UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
FACULDADE DE COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

ANGELO HENRIQUE PERES CESTARI JUNIOR

ANÁLISE DE ALTERNATIVAS PARA MITIGAR ALUCINAÇÕES EM MODELOS DE LINGUAGEM DE LARGA ESCALA

CAMPO GRANDE-MS
2025

ANGELO HENRIQUE PERES CESTARI JUNIOR

ANÁLISE DE ALTERNATIVAS PARA MITIGAR ALUCINAÇÕES EM MODELOS DE LINGUAGEM DE LARGA ESCALA

Trabalho de Conclusão de Curso de Graduação em Engenharia de Computação, da Faculdade de Computação da Universidade Federal de Mato Grosso do Sul (FACOM/UFMS) como registro para a obtenção do título de Bacharel em Engenharia de Computação
Orientador: Prof. Dr. Renato Porfirio Ishii

CAMPO GRANDE-MS
2025

Aos meus pais, Angelo e Andreia, é uma honra ser filho de vocês. Obrigado por acreditarem nos meus planos, por investirem tanto na minha educação e por sempre estarem presentes, mesmo de longe, em todos os momentos que precisei. Essa conquista é de vocês também.

AGRADECIMENTOS

Agradeço primeiramente a Deus por todas as bênçãos que Ele me concedeu, especialmente pela saúde e por uma família maravilhosa. Sem fé, nada somos, e Seu suporte foi fundamental para esta conquista.

À minha querida irmãzinha Alessandra, obrigado por todas as conversas, brincadeiras e até pelas brigas que tivemos. Você sempre foi um exemplo de dedicação nos estudos, e superar suas notas sempre foi motivo para um esforço extra.

À minha avó Chica, sempre preocupada se eu estava me alimentando direito, se estava me cuidando e, principalmente, pelo carinho em preparar todas as caixas térmicas recheadas de comida para mim. Seus parmegianas foram fundamentais para a conclusão deste trabalho.

À minha namorada Kamilly, que me encontrou em um momento não muito focado nos estudos e me colocou de volta aos trilhos. Sua companhia tornou-me uma pessoa muito melhor e mais motivada a batalhar por todos os meus sonhos.

Aos meus amigos, em especial Gabriel Zaneti e Miguel Zaneti, por estarem sempre à disposição, mesmo de longe, para ouvir os meus problemas, ajudar a resolvê-los e serem sempre bons conselheiros.

Por fim, agradeço a todos os meus professores ao longo dessa jornada por transmitirem seus conhecimentos com tanto carinho e dedicação. Em especial ao meu orientador, Prof. Dr. Renato Porfirio Ishii, que não apenas me orientou neste trabalho, mas também me aconselhou em caminhos futuros.

RESUMO

Modelos de linguagem de larga escala (LLMs) têm se tornado centrais em aplicações que exigem geração de texto natural, resolução de problemas e apoio à tomada de decisão. Apesar de seus avanços, esses modelos permanecem suscetíveis ao fenômeno das alucinações, caracterizado pela produção de respostas incorretas ou não verificáveis, o que limita sua utilização em domínios sensíveis. Este trabalho investiga diferentes estratégias de mitigação de alucinações aplicadas ao modelo Llama3.2 3B, avaliando abordagens baseadas em recuperação de informação (RAG), reranqueamento (MMR e reranqueamento neural), verificação interna (*Chain-of-Verification*) e pós-edição com agente revisor (*Answer + Reviewer*). Para isso, foram conduzidos experimentos padronizados nos benchmarks TruthfulQA e ARC Challenge, que avaliam, respectivamente, veracidade e capacidade de raciocínio. Os resultados obtidos indicam que técnicas de recuperação de informação mostram ganhos relevantes no ARC Challenge, especialmente o RAG com reranqueamento. Por outro lado, a técnica de verificação interna, representada pelo CoVe, obteve resultados surpreendentes, em comparação com o modelo base, no TruthfulQA. Conclui-se, portanto, que a mitigação efetiva das alucinações depende da natureza da tarefa, e que combinações híbridas entre recuperação e verificação interna representam um caminho promissor para o desenvolvimento de modelos mais confiáveis.

Palavras-chave: modelos de linguagem de larga escala; mitigar alucinações; RAG; verificação interna

ABSTRACT

Large Language Models (LLMs) have become central to applications requiring natural language generation, problem solving, and decision support. Despite recent advances, these models remain susceptible to the phenomenon of hallucinations, characterized by the production of incorrect or unverifiable responses, which limits their use in sensitive domains. This work investigates different hallucination-mitigation strategies applied to the Llama3.2 3B model, evaluating approaches based on information retrieval (RAG), reranking (MMR and neural reranking), internal verification (Chain-of-Verification), and post-editing with a reviewer agent (Answer + Reviewer). Standardized experiments were conducted on the TruthfulQA and ARC Challenge benchmarks, which evaluate accuracy and reasoning ability, respectively. The results indicate that information-retrieval techniques yield relevant gains in the ARC Challenge—particularly RAG with reranking. Conversely, the internal verification technique, represented by CoVe, achieved remarkably strong results on the TruthfulQA compared with the baseline model. Therefore, we conclude that effective hallucination mitigation depends on the nature of the task and that hybrid combinations of retrieval and internal verification represent a promising direction for developing more reliable language models.

Keywords: large language models; mitigate hallucinations; retrieval augmented generation; intern verification

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARC	AI2 Reasoning Challenge
CoVe	Chain-of-Verification
GenAI	Inteligência Artificial Generativa
GPT	Generative Pre-trained Transformer
IA	Inteligência Artificial
LLMs	Large Language Models
MMR	Maximal Marginal Relevance
RAG	Retrieval-Augmented Generation
RLHF	Reinforcement Learning from Human Feedback

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, o termo *Inteligência Artificial* (IA) tem ocupado um espaço crescente nas discussões científicas, tecnológicas e sociais. Embora seja apresentada como uma área recente, suas origens remontam ao período da Segunda Guerra Mundial, quando pesquisadores começaram a investigar a possibilidade de criar máquinas capazes de simular o raciocínio humano (Turing, 1950). Desde então, a IA passou por diversas fases de desenvolvimento, alternando períodos de avanço e de estagnação. Essa trajetória histórica preparou o terreno para uma mudança de paradigma mais recente, na qual técnicas avançadas de aprendizado profundo redefiniram as capacidades dos sistemas inteligentes.

É nesse novo cenário que se insere a atual revolução da *Inteligência Artificial Generativa* (GenAI), viabilizada principalmente pelos modelos de linguagem de larga escala (*large language models*, LLMs). Esses modelos são sistemas de aprendizado profundo treinados com enormes quantidades de dados, capazes de compreender e gerar textos coerentes, traduzir idiomas, resumir documentos, responder a perguntas, resolver problemas e até auxiliar em tarefas complexas, como geração de código e apoio a decisões (Bommasani, 2021).

Com o lançamento do *ChatGPT* (Radford et al., 2018), desenvolvido pela OpenAI, os LLMs tornaram-se amplamente conhecidos pelo público e pela comunidade científica, demonstrando capacidades antes consideradas inacessíveis. As implicações desses avanços transcendem o domínio tecnológico, influenciando profundamente o modo como os indivíduos trabalham, aprendem e interagem, além de suscitar discussões sobre os limites éticos e sociais da automação inteligente.

Contudo, o avanço desses modelos trouxe também desafios significativos. Entre eles, destaca-se o fenômeno das *alucinações*, que ocorre quando o modelo produz respostas factualmente incorretas, inconsistentes ou completamente inventadas, limitando a confiabilidade e utilização dessas tecnologias.

1.1 JUSTIFICATIVA

Para compreender a relevância prática do problema das alucinações e justificar a necessidade de investigar técnicas de mitigação, é importante observar como os LLMs já vêm sendo incorporados a diferentes setores profissionais. Entre os contextos mais sensíveis, destacam-se o jurídico e o da saúde, nos quais erros podem produzir impactos diretos, imediatos e potencialmente graves. Nesses ambientes, a confiabilidade das respostas geradas pelos modelos não é apenas desejável, mas um requisito normativo, ético e operacional. Assim, nas subseções seguintes, são exemplificados esses dois domínios nos quais o fenômeno das alucinações assume particular relevância, ilustrando o porquê a mitigação desse problema é essencial para o uso responsável e seguro dessas tecnologias.

1.1.1 Área Jurídica

No campo jurídico, o uso de sistemas baseados em IA tem se expandido de forma acelerada, abrangendo desde a automação de tarefas administrativas até a análise de documentos e jurisprudências (DSA, 2024). Ferramentas alimentadas por LLMs demonstram potencial para auxiliar na redação de petições, elaboração de pareceres e busca de precedentes, promovendo maior eficiência e economia de tempo. Contudo, a utilização de modelos que ainda apresentam alucinações representa um risco significativo para a prática jurídica, uma vez que informações imprecisas podem comprometer a fundamentação de decisões, induzir interpretações equivocadas e gerar insegurança jurídica.

Casos recentes evidenciam de forma concreta os riscos decorrentes da utilização indiscriminada de modelos de linguagem em atividades jurídicas. Um exemplo amplamente divulgado ocorreu nos Estados Unidos, quando um advogado apresentou ao Tribunal de Recursos do Segundo Distrito da Califórnia uma petição com o auxílio do ChatGPT. Durante a análise, constatou-se que 21 das 23 citações jurídicas incluídas no documento eram falsas (Melo, 2025). O episódio culminou na aplicação de uma multa de 10 mil dólares ao profissional e evidenciou a necessidade de mecanismos de verificação e controle de confiabilidade nas ferramentas baseadas em IA.

