estruturais foram identificadas. RAG e RLHF são as técnicas com maior evidência empírica; CoVe, ITI e DoLa representam avanços mais recentes com resultados promissores, mas validação ainda limitada. A análise transversal das categorias revelou que nenhuma técnica isolada é suficiente para todos os tipos de alucinação e contextos de uso, e que a ausência de comparações diretas entre abordagens em condições experimentais padronizadas é a principal barreira ao progresso cumulativo do campo.

A seção de ameaças à validade destaca limitações metodológicas que devem ser consideradas na interpretação dos resultados, particularmente a heterogeneidade conceitual na definição de alucinação entre os estudos e o potencial viés de publicação em favor de resultados positivos. O uso de preprints foi reconhecido como uma limitação necessária, mitigada pelo critério de citação substancial. Essas limitações não invalidam as conclusões, mas reforçam a necessidade de protocolos de avaliação padronizados.

Três contribuições principais são oferecidas por este trabalho: (i) uma síntese comparativa estruturada das abordagens de mitigação, organizada em categorias que facilitam a seleção de técnicas em contextos específicos; (ii) a quantificação do desequilíbrio de pesquisa entre tipos de alucinação, com 78% dos estudos focando exclusivamente em alucinações factuais; e (iii) a identificação da lacuna em línguas de menor recurso como oportunidade concreta e urgente para a comunidade brasileira de

PLN, considerando a ampla adoção de LLMs no país e a ausência de benchmarks dedicados ao português. Trabalhos futuros devem priorizar a construção de benchmarks padronizados para o português, a investigação de integrações sistemáticas entre múltiplas estratégias de mitigação e a extensão das metodologias existentes ao regime multimodal e a sistemas agênticos.

Referências

Agrawal, G. et al. (2023) Do LLMs know when they're hallucinating? A study of LLM self-awareness. arXiv preprint arXiv:2305.11747.

Bang, Y. et al. (2023) A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. arXiv preprint arXiv:2302.04023.

Borgeaud, S. et al. (2022) Improving language models by retrieving from trillions of tokens. In: ICML 2022, Proceedings of Machine Learning Research, v. 162, p. 2206-2240.

Chuang, Y.-S. et al. (2023) DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. arXiv preprint arXiv:2309.03883.

Dhuliawala, S. et al. (2023) Chain-of-Verification Reduces Hallucination in Large Language Models. arXiv preprint arXiv:2309.11495.

Du, Y. et al. (2023) Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv preprint arXiv:2305.14325.

Gao, L. et al. (2023) RARR: Researching and Revising What Language Models Say, Using Language Models. arXiv preprint arXiv:2210.08726.

Guo, Z. et al. (2022) Calibrating Sequence Likelihood Improves Conditional Language Generation. arXiv preprint arXiv:2210.00045.

Han, X. and Tsvetkov, Y. (2022) ORCA: interpreting prompted language models via locating supporting data evidence. CoRR, abs/2205.12600.

Huang, L. et al. (2023) A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv preprint arXiv:2311.05232.

Ji, Z. et al. (2023) Survey of hallucination in natural language generation. ACM Computing Surveys, v. 55, n. 12, p. 248:1-248:38.

Kadavath, S. et al. (2022) Language Models (Mostly) Know What They Know. arXiv preprint arXiv:2207.05221.

Kitchenham, B. and Charters, S. (2007) Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report, Keele University.

Lazaridou, A. et al. (2022) Internet-Augmented Language Models through Few-Shot Prompting for Open-Domain Question Answering. arXiv preprint arXiv:2203.05115.

Lewkowycz, A. et al. (2022) Solving Quantitative Reasoning Problems with Language Models. Advances in Neural Information Processing Systems, v. 35.

Li, K. et al. (2023) Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. arXiv preprint arXiv:2306.03341.

Lin, S. et al. (2022) TruthfulQA: Measuring How Models Mimic Human Falsehoods. In: ACL 2022, p. 3214-3252.

Manakul, P.; Liusie, A. and Gales, M. J. F. (2023) SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv preprint arXiv:2303.08896.

Maynez, J. et al. (2020) On Faithfulness and Factuality in Abstractive Summarization. Proceedings of ACL 2020.

Medeiros, L. S. F. and Oliveira, H. T. A. (2025) Comparação de Modelos de Embeddings e LLMs para Geração Aumentada por Recuperação em Português. In: Anais do SEMISH 2025, p. 429-440.

Min, S. et al. (2023) FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. arXiv preprint arXiv:2305.14251.

Mundler, N. et al. (2023) Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. arXiv preprint arXiv:2305.15852.

Nakano, R. et al. (2021) WebGPT: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332.

Ouyang, L. et al. (2022) Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, v. 35.

Peng, B. et al. (2023) Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. arXiv preprint arXiv:2302.12813.

Shuster, K. et al. (2021) Retrieval Augmentation Reduces Hallucination in Conversation. arXiv preprint arXiv:2104.07567.

Stiennon, N. et al. (2020) Learning to summarize with human feedback. Advances in Neural Information Processing Systems, v. 33.

Sun, W. et al. (2023) Contrastive learning reduces hallucination in conversations. In: AAAI 2023, p. 13618-13626.

Tapajós, R. et al. (2025) Avaliação de comportamento de modelos de linguagem pré-treinados em português: um estudo comparativo. In: Anais do STIL 2025.

Tonmoy, S. M. et al. (2024) A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. arXiv preprint arXiv:2401.01313.

Vaswani, A. et al. (2017) Attention Is All You Need. Advances in Neural Information Processing Systems, v. 30.

Wei, J. et al. (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems, v. 35.

Weng, M. et al. (2023) Large Language Models Are Better Reasoners with Self-Verification. arXiv preprint arXiv:2212.09561.

Wolf, T. et al. (2020) Transformers: State-of-the-art natural language processing. In: EMNLP 2020 System Demonstrations, p. 38-45.

Yao, S. et al. (2023) Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv preprint arXiv:2305.10601.

Yu, W. et al. (2023) Improving language models via plug-and-play retrieval feedback. CoRR, abs/2305.14002.

Zhao, Z. et al. (2021) Calibrate Before Use: Improving Few-Shot Performance of Language Models. Proceedings of ICML 2021.