A confiabilidade é, portanto, requisito indispensável para a adoção segura dessas tecnologias no ambiente jurídico. A mitigação das alucinações torna-se essencial para que as LLMs possam atuar como ferramentas de apoio efetivo à tomada de decisão, sem violar princípios como a veracidade, a boa-fé e a segurança das informações. O aprimoramento desses modelos possibilita ampliar seu uso em tarefas de suporte jurídico, sempre preservando a integridade técnica e ética das informações produzidas.

1.1.2 Área da Saúde

Na área da saúde, os modelos de linguagem também vêm sendo aplicados com o objetivo de apoiar diagnósticos, sintetizar informações clínicas, gerar relatórios médicos e auxiliar na educação de pacientes e profissionais. A capacidade dos LLMs de compreender e gerar textos em linguagem natural os torna especialmente promissores para a gestão e análise de grandes volumes de dados médicos. No entanto, o risco de alucinações impõe sérias limitações a sua adoção plena, uma vez que a produção de informações incorretas ou imprecisas pode afetar diretamente a saúde do paciente e comprometer a tomada de decisões clínicas (Bommasani, 2021).

Para evidenciar essa realidade, a pesquisa *TIC Saúde 2024* (Portilho, 2024) revelou que 17 em cada 100 médicos brasileiros já utilizam ferramentas de inteligência artificial generativa (GenAI) em suas rotinas profissionais. De acordo com o levantamento, os principais usos estão relacionados ao suporte em pesquisas e à elaboração de relatórios médicos, evidenciando a utilização dessas tecnologias no cotidiano hospitalar.

Dessa forma, a redução das alucinações é um passo fundamental para que a GenAI se torne uma aliada confiável no campo da saúde. A implementação de estratégias eficazes de mitigação contribuirá para aumentar a precisão das respostas e consolidar a confiança dos profissionais de saúde no uso dessas ferramentas.

1.2 OBJETIVOS

Mitigar alucinações em modelos de linguagem de larga escala é uma tarefa complexa e desafiadora. As alucinações podem ocorrer em diferentes níveis, desde a produção de informações factualmente incorretas até a geração de respostas logicamente inconsistentes. Diversas abordagens têm sido propostas para reduzir esse problema, incluindo ajustes no treinamento (Ouyang et al., 2022), mecanismos de verificação factual (Dhuliawala et al., 2024) e técnicas de recuperação de informações externas (Lewis et al., 2020).

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é **analisar e comparar diferentes alternativas de mitigação de alucinações em modelos de linguagem de larga escala**, avaliando sua efetividade quanto à precisão factual e à confiabilidade das respostas produzidas.

1.2.2 Objetivos Específicos

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- Revisar as principais estratégias de mitigação apresentadas na literatura recente, classificando-as de acordo com a taxonomia proposta por Silva (2025);
- Testar diferentes abordagens de mitigação selecionadas, como o Retrieval-Augmented Generation (RAG), métodos de pós-edição (*post-editing*) e o uso de multiagentes de IA.
- Avaliar comparativamente os resultados obtidos por meio de métricas de consistência factual, precisão e coerência, identificando as técnicas mais promissoras para redução de alucinações.

Esses objetivos visam contribuir para o aprofundamento do conhecimento sobre as alternativas existentes de mitigação de alucinações, bem como para a implementação e avaliação prática dessas técnicas.

O capítulo 2 apresenta os trabalhos relacionados, descrevendo a revisão sistemática da literatura e destacando os estudos que fundamentam as estratégias selecionadas para os experimentos. Em seguida, o capítulo 3 reúne a fundamentação teórica, abordando os conceitos essenciais sobre modelos de linguagem, arquitetura *Transformers*, família GPT e o fenômeno das alucinações.

O capítulo 4 apresenta detalhadamente os métodos empregados para atingir esses objetivos, descrevendo cada estratégia de mitigação e os critérios adotados para avaliar seu desempenho. No capítulo 5, são expostos e discutidos os resultados experimentais, analisando comparativamente o impacto de cada método. Por fim, o capítulo 6 reúne as conclusões deste trabalho e aponta direções promissoras para pesquisas futuras.

2 TRABALHOS RELACIONADOS

Para compreender de forma aprofundada o problema das alucinações em modelos de linguagem, realizou-se inicialmente uma revisão sistemática da literatura, com o objetivo de identificar e classificar as principais abordagens de mitigação propostas na literatura recente. As soluções encontradas foram agrupadas em quatro categorias, conforme a taxonomia proposta por Silva (2025), detalhada em breve.

2.1 REVISÃO SISTEMÁTICA DA LITERATURA

2.1.1 Planejamento

O processo de revisão, representado na Figura 1, foi planejado de forma a equilibrar abrangência e precisão na busca de publicações relevantes. Inicialmente, foram definidos critérios mínimos de seleção para garantir a pertinência temática e a qualidade metodológica dos estudos incluídos.

Em seguida, realizou-se a coleta dos artigos em bases científicas específicas. Após essa etapa, foi conduzida uma avaliação criteriosa dos trabalhos identificados, com o objetivo de selecionar apenas aqueles que atendessem aos requisitos estabelecidos.

Por fim, efetuou-se uma leitura analítica e classificatória, a partir da qual os artigos foram organizados nas categorias definidas, possibilitando uma visão estruturada das estratégias contemporâneas de mitigação de alucinações em modelos de linguagem.

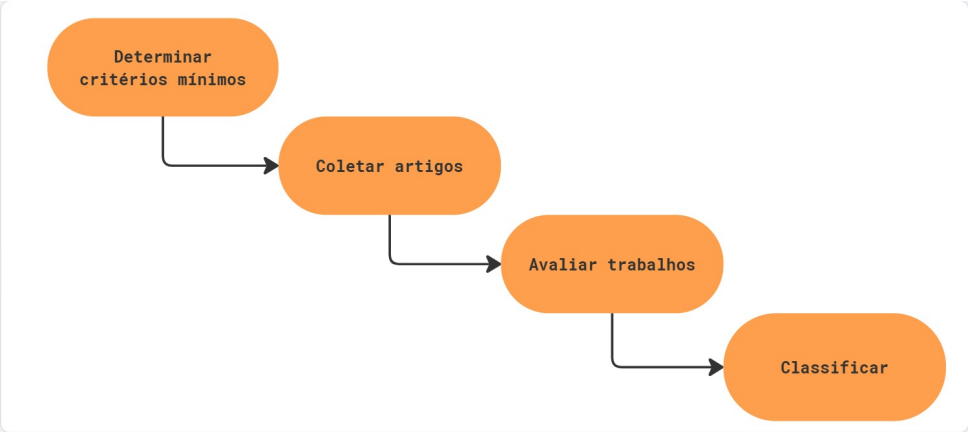

Figura 1 – Fluxograma da revisão sistemática da literatura.

2.1.2 Critérios mínimos adotados

Para garantir a seleção de artigos recentes, relevantes e alinhados ao escopo desta pesquisa, foram estabelecidos os seguintes critérios mínimos de inclusão:

- **Período de publicação:** entre os anos de 2023 e 2025;
- **Idioma:** publicações disponíveis em língua inglesa ou portuguesa;
- **Fonte de publicação:** artigos encontrados em, pelo menos, um dos seguintes repositórios científicos reconhecidos:
  - *ACM Digital Library*;
  - *ArXiv*;
  - *IEEE Digital Library*;
  - *Science@Direct*.

2.1.3 Coleta de artigos

A etapa de coleta teve como objetivo reunir os estudos mais relevantes sobre técnicas de mitigação de alucinações em modelos de linguagem. Para isso, foi empregada a plataforma *Google Scholar*, por sua ampla cobertura e constante atualização de publicações científicas.

A busca foi conduzida utilizando a seguinte *query* de busca:

```
("Retrieval-Augmented Generation"OR RAG) AND "hallucinations"AND
                        "mitigation")
```

Essa combinação de termos foi definida de modo a abranger publicações que tratassem explicitamente da geração aumentada por recuperação e das estratégias de redução de alucinações em modelos de linguagem, porém diversas abordagens diferentes foram encontradas, mais discutidos posteriormente.

Como resultado inicial, foram retornados 3.520 artigos. Após a aplicação dos critérios mínimos de inclusão definidos anteriormente, permaneceram 473 trabalhos válidos, dos quais 6 foram identificados como duplicados e, portanto, removidos. O conjunto final de 467 artigos constituiu a base de análise para a etapa subsequente de avaliação e classificação.

2.1.4 Avaliação dos artigos

A definição dos critérios de inclusão e exclusão constitui uma etapa fundamental no processo de revisão sistemática da literatura, uma vez que garante a objetividade, a reprodutibilidade e o direcionamento adequada da seleção de estudos. Esses critérios permitem filtrar os trabalhos mais relevantes, assegurando que apenas publicações com potencial de contribuição direta ao tema sejam analisadas em profundidade.

A verificação dos critérios foi realizada manualmente, mediante a leitura dos títulos, resumos e, quando necessário, do texto completo dos estudos encontrados. Para isso, utilizaram-se palavras-chave como *github*, *OpenAI*, *API*, *mitigation* e *HuggingFace*, processo que demandou aproximadamente dois meses.

Os critérios adotados neste estudo estão descritos a seguir.

2.1.4.1 Critérios de Inclusão

Foram priorizados artigos que apresentavam caráter aplicado e contribuíam de forma prática para o desenvolvimento de técnicas de mitigação de alucinações. Além disso, buscou-se incluir estudos que disponibilizassem seus códigos de implementação ou resultados experimentais de forma pública, permitindo a replicação e a validação dos experimentos descritos.

Assim, foram incluídos:

- Estudos que apresentavam implementações ou repositórios públicos disponíveis;
- Trabalhos com foco explícito em estratégias de mitigação de alucinações em modelos de linguagem;
- Publicações que reportavam métricas quantitativas de avaliação, permitindo comparação de desempenho entre abordagens.

2.1.4.2 Critérios de Exclusão

Foram excluídos os estudos que não atendiam aos requisitos de reprodutibilidade e transparência, bem como aqueles cujo escopo não contemplava diretamente a mitigação de alucinações.

Dessa forma, foram desconsiderados:

- Artigos cujos códigos, modelos ou conjuntos de dados não estavam disponíveis publicamente;
- Trabalhos que abordavam apenas a detecção de alucinações, sem propor mecanismos de mitigação;
- Estudos conceituais ou teóricos sem experimentação empírica verificável.

2.1.5 Classificação dos artigos

Embora a *query* de busca tenha sido elaborada com o intuito de recuperar exclusivamente artigos que abordassem o método RAG, observou-se a presença de diversas alternativas voltadas à mitigação de alucinações. Diante desse cenários, os 16 artigos selecionados foram classificados nas seguintes categorias (Silva, 2025): *Otimização de Modelo*, *Aprimoramento de Inferência e Saída*, *Conhecimento Externo e Interação* e *Métodos Experimentais*.

Cada categoria é detalhada a seguir.

2.1.5.1 Otimização de Modelo

Engloba técnicas que aprimoram a arquitetura do modelo e o seu processo de aprendizado, explorando estratégias como o enriquecimento de dados, ajuste fino supervisionado (*fine-tuning*) e a experimentação com diferentes configurações de hiperparâmetros.

O objetivo principal dessa categoria é aumentar a robustez e a capacidade preditiva dos modelos, reduzindo a ocorrência de respostas imprecisas ou incoerentes.

2.1.5.2 Aprimoramento de Inferência e Saída

Reúne abordagens aplicadas na etapa de inferência com o propósito de refinar as saídas geradas pelo modelo. Inclui técnicas de pós-processamento, como filtros baseados em regras ou correção automática de inconsistências por meio de feedback em tempo real. Estratégias como o Reinforcement Learning from Human Feedback (RLHF) exemplificam bem essa categoria, uma vez que utilizam o retorno humano para ajustar o comportamento do modelo e otimizar a qualidade das respostas.

2.1.5.3 Conhecimento Externo e Interação

Abrange métodos que integram fontes de dados externas ou mecanismos de feedback interativo para enriquecer ou corrigir as respostas geradas. Essa abordagem tem como premissa a incorporação de informações atualizadas e contextualmente relevantes, mitigando limitações inerentes ao conhecimento estático do modelo. Um dos exemplos mais representativos dessa categoria é o RAG (Lewis et al., 2020) que combina geração neural com recuperação de informações para aprimorar a precisão factual.

2.1.5.4 Métodos Experimentais

Compreende um conjunto de abordagens inovadoras e exploratórias que buscam compreender e mitigar as alucinações a partir de novas perspectivas de análise do comportamento dos modelos. Nessa categoria, destacam-se investigações envolvendo sistemas multiagentes, análise de estados internos de representação e técnicas inspiradas em princípios metacognitivos. Tais métodos ainda se encontram em estágio experimental, mas demonstram potencial promissor para o desenvolvimento de modelos mais interpretáveis e confiáveis.

2.1.6 Síntese e direcionamento experimental

Dentre todas as categorais identificadas, este estudo decidiu não incluir nos experimentos a categoria *Otimização de Modelo*, por exigir recursos computacionais elevados e etapas de treinamento extensivo que vão além do escopo deste trabalho. Essa exclusão

permite concentrar a análise nas estratégias que são facilmente aplicáveis e executáveis em ambiente local, mantendo a viabilidade da avaliação empírica.

Dessa forma, os experimentos desenvolvidos nesta pesquisa focam em três das quatro categorias identificadas: *Conhecimento Externo e Interação*, *Aprimoramento de Inferência e Saída* e *Métodos Experimentais*. Cada uma delas representa uma linha de investigação consolidada na literatura e alinhada aos desafios apontados pela revisão.

Na próxima seção, são apresentados os principais trabalhos selecionados a partir da revisão sistemática da literatura, os quais constituíram a base conceitual e metodológica para o desenvolvimento dos experimentos conduzidos neste estudo.

2.2 ESTUDOS NORTEADORES

No estudo *Survey of Hallucination in Natural Language Generation* (Ji et al., 2023), os autores realizam a revisão mais abrangente da área, consolidando definições, causas, tipologias e métricas relacionadas ao fenômeno das alucinações. O trabalho diferencia dois tipos principais de alucinações, *intrínsecas*, quando o modelo contradiz informações presentes na entrada, e *extrínsecas*, quando o modelo produz conteúdo que não pode ser verificado por fontes disponíveis. Além disso, o estudo discute os fatores arquiteturais, cognitivos e contextuais que intensificam o problema, bem como limitações dos métodos existentes e direções futuras. Essa revisão fornece a base conceitual que orienta toda a discussão deste trabalho e justifica a seleção das categorias experimentais analisadas.

Na categoria de *Conhecimento Externo e Interação*, destaca-se o RAG (Lewis et al., 2020), considerado uma das abordagens mais influentes para mitigar alucinações extrínsecas. O RAG auxilia o modelo a incorporar informações recuperadas de repositórios externos confiáveis, reduzindo a dependência exclusiva do conhecimento paramétrico. Neste estudo, além do RAG simples, investigaram-se variações baseadas em reranqueamento de documentos, como o uso de MMR (Maximal Marginal Relevance) e de modelos *cross-encoder*, a fim de analisar como diferentes estratégias de seleção de contexto afetam a veracidade das respostas.

No conjunto de técnicas enquadradas na categoria de *Aprimoramento de Inferência e Saída*, destaca-se o *Chain-of-Verification* (CoVe) (Dhuliawala et al., 2024). Essa abordagem representa um avanço importante ao propor que o modelo realize uma verificação estruturada de suas próprias respostas por meio da decomposição em passos verificáveis. O CoVe explora exclusivamente o conhecimento interno do modelo, funcionando como uma forma de auditoria autorregressiva que reduz inconsistências sem recorrer a bases externas.

Por fim, a categoria de *Métodos Experimentais* inclui estratégias de pós-edição baseadas em agentes, como o pipeline *Answer + Reviewer*, no qual uma resposta inicial é submetida a um agente verificador encarregado de corrigir imprecisões ou rejeitar afirmações não suportadas.

3 FUNDAMENTAÇÃO TEÓRICA

Para compreender plenamente como as alucinações surgem em modelos de linguagem de larga escala, é necessário, antes, examinar os fundamentos arquiteturais que possibilitaram o desenvolvimento dessas tecnologias. Os avanços recentes em LLMs são fruto de inovações estruturais que redefiniram a forma como máquinas processam e representam a linguagem. Entre essas inovações, destaca-se a arquitetura *Transformer* (Vaswani et al., 2017), que serviu de base para modelos posteriores cada vez mais sofisticados, como a família *GPT*.

Assim, este capítulo apresenta, de forma progressiva, os principais elementos conceituais que sustentam o funcionamento das LLMs. A seção 3.1 apresenta a arquitetura *Transformer*, que introduziu o mecanismo de atenção como núcleo do processamento moderno de linguagem. Já a seção 3.2, introduz os modelos *GPT*, que expandiram essa arquitetura para tarefas autorregressivas de larga escala. E por fim, na seção 3.3, é apresentado o fenômeno das alucinações, analisado à luz dessas estruturas e de suas limitações intrínsecas.

A seguir, inicia-se pela arquitetura que tornou essa revolução possível: os *Transformers*.

3.1 TRANSFORMERS

O marco fundamental para o desenvolvimento dos modelos de linguagem de larga escala foi a publicação do artigo *Attention is All You Need* (Vaswani et al., 2017), que introduziu a arquitetura *Transformer* (Figura 2). Esse modelo substituiu estruturas recorrentes e convolucionais tradicionais por um mecanismo de atenção capaz de capturar relações entre palavras de forma paralela e altamente eficiente.

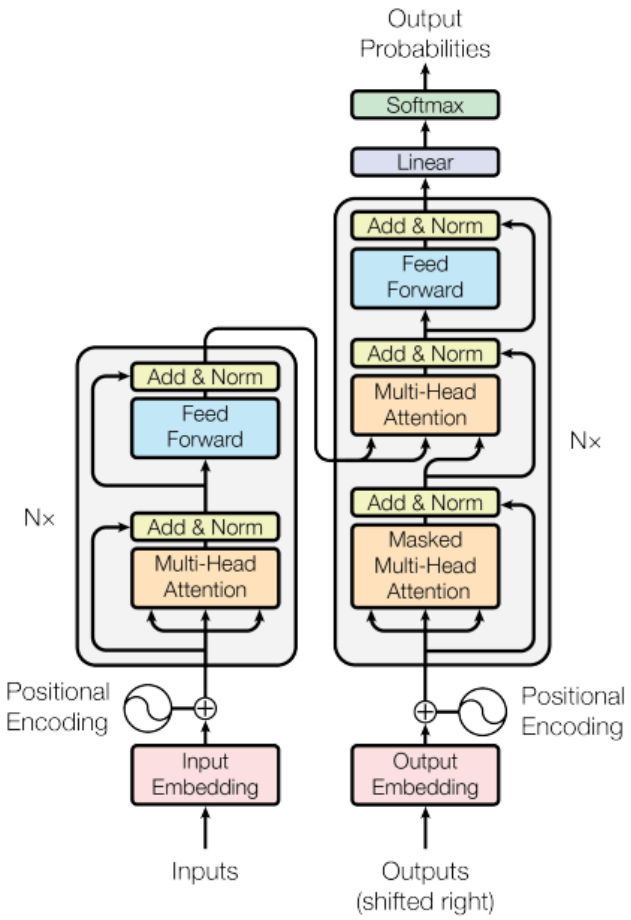

Figura 2 – Arquitetura *Transformer*.

Fonte: (Vaswani et al., 2017).

A arquitetura *Transformer*, figura 2, é composta, em sua forma original, por dois blocos principais: *encoder* (à esquerda) e *decoder* (à direita). O *encoder* recebe como entrada uma sequência de tokens, que são inicialmente convertidos em vetores densos por meio da camada de *embeddings*. Esses vetores são altamente enriquecidos com informações de posição por meio dos chamados *positional encodings*, necessários porque, diferentemente das redes recorrentes, o *Transformer* não processa a entrada de maneira sequencial, mas sim em parelelo. O objetivo do *encoder* é produzir uma representação contextualizada de toda as sequência de entrada.

Cada camada do *encoder* é composta, essencialmente, por dois componentes: o mecanismo de atenção multi-cabeças (*multi-head attention*) e uma rede *feed-forward* totalmente conectada. O mecanismo de atenção permite que o modelo atribua pesos diferentes às palavras da sequência, identificando quais tokens são mais relevantes para a interpretação de cada posição. O uso de múltiplas cabeças de atenção possibilita que o modelo capture, simultaneamente, diferentes tipos de relação semântica entre os tokens. Em se-

guida, a rede *feed-forward* atua de forma ponto a ponto sobre cada posição, refinando as representações geradas pela atenção.

O funcionamento desse mecanismo é formalizado por meio do *Scaled Dot-Product Attention*, definido por:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Nessa sequência, Q, K e V representam as matrizes de *queries*, *keys* e *values*. O produto QK^T calcula a similaridade entre cada token e os demais, enquanto a divisão por $\sqrt{d_k}$ estabiliza os valores resultantes, evitando que fiquem numericamente muito grandes. Em seguida, a função *softmax* transforma essas similaridades em pesos de atenção, que são aplicados sobre os vetores em V, produzindo uma representação contextualizada para cada posição da sequência.

Esses componentes são envolvidos por conexões residuais (*residual connections*) e camadas de normalização (*layer normalization*), que contribuem para a estabilização do treinamento em modelos profundos. As conexões residuais ajudam a mitigar o problema do *vanishing gradient*, permitindo que o gradiente flua mais facilmente entre as camadas, enquanto a normalização melhora a estabilidade numérica e a convergência do modelo.

O *decoder*, por sua vez, é responsável pela geração da saída, token a token. Ele também utiliza camadas de atenção multi-cabeças, porém com duas variantes. A primeira é a atenção mascarada (*masked multi-head attention*), que garante que cada posição só possa "enxergar" tokens anteriores na sequência gerada, preservando a natureza autorregressiva do modelo. A segunda é a atenção sobre as representações produzidas pelo *encoder*, o que permite ao *decoder* condicionar a geração da saída à sequência de entrada. Ao final, as representações do *decoder* são projetadas para o vocabulário por meio de uma camada linear seguida de uma função *softmax*, que produz uma distribuição de probabilidade sobre o próximo token a ser gerado.

O uso intensivo do mecanismo de atenção, aliado ao processamento paralelo das sequências, possibilitou treinar modelos com grande profundidade e escala, aumentando significativamente a capacidade de representar dependências de longo alcance no texto. Essa característica tornou a arquitetura *Transfomer* a base de praticamente todos os modelos de linguagem de larga escala modernos.

3.2 GENERATIVE PRE-TRAINED TRANSFORMER (GPT)

A evolução natural da arquitetura *Transformer* levou ao desenvolvimento dos modelos autorregressivos da família GPT, cuja versão mais influente é o GPT-3 (Brown et al., 2020). Esse modelo utiliza exclusivamente o bloco de *decoder* do *Transformer*, gerando texto de maneira autorregressiva, isto é, prevendo cada token com base nos anteriores. O GPT-3 foi treinado com uma variedade massiva de textos extraídos da

internet, livros digitalizados e repositórios públicos, permitindo ao modelo capturar padrões linguísticos, relações semânticas complexas e conhecimento factual disperso. Seu pré-treinamento não supervisionado, seguido da ausência de *fine-tuning* específico, tornou-o um dos primeiros modelos capazes de executar uma ampla gama de tarefas sem ajustes paramétricos adicionais, apenas com instruções fornecidas no próprio *prompt*.

Enquanto o Transformer original emprega mecanismos de atenção para processar entradas completas, o GPT adapta essa arquitetura para operar de forma inteiramente autorregressiva, empregando atenção mascarada para garantir que cada posição só tenha acesso aos tokens anteriores durante a geração. A Figura 3 ilustra essa configuração, destacando a repetição de blocos compostos por *masked multi self attention*, camadas de normalização e redes *feed-forward*, além da forma como diferentes tarefas são convertidas em sequências textuais únicas.

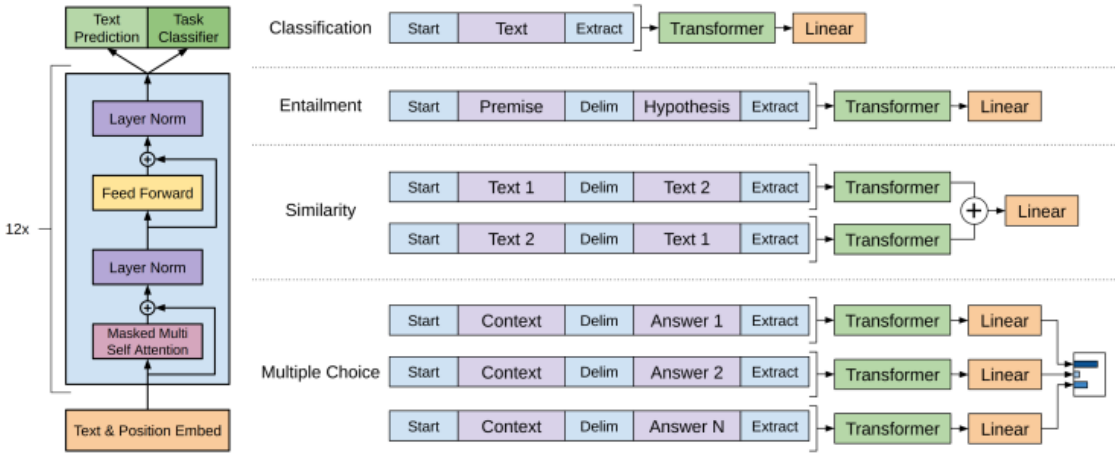

Figura 3 – Arquitetura GPT.

Fonte: (Radford et al., 2018).

Como mostra a figura acima, o GPT processa qualquer tarefa - classificação, inferência, similaridade ou múltipla escolha - como uma única sequência de texto, reforçando o papel central do *prompt* na determinação do comportamento do modelo. A ausência do bloco de *encoder* elimina o uso do mecanismo de *cross-attention*, característico do *Transformer* completo, tornando o GPT um modelo inteiramente autorregressivo. Essa abordagem simplificada, porém extremamente escalável, contribui para a versatilidade do GPT e explica sua capacidade de generalizar comportamentos a partir de instruções mínimas.

Um dos principais avanços introduzidos pelo GPT-3 é sua escala sem precedentes, composta por 175 bilhões de parâmetros, o que representou um salto exponencial em relação a modelos anteriores. Os autores demonstraram que o aumento da escala resulta no surgimento de capacidades emergentes, isto é, habilidades que não podiam ser previstas a partir do desempenho de modelos menores, tais como coerência em diálogos longos, racio-

cínio básico, elaboração de textos estruturados e interpretação contextual avançada. Além disso, o GPT-3 inaugurou o paradigma do *few-shot learning*, no qual o modelo aprende a executar uma tarefa com apenas alguns exemplos fornecidos no *prompt*, sem necessidade de treinamento adicional. Esse comportamento também se atende às modalidades *zero-shot* e *one-shot*, evidenciando que o modelo internalizou padrões de linguagem suficientemente gerais para extrapolar a partir de instruções mínimas. Essa capacidade influenciou diretamente a criação de aplicações como o ChatGPT, que exploram a versatilidade do modelo por meio de técnicas avançadas de engenharia de *prompt*.

Apesar de seu desempenho impressionante, destacam limitações fundamentais do GPT-3 que se relacionam diretamente com o fenômeno das alucinações. Por se tratar de um modelo puramente autorregressivo e probabilístico, o GPT-3 não possui mecanismos internos de verificação factual, o que o leva a gerar respostas plausíveis, porém incorretas, sempre que as instruções do usuário são ambíguas, quando faltam evidências no conjunto de treinamento ou quando o modelo enfrenta lacunas informacionais. Os autores também identificaram a tendência do modelo de reforçar vieses presentes nos dados, fabricar detalhes inexistentes e apresentar excesso de confiança em respostas falsas - características que tornaram o GPT-3 um dos primeiros modelos a evidenciar com clareza o comportamento conhecido como *alucinação*.

3.3 ALUCINAÇÕES EM MODELOS DE LINGUAGEM

O desenvolvimento dos modelos de linguagem de larga escala ampliou de forma significativa sua capacidade de produzir textos coerentes, detalhados e contextualmente adequados. Contudo, essa mesma capacidade trouxe maior visibilidade a um problema intrínseco a tais sistemas: o fenômeno das *alucinações*. O termo refere-se à produção de conteúdos incorretos, logicamente inconsistentes ou totalmente inventados, mesmo quando o modelo aparenta compreender perfeitamente a tarefa. Segundo a revisão conduzida por Ji et al. (2023), as alucinações constituem uma limitação estrutural dos modelos autorregressivos e permeiam tanto modelos de médio porte quanto arquiteturas avançadas como o GPT-3 e seus sucessores.

De modo geral, as alucinações podem ser divididas em duas grandes categorias, conforme a relação entre a resposta gerada e as informações disponíveis (Ji et al., 2023):

1. **Alucinações intrínsecas**

   Ocorrem quando o modelo produz conteúdo que contradiz diretamente o texto de entrada ou o próprio contexto fornecido. São comuns em tarefas de sumarização, tradução e perguntas e respostas, manifestando-se, por exemplo, na troca de datas, nomes, valores ou eventos claramente presentes na entrada.

2. **Alucinações extrínsecas**

> Manifestam-se quando a resposta gerada não pode ser verificada com base no contexto, seja por ausência de evidências, seja por extrapolação indevida. A resposta pode até ser verdadeira, mas não está fundamentada em nenhuma informação disponível. Esse tipo de alucinação é particularmente relevante em tarefas abertas e em perguntas que exploram mitos ou crenças amplamente difundidas.

Além dessa classificação central, estudos recentes propõem subdivisões adicionais, incluindo alucinações factuais (erros de conhecimento), contextuais (uso indevido do enunciado), lógicas (raciocínios inválidos) e de saturação (quando o modelo recusa-se a admitir desconhecimento). Essa pluralidade de formas evidencia a profundidade do problema e sua abrangência em diferentes tipos de tarefa.

Diversos fatores estruturais contribuem para o surgimento das alucinações. O primeiro deles decorre da natureza autorregressiva dos modelos: a geração de cada token baseia-se em distribuições probabilísticas condicionadas unicamente aos tokens anteriores, sem mecanismos internos de verificação factual. Isso significa que pequenas incertezas acumulam-se ao longo da sequência, abrindo espaço para desvios semânticos ou invenção de detalhes. Outro fator relevante é o processo de pré-treinamento, que utiliza grandes bases de dados extraídas da internet. Esse corpus massivo contém informações contraditórias, imprecisas, desatualizadas ou simplesmentes falsas, que acabam sendo internalizadas nos parâmetros do modelo.

A compressão do conhecimento nesses parâmetros também gera limitações importantes. Embora os modelos capturem padrões de linguagem em larga escala, eles perdem granularidade factual, apresentando dificuldades em recuperar informações específicas com precisão. Além disso, o processo de geração não inclui qualquer forma de *grounding*, isto é, comparação sistemática entre a resposta produzida e fontes externas confiáveis. Desse modo, o modelo baseia-se exclusivamente em seus próprios parâmetros, o que podem, naturalmente, conter vieses e lacunas.

Somam-se a essas causas estruturais diversos fatores contextuais que amplificam a probabilidade de alucinação. Instruções ambíguas ou *prompts* incompletos induzem o modelo a preencher lacunas com inferências estatísticas; perguntas sugestivas favorecem o *prompt bias*, levando o modelo a confirmas falsas premissas; e tarefas de múltiplas etapas aumentam a chance de inconsistências lógicas ao longo da geração.

Compreender esse fenômeno é fundamental para avaliar de forma crítica o uso dos modelos de linguagem, especialmente em aplicações sensíveis. É também o ponto de partida para o desenvolvimento de estratégias de mitigação, que buscam reduzir a incidência de erros por meio de verificação interna, recuperação de conhecimento externa ou técnicas de pós-edição. A fundamentação apresentada nesta seção estabelece o arcabouço necessário para a escolha e a implementação dos métodos experimentais que serão discutidos nos capítulos seguintes.

4 METODOLOGIA

A metodologia adotada neste trabalho foi estruturada com o objetivo de avaliar, de forma sistemática e comparável, diferentes estratégias de mitigação de alucinações em modelos de linguagem de larga escala. Para isso, foram definidos procedimentos experimentais padronizados, envolvendo a preparação do ambiente computacional, a construção do banco vetorial, a seleção dos conjuntos de dados, a implementação dos métodos de mitigação e a definição dos critérios de avaliação.

Os experimentos foram conduzidos a partir de um modelo base, o Llama3.2 3B, operando em condições controladas, às quais foram progressivamente incorporadas técnicas de mitigação cada vez mais sofisticadas. Entre elas, destacam-se:

(i) métodos baseados em recuperação de conhecimento externo (RAG);

(ii) técnicas de verificação interna estruturada, como o CoVe; e

(iii) estratégias de pós-edição com múltiplos agentes, representados pelo pipeline *Answer + Reviewer*.

O presente capítulo descreve, em detalhes, todo o processo metodológico utilizado: desde a configuração do ambiente experimental até os protocolos adotados para garantir a reprodutibilidade dos resultados. As seções seguintes apresentam os componentes essenciais desse processo.

4.1 AMBIENTE E FERRAMENTAS

O ambiente experimental foi configurado com o objetivo de assegurar reprodutibilidade e controle sobre as variáveis envolvidas nos experimentos. As ferramentas, modelos e recursos empregados estão listados a seguir:

- **Modelos de linguagem:** foi utilizado o *Llama3.2 3B* como modelo base para geração das respostas, operando com parâmetros determinísticos (`temperature = 0` e `top_p = 1`) a fim de eliminar variações estocásticas durante a inferência.
- **Modelos auxiliares:** o modelo *all-MiniLM-L6-v2* foi empregado para geração de embeddings semânticos utilizados no processo de recuperação vetorial, enquanto o *ms-marco-MiniLM-L6-v2* foi utilizado como *cross-encoder* para reranqueamento dos documentos recuperados.
- **Banco vetorial:** o *ChromaDB* foi utilizado como sistema de armazenamento e indexação dos embeddings extraídos dos documentos da Wikipedia, permitindo recuperação eficiente e ordenada de evidências textuais.
- **Conjuntos de dados:** foram utilizados os dados do *TruthfulQA* e *Arc Challenge*, empregados para avaliar, respectivamente, a veracidade factual e a capacidade de raciocínio das respostas geradas pelo modelo.

- **Métricas de avaliação:** o desempenho foi quantificado por meio das métricas *MC1* e *MC2* (para o TruthfulQa), que medem a capacidade do modelo de selecionar a alternativa mais verdadeira e de atribuir maior pontuação factual a opção correta; e pela *acurácia*, empregada na avaliação do *ARC Challenge*.
- **Infraestrutura computacional:** todos os experimentos foram executados em ambiente local equipado com uma GPU *NVIDIA RTX 4070*, com 8GB de memória dedicada, garantindo recursos computacionais adequados para execução dos métodos avaliados.

4.2 BANCO VETORIAL

Para a composição do banco vetorial utilizado neste trabalho, adotou-se o conjunto de dados *wikimedia/wikipedia*, disponibilizado pela plataforma HuggingFace (Foundation, s.d.), que contém o conteúdo completo da Wikipédia em inglês extraído por meio do *Wikimedia Dumps*. Esse dataset é amplamente utilizado em sistemas de recuperação e em métodos de RAG, constituindo uma das fontes de conhecimento externo mais comuns em estudos de mitigação de alucinações.

O corpus contém mais de 6,41 milhões de documentos, cada um estruturado em três campos principais: *URL*, *título* e *texto*. Essa organização facilita tanto indexação quanto a inspeção posterior dos trechos recuperados nos experimentos.

4.2.1 Processo de Contrução do Banco Vetorial

O banco vetorial utilizado nos métodos RAG, *RAG + MMR* e *RAG + Reranqueamento* foi construído seguindo três etapas principais:

1. **Pré-processamento dos textos:** os documentos do dataset foram carregados e processados, preservando o conteúdo integral das páginas.
2. **Geração dos embeddings:** cada documento foi convertido em um vetor semântico utilizando o modelo *all-MiniLM-L6-v2*, amplamente utilizado em tarefas de similaridade textual e recuperação densa.
3. **Indexação no ChromaDB:** os embeddings e seus metadados (título, URL e texto) foram armazenados no *ChromaDB*, possibilitando buscas eficientes por similaridade do cosseno.

Esse pipeline permite que cada pergunta seja convertida em um *embedding* e comparada ao banco vetorial, retornando os documentos mais relevantes para compor o contexto utilizado pelo modelo.

4.3 CONJUNTO DE DADOS

A avaliação das técnicas de mitigação de alucinações foi conduzida por meio de dois conjuntos de dados amplamente utilizados na literatura: *TruthfulQA* e *ARC Challenge*. Cada um deles examina dimensões distintas do desempenho dos modelos de linguagem, permitindo uma análise abrangente que contempla tanto a veracidade factual das respostas quanto a capacidade de raciocínio e inferência.

4.3.1 TruthfulQA

O *TruthfulQA* (Lin; Hilton; Evans, 2022) é atualmente um dos principais benchmarks para avaliação da veracidade em modelos de linguagem. O conjunto foi construído com o objetivo de medir a tendência dos modelos a reproduzir afirmações falsas, desinformação popular ou mitos amplamente difundidos. Desse modo, ele avalia diretamente a propensão dos modelos a alucinar.

O dataset possui duas modalidades: (i) respostas abertas e (ii) múltiplas escolhas. Neste trabalho, utilizou-se exclusivamente a versão *multiple choice*, que apresenta, para cada pergunta, alternativas contendo tanto uma resposta verdadeira quanto respostas falsas elaboradas para induzir erros. Esse formato permite isolar a habilidade do modelo de identificar a afirmação factualmente correta em um contexto controlado, facilitando a mensuração objetiva da ocorrência de alucinações.

Por sua natureza, o TruthfulQA é particularmente adequado para avaliar cenários em que o modelo deve resistir a informações incorretas e evitar inferências não fundamentadas.

4.3.2 ARC Challenge

Para complementar a avaliação de veracidade factual, adotou-se o benchmark AI2 Reasoning Challenge (ARC) (Clark et al., 2018). Diferentemente do TruthfulQA, que mede principalmente a adesão à veracidade factual, o ARC avalia a capacidade de raciocínio científico e inferência lógica.

O conjunto é composto por questões de múltipla escolha originalmente utilizadas em exames escolares de ciências nos Estados Unidos, divididos em dois subconjuntos: *ARC Easy* e *ARC Challenge*. Este último, utilizado neste trabalho, contém questões substancialmente mais complexas, que requerem raciocínio multietapas, compreensão de fenômenos naturais e análise contextual mais profunda.

Embora o ARC Challenge não seja um benchmark de veracidade, ele é relevante para identificar alucinações relacionadas a falhas de inferência, interpretações equivocadas do enunciado e inconsistências lógicas, aspectos que não são capturados pelo TruthfulQA. Em conjunto, ambos os datasets fornecem uma visão abrangente do comportamento dos modelos frente a diferentes tipos de desafios cognitivos.

4.4 AVALIAÇÕES E MÉTRICAS

A avaliação das técnicas de mitigação de alucinações foi conduzida com base em três métricas principais: *MC1* e *MC2*, aplicadas ao conjunto *TruthfulQA*, e *acurácia*, empregada no *ARC Challenge*. Essas métricas capturam dimensões complementares do desempenho dos modelos de linguagem, permitindo uma análise equilibrada entre *veracidade factual* e *capacidade de raciocínio*.

4.4.1 Métrica MC1

A métrica *MC1* (Multiple-Choice 1) consiste na seleção direta da alternativa considerada verdadeira pelo benchmark. Para cada questão, o modelo deve escolher uma única resposta dentre as opções apresentadas. Diferentemente de métricas baseadas em geração livre, o *MC1* avalia exclusivamente a capacidade do modelo de identificar a alternativa correta diante de respostas concorrentes, muitas das quais são formuladas intencionalmente para induzir erros.

Por exemplo, dada uma questão de múltipla escolha e suas respectivas alternativas, o modelo deve indicar diretamente aquela que considera correta. A métrica MC1 avalia essa escolha de forma idêntica a acurácia tradicional: calcula-se simplesmente a proporção de respostas em que o modelo selecionou a alternativa correta, dividida pelo total de questões avaliadas. Em outras palavras, o MC1 mede apenas a taxa de acertos diretos, sem levar em conta o grau de confiança do modelo ou sua capacidade de diferenciar a factualidade entre as alternativas disponíveis.

Assim, o *MC1* fornece uma medida direta da precisão factual, sendo especialmente útil para quantificar a propensão do modelo a selecionar respostas alinhadas com mitos populares ou informações incorretas.

4.4.2 Métrica MC2

A métrica *MC2* (Multiple-Choice 2), proposto originalmente por Lin, Hilton e Evans (2022), avalia a consistência factual do modelo em um nível mais fino. No benchmark oficial, essa métrica utiliza probabilidades token a token para medir preferência do modelo pela alternativa correta. Entretanto, como essa abordagem depende de detalhes internos do modelo, ela não pode ser aplicada diretamente ao *Llama3.2 3B*.

Para contornar essa limitação, este estudo adotou uma aproximação amplamente utilizada na literatura recente: cada alternativa da questão recebe um *score* de factualidade entre 0 e 100, atribuído pelo próprio modelo. A alternativa com maior pontuação é então selecionada como resposta final.

Essa aproximação preserva a essência da métrica original, permitindo avaliar:

- a confiança relativa do modelo em cada alternativa;

- sua capacidade de distinguir afirmações factuais de afirmações plausíveis, porém incorretas;
- sua sensibilidade a mitos amplamente difundidos.

Como exemplo, suponha uma questão com quatro alternativas. O modelo pode não selecionar explicitamente a resposta correta no *MC1*, mas ao atribuir pontuações às alternativas, pode julgar a resposta verdadeira como a mais factual, resultando em acerto no *MC2*. Assim, essa métrica captura um nível mais fino de discernimento factual, avaliando não apenas a escolha final do modelo, mas sua capacidade de diferenciar entre afirmações verdadeiras, parcialmente verdadeiras ou claramente falsas.

4.4.3 Acurácia (ARC Challenge)

No conjunto de dados *ARC Challenge*, a avaliação é realizada por meio de acurácia simples, isto é, o percentual de questões em que o modelo seleciona a alternativa correta. Ao contrário do TruthfulQA, que testa explicitamente a veracidade factual, o ARC Challenge avalia a capacidade do modelo de raciocinar sobre fenômenos científicos, interpretar descrições e resolver problemas multietapas.

Embora não seja um teste de factualidade estrita, erros no ARC Challenge podem envolver inferências incorretas, associações imprecisas e conclusões que não seguem logicamente da premissa.

4.4.4 Complementaridade das métricas

O uso combinado das três métricas, *MC1*, *MC2* e *acurácia*, permite uma avaliação equilibrada entre:

- **Veracidade factual** - o quanto o modelo evita de afirmações falsas (TruthfulQA).
- **Coerência inferencial** - o quanto o modelo raciocina corretamente sobre o enunciado (ARC Challenge)

Assim, os resultados apresentados refletem tanto a fidelidade factual quanto a robustez do raciocínio, fornecendo uma visão abrangente do impacto das técnicas de mitigação avaliadas.

4.5 MÉTODOS EXPERIMENTADOS

4.5.1 Método 1 - Baseline (Modelo Original)

O primeiro método avaliado corresponde ao uso direto do modelo Llama3.2 3B sem qualquer técnica adicional de mitigação de alucinações. Esse método estabelece a linha de base para comparação com as demais abordagens desenvolvidas neste estudo, permitindo

mensurar a capacidade intrínsica do modelo de responder corretamente às questões sem apoio externo, sem recuperação de informação e sem etapas auxiliares de verificação.

O modelo foi instruído apenas a selecionar a opção mais verdadeira com base em seu próprio conhecimento pré-treinado. O prompt utilizado foi propositalmente minimalista, de modo a não introduzir qualquer mecanismo de mitigação:

Answer the following multiple-choice question:

Question: {question}

Choices: {1.Choice1, 2.Choice2, ...}

Answer only with the number of the correct choice (1,2,3,...).

Dado que o método base não utiliza qualquer forma de recuperação de conhecimento externo, reranqueamento ou verificação interna, seu desempenho reflete exclusivamente o conhecimento armazenado nos parâmetros do modelo e sua propensão natural a reproduzir fatos ou alucinações.

4.5.2 Método 2 - Recuperação de Informação (RAG)

Para o segundo método, o RAG foi implementado por meio de um pipeline composto por três etapas principais:

i recuperação de documentos relevantes;

ii construção de um *prompt* enriquecido com o contexto externo; e

iii geração da resposta.

Para a etapa de recuperação, foi utilizado o banco vetorial baseado no *ChromaDB*, mencionado anteriormente. Cada pergunta dos benchmarks foi utilizada como mecanismo de busca vetorial, que retornou os cinco documentos semanticamente mais próximos, segundo a similaridade do cosseno.

O contexto recuperado foi então incorporado ao *prompt*, exemplificado abaixo, instruindo o modelo a utilizar exclusivamente essas informações para selecionar a alternativa correta.

You are answering a multiple-choice factuality question.

Use ONLY the information in the context below to decide which answer is most truthful.

Context: {documentos recuperados}

Question: {question} Choices: {1. choice_1, 2. choice_2, ...}

Answer only with the number of the correct choice.

A aplicação do RAG tem como motivação principal investigar se o fornecimento explícito de evidências externas é capaz de reduzir a incidência de respostas inventadas ou

factualmente incorretas. Entretanto, como a eficácia dessa abordagem depende fortemente da qualidade e relevância do contexto recuperado, não se presume que o RAG melhore o desempenho em todos os tipos de pergunta, especialmente em benchmarks projetados para desafiar o conhecimento enciclopédico dos modelos.

Assim, embora o método RAG tenda teoricamente a melhorar a veracidade das respostas, sua eficácia prática em benchmarks não é garantida e deve ser verificada empiricamente. Os resultados obtidos com esse método permitem avaliar se o uso de informações externas é suficiente para reduzir alucinações ou se técnicas adicionais, como reranqueamento, verificação em cadeia ou sistemas multiagentes, são necessários para alcançar melhorias consistentes.

Durante a implementação do método RAG simples, observou-se que a utilização do parâmetro `search_type - "mmr"` (*Maximal Marginal Relevance*) no mecanismo de recuperação do ChromaDB produziu resultados superiores aos obtidos com a busca vetorial padrão baseada apenas em similaridade. O MMR atua como uma forma de reranqueamento leve e não supervisionado, selecionando documentos que equilibrem relevância e diversidade, reduzindo a redundância entre os textos recuperados. Embora não constitua um método independente - e tampouco utilize modelos neurais de reranquamento, como os empregados no Método 3 - o MMR ainda pode ser compreendido como um refinamento na etapa de recuperação, contribuindo para um contexto mais informativo em certos tipos de pergunta. Por essa razão, seus efeitos serão considerados de forma complementar na análise dos resultados do método RAG.

4.5.3 Método 3 - RAG + Reranqueamento

Embora o método 2 apresente ganhos potenciais na mitigação de alucinações ao fornecer ao modelo um conjunto de evidências externas, sua eficácia depende diretamente da qualidade do material recuperado. Em cenários em que a busca inicial não retorna os dados mais relevantes, o modelo pode receber um contexto pouco informativo ou até mesmo enganoso, o que reduz o impacto positivo da técnica. Para enfrentar essa limitação, este estudo implementa uma segunda variação baseada no uso combinado de recuperação e reranqueamento.

O método consiste em duas etapas principais: (i) uma recuperação inicial mais ampla, usando a busca vetorial aproximada, e (ii) um reranqueamento subsequente que ordena os documentos recuperados de acordo com sua relevância contextual em relação à pergunta. Para a etapa de busca, utilizou-se o mesmo banco vetorial empregado no método anterior, baseado em *ChromaDB* e embeddings *all-MiniLM-L6-v2*. No entanto, ao invés de recuperar somente os cinco documentos mais similares, ampliou-se a busca para os vinte documentos mais próximos, de forma a capturar um espectro mais amplo de possíveis informações relacionadas.

Na etapa de reranqueamento, foi utilizado um modelo *cross-encoder* treinado no

conjunto *MS MARCO* (*cross-encoder/ms-marco-MiniLM-L6-v2*). Esse tipo de modelo recebe a pergunta e cada documento como entrada conjunta, aprendendo a validar, de maneira contextual, a relevância do documento para aquela pergunta específica. Diferentemente da recuperação vetorial, que opera apenas sobre similaridade de embeddings, o cross-encoder utiliza atenção cruzada entre a consulta e o documento para produzir uma estimativa mais precisa de relevância semântica.

O pipeline implementado funciona da seguinte maneira: para cada pergunta, os vinte documentos inicialmente recuperados são reranqueados pelo cross-encoder, ordenados pela relevância e, então, apenas os cinco melhores documentos são selecionados e concatenados para compor o contexto final utilizado pelo modelo. Esse contexto refinado é incorporado ao mesmo formato de *prompt* do RAG simples, instruindo o modelo a basear sua decisão exclusivamente nas informações fornecidas.

Esse técnica é motivada pelos achados de Lewis et al. (2020) e estudos posteriores, que apontam que modelos de linguagem - especialmente aqueles de menor porte, como o Llama3.2 3B - podem ignorar informações relevantes quando o contexto contém ruído contextual ou documentos mal alinhados à pergunta. O reranqueamento atua justamente reduzindo esse ruído, aumentando a probabilidade de que o contexto fornecido contenha evidências diretamente relacionadas à resposta correta.

Apesar disso, o método apresenta desafios importantes quando aplicados ao *TruthfulQA*. Como o benchmark é composto por perguntas projetadas para explorar mitos populares e desinformação amplamente difundida, nem sempre a Wikipédia contém informações explícitas que refutem tais crenças. Em alguns casos, documentos bem ranqueados podem mencionar controversas públicas, debates científicos ou afirmações incorretas amplamente difundidas, o que pode influenciar negativamente o processo de pontuação no *MC2*. Assim, embora o reranqueamento tenda a reduzir ruído e aumentar a relevância média do contexto, seu impacto efetivo na mitigação de alucinações deve ser verificado empiricamente.

No contexto deste trabalho, o *RAG + Rerank* representa uma abordagem intermediária entre a recuperação simples e métodos mais sofisticados de mitigação, como agentes verificadores e *chain-of-verification*. Sua inclusão no estudo permite avaliar até que ponto o uso de um contexto mais refinado contribui para melhorar o desempenho factual do modelo, medidas pelas métricas *MC1* e *MC2*.

4.5.4 Método 4 - Chain-of-Verification (CoVe)

A quarta abordagem avaliada neste estudo consiste na aplicação do método conhecido como CoVe (Chain-of-Verification) (Dhuliawala et al., 2024), uma técnica de mitigação de alucinações baseada na verificação interna da resposta gerada pelo próprio modelo. Diferentemente dos métodos baseados em recuperação de informações externas, como o RAG, o CoVe opera exclusivamente sobre o conhecimento presente nos parâmetros

do modelo, realizando um processo estrutuuturado em múltiplas etapas que visa identificar inconsistências, corrigir afirmações falsas e reduzir a produção de conteúdo não factual.

O CoVe parte do pressuposto de que um modelo de linguagem, apesar de suscetível a alucinações, é capaz de avaliar criticamente suas próprias respostas quando incentivado a decompor a informação em etapas verificáveis. Esse mecanismo foi inspirado em estudos recentes sobre autoavaliação e auto-correção em LLMs, que demonstram que modelos podem detectar e corrigir parte de seus próprios erros quando submetidos a um processo explícito de verificação estruturada.

No presente trabalho, o CoVe foi implementado por meio de um pipeline composto por três etapas:

1. **Geração da resposta inicial:** o modelo responde diretamente à pergunta, de maneira semelhante ao método base. Essa resposta é utilizada como entrada para as etapas subsequentes.
2. **Geração das etapas de verificação:** o modelo recebe a pergunta e a resposta inicial e é instruído a decompor a resposta em um conjunto de fatos ou afirmações verificáveis. O objetivo é transformar a resposta inicial em uma coleção de elementos menores que possam ser avaliados individualmente.
3. **Produção da resposta final corrigida:** utilizando as etapas de verificação, o modelo reavalia sua resposta original, corrindo informações incorretas ou não suportadas. Caso identifique ausência de evidências suficientes, o modelo é instruído a utilizar respostas conservadoras, como "I don't know", evitando a fabricação de conteúdo não factual.

A resposta utilizada para avaliação é sempre a resposta final corrigida, e não a resposta primária. Para o *TruthfulQA*, que utiliza formato de múltipla escolha, o CoVe foi integrado ao protocolo de avaliação da seguinte forma: após a geração da resposta final, o modelo recebe as alternativas disponíveis e seleciona apenas o número correspondente à opção que mais se alinha ao conteúdo verificado. No cálculo do *MC2*, cada alternativa é comparada com a resposta final por meio de uma mecanismo de pontuação factual, permitindo avaliar até que ponto a resposta corrigida converge para a alternativa correta.

A principal vantagem do CoVe é que ele não depende de bases de conhecimento externas nem de mecanismos complexos de recuperação. Sua eficácia decorre da capacidade do modelo de examinar criticamente suas próprias respostas, identificar potenciais erros e reconstruir a resposta com maior precisão factual. Esse tipo de de técnica é especialmente relevante para modelos de menor porte, como o Llama3.2 3B utilizado neste estudo, que frequentemente apresentam dificuldades em integrar informações externas (como no RAG), mas respondem bem a mecanismos de verificação interna.

No entanto, o CoVe apresenta limitações. Como todo processo baseado em autoavaliação, ele depende fortemente da capacidade do modelo de reconhecer seus próprios erros, o que nem sempre ocorre de maneira consistente. Além disos, o método requer múltiplas

chamadas ao modelo para cada pergunta, geralmente três, o que aumenta significativamente o custo computacional em comparação ao método base.

Apesar dessas limitações, o CoVe constitui uma estratégia promissora para mitigação de alucinações, oferecendo uma abordagem complementar às técnicas baseadas em recuperação. Sua inclusão neste estudo permite avaliar até que ponto a verificação interna pode melhorar a precisão factual das respostas sem apoio de fontes externas.

4.5.5 Método 5 - Agente Revisor (Answer + Reviewer)

O último método avaliado neste estudo consiste em uma estratégia de mitigação baseada em pós-edição (*post-editing*), utilizando uma arquitetura de dois agentes: um agente gerador (*Answerer*) e um agente revisor (*Reviewer*). Essa abordagem busca reduzir alucinações por meio da revisão crítica da resposta inicial, permitindo que o modelo produza uma versão corrigida ou mais conservadora quando não houver evidências suficientes para uma conclusão segura.

O processo ocorre em duas etapas principais. Na primeira, o agente gerador recebe a pergunta e produz uma resposta inicial, sem acesso a informações externas e de modo semelhantes ao método base. Essa resposta pode conter erros factuais, afirmações não verificáveis ou excesso de confiança. Na segunda etapa, o agente revisor recebe a pergunta e a resposta inicial e é instruído a avaliá-la criticamente. O revisor deve:

- identificar afirmações incorretas ou não suportadas;
- corrigir trechos imprecisos quando possível;
- adotar respostas conservadoras (*I don't know*) quando a resposta inicial não puder ser justificada.

Diferentemente do método CoVe, apresentado anteriormente, o método de pós-edição direta não solicita ao modelo que decomponha a resposta em etapas verificáveis ou microafirmações factuais. Em vez disso, o agente revisor opera diretamente sobre a resposta final, realizando a análise holística sem produzir explicações intermediárias. Por esse motivo, o método pode ser considerado mais simples e computacionalmente mais leve que o CoVe, ainda que potencialmente menos robusto em cenários que exijam verificação profunda.

Após a revisão, a resposta final corrigida é utilizada para a avaliação nas métricas *MC1* e *MC2*, seguindo o mesmo protocolo dos métodos anteriores. No caso do *MC1*, o modelo recebe a resposta revisada e as alternaticas da questão e deve selecionar o número correspondente à opção compatível com o conteúdo corrigido. Para o *MC2*, cada alternativa é comparada com a resposta final pelo próprio modelo, que atribui uma pontuação de factualidade entre 0 a 100, sendo selecionada aquela com maior pontuação.

As vantagens do método incluem sua simplicidade e sua capacidade de reduzir respostas inventadas por meio da adoção de uma postura mais cautelosa pelo agente

revisor. No entanto, sua eficácia depende da habilidade do modelo em detectar seus próprios erros de forma direta, sem a estruturação adicional oferecida pelas verificações intermediárias do CoVe. Ainda assim, o método oferece uma perspectiva complementar às demais técnicas avaliadas, permitindo analisar até que ponto a pós-edição interna, sem evidências externas, pode contribuir para reduzir alucinações.

5 RESULTADOS E DISCUSSÕES

5.1 APRESENTAÇÃO GERAL DOS RESULTADOS

Nesta seção são apresentados e discutidos os resultados obtidos pelos seis métodos avaliados sobre os benchmarks *TruthfulQA* e *ARC Challenge*. Todos os experimentos foram conduzidos de forma padronizada, seguindo o protocolo descrito no capítulo anterior, permitindo comparar diretamente o impacto de cada técnica de mitigação de alucinações.

Os métodos avaliados incluem: (i) modelo base sem mitigação, (ii) RAG simples, (iii) RAG com reranqueamento por diversidade (MMR), (iv) RAG com reranqueamento neural (cross-encoder), (v) CoVe (Chain-of-Verification) e (vi) o método de pós-edição com agente revisor (*Answer + Reviewer*).

Para todos os métodos, os resultados são apresentados, na tabela 1, utilizando as métricas *MC1* e *MC2* no *TruthfulQA*, e acurácia no *ARC Challenge*, todas expressas em uma escala de 0 a 1 para facilitar comparação. Em seguida, as técnicas são analisadas comparativamente, destacando os ganhos e limitações observadas.

Método	**MC1**	**MC2**	**Arc Challenge**
Modelo Base	0.4350	0.2650	0.3814
RAG Simples	0.2700	0.3600	0.5657
RAG + MMR	0.3200	0.3750	0.5486
RAG + Rerank	0.3000	0.3800	0.5742
CoVe	0.7750	0.5600	0.4351
Answer + Reviewer	0.2800	0.5000	0.3421

Tabela 1 – Resultados no TruthfulQA (200 amostras) (MC1 e MC2) e ARC Challenge (1172 amostras)

5.1.1 Baseline

Os resultados obtidos no TruthfulQA mostram que o modelo apresenta um desempenho moderado. O valor do MC1 atingiu 0,4350 indicando que o modelo acertou diretamente 43% das questões ao selecionar a alternativa correta. Por outro lado, a métrica MC2, que avalia a consistência factual das alternativas a partir de *scores* gerados pelo próprio modelo, foi inferior (0,2650). Essa diferença sugere que, embora o modelo seja capaz de identificar corretamente algumas respostas, ele apresenta dificuldade para discriminar, de maneira mais fina, o grau de factualidade das opções disponíveis. Esse comportamento está alinhado com limitações esperadas em modelos menores, cuja capacidade de verificação interna é reduzida

No ARC Challenge, o modelo base alcançou acurácia de 0,3814. Esse valor é condizente com o desempenho típico de modelos de pequeno porte em tarefas de raciocínio escolar, que exigem inferências seguidas e conhecimento contextual que muitas vezes não está disponível de forma explícita nos parâmetros do modelo. Assim, os resultados

reforçam que o modelo base apresenta um desempenho razoável, porém limitado, tanto em veracidade factual quanto em raciocínio.

5.1.2 RAG Simples

Os resultados obtidos com o método *RAG Simples* evidenciam que a recuperação de informações externas não garante, por si só, melhorias consistentes na veracidade factual, especialmente em benchmarks projetados para induzir erros, como o *TruthfulQA*.

No *TruthfulQA*, a métrica *MC1* caiu de 0,4350 (modelo base) para 0,2700, indicando que o modelo passou a selecionar com menor frequência a alternativa correta. Esse comportamento sugere que o contexto recuperado da Wikipédia nem sempre é relevante para as questões propostas, que frequentemente envolvem mitos populares, armadilhas semânticas ou afirmações não encontradas em textos enciclopédicos. Nesses casos, o contexto externo introduz ruído adicional, competindo com o conhecimento interno do modelo e prejudicando a decisão final.

Por outro lado, a métrica *MC2* apresentou melhora, passando de 0,2650 para 0,3600. Esse aumento indica que, embora o modelo tenha errado mais frequentemente a alternativa correta, ele passou a avaliar as alternativas com maior sensibilidade factual. Ou seja, o contexto externo ajudou o modelo a distinguir melhor nuances de veracidade, ainda que isso não tenha sido suficiente para melhorar o acerto direto (*MC1*).

O cenário é significativamente diferente no *ARC Challenge*. A acurácia aumentou de 0,3814 para 0,5657, demonstrando um ganho expressivo. Esse resultado sugere que, ao contrário do TruthfulQA, o ARC contém questões cujo conteúdo está mais alinhado ao material enciclopédico recuperado. Em tarefas de raciocínio científico, a Wikipédia fornece informações diretamente úteis, o que auxilia o modelo a preencher lacunas e tomar decisões mais fundamentadas.

5.1.3 RAG com MMR

No método RAG + MMR, avaliou-se o impacto de aplicar o mecanismo MMR (Maximal Marginal Relevance) na etapa de recuperação de documentos. Diferentemente do RAG simples, que seleciona apenas os documentos mais similares à pergunta, o MMR busca equilibrar relevância e diversidade, reduzindo redundâncias e evitando que o contexto recuperado seja formado por textos muito parecidos entre si.

Os resultados no TruthfulQA mostram que essa estratégia traz ganhos moderados. A métrica MC1 aumentou de 0,2700 (RAG simples) para 0,3200, indicando que documentos menos redundantes ajudam o modelo a escolher a alternativa correta com mais frequência. De forma semelhante, a métrica MC2 elevou-se de 0,3600 para 0,3750, sugerindo que o modelo conseguiu atribuir pontuações factuais mais consistentes às alternativas após receber um contexto mais variado e informativo. Mesmo assim, o desempenho continua abaixo do modelo base em MC1, refletindo que o TruthfulQA contém perguntas

cujo conteúdo não aparece de forma explícita na Wikipédia, limitando o benefício da recuperação externa.

No ARC Challenge, o método apresentou uma leve redução na acurácia em relação ao RAG simples: de 0,5657 para 0,5486. Essa diferença relativamente pequena sugere que, embora o MMR elimine redundâncias, nem sempre a diversidade adicional favorece o raciocínio exigido pelo benchmark. Em algumas questões, documentos mais variados podem ser menos diretamente relevantes para a resolução da tarefa, prejudicando a inferência mesmo quando o contexto recuperado é mais amplo.

5.1.4 RAG com Reranqueamento Neural

No TruthfulQA, os resultados mostram um comportamento misto. A métrica MC1 atingiu 0,3000, ligeiramente inferior ao RAG + MMR e ainda abaixo do desempenho do modelo base (0,4350). Por outro lado, o MC2 apresentou leve aumento, alcançando 0,3800. Essa piora no desempenho da escolha da alternativa correta, mas melhora na consistência factual relativa sugere que o cross-encoder consegue refinar parte do contexto mais relevante, porém esse refinamento não se traduz, necessariamente, em maior capacidade de selecionar a alternativa correta em perguntas projetadas para provocar respostas enganosas. Como no RAG simples, a dependência de informações externas permanece sensível à natureza das perguntas do TruthfulQA, que frequentemente não se alinham ao conteúdo enciclopédico disponível na Wikipédia.

No *ARC Challenge*, o método apresentou seu melhor desempenho, alcançando acurácia de 0,5742, o valor mais elevado entre todos os métodos baseados em recuperação. Esse resultado demonstra que o reranqueamento neural é especialmente efetivo em tarefas que envolvam raciocínio científico e interpretação de conteúdos escolares, onde a precisão e a adequação contextual das evidências recuperadas são determinantes para a escolha correta. Ao reduzir documentos pouco pertinentes e privilegiar textos altamente relacionados ao enunciado, o método consegue fornecer ao modelo um suporte informacional mais alinhado à tarefa.

5.1.5 Chain-of-Verification (CoVe)

No *TruthfulQa*, o CoVe apresentou o melhor desempenho entre todos os métodos avaliados. O valor de *MC1* atingiu 0,7750, representando uma melhora substancial em relação ao modelo base e superando inclusive os métodos baseados em recuperação. A métrica *MC2* também apresentou o maior valor até então (0,5600), indicando que o CoVe não apenas aumenta a probabilidade de o modelo selecionar a alternativa correta, mas também melhora sua capacidade de atribuir maior veracidade às alternativas corretas durante o processo de pontuação. Esses resultados mostram que a decomposição estruturada da resposta em verificações auxilia o modelo a refletir sobre sua própria saída, reduzindo alucinações intrínsecas.

No ARC Challenge, porém, o desempenho do CoVe apresentou uma queda, com acurácia de 0,4351, valor inferior ao obtido pelos métodos baseados em RAG simples e RAG com reranqueamento, mas ainda acima do baseline. Esse resultado indica que, embora a verificação interna seja extremamente eficaz para questões de factualidade, ela nem sempre melhora o desempenho em tarefas de raciocínio científico e inferencial, onde o acesso a informações externas (como artigos da Wikipédia) tende a oferecer maior suporte para a escolha correta.

5.1.6 Answer + Reviewer

No TruthfulQA, o método apresentou um comportamento misto. O valor de MC1 foi de 0,2800, inferior ao modelo base (0,4350), indicando que a revisão nem sempre resultou na escolha da alternativa correta. Esse resultado sugere que, em algumas situações, o revisor pode ter adotado respostas excessivamente conservadoras — como "I don't know"— ou reinterpretado a resposta inicial de forma a se distanciar da alternativa correta. Por outro lado, a métrica MC2 atingiu 0,5000, um valor expressivamente superior ao modelo base e maior que todas as variantes de RAG. Isso indica que, embora o revisor nem sempre identifique a alternativa correta, ele é mais eficaz em atribuir scores de factualidade que distinguem melhor as respostas verdadeiras das falsas. Ou seja, o revisor melhora a consistência factual global, mas não necessariamente a precisão direta.

No ARC Challenge, o método apresentou acurácia de 0,3421, inferior ao modelo base (0,3814). Esse resultado revela uma limitação importante: como o ARC exige raciocínio científico e inferência multietapas, respostas reescritas de forma conservadora ou excessivamente sintética podem dificultar o mapeamento correto entre a resposta revisada e as alternativas do enunciado. Assim, a pós-edição pode suavizar a precisão quando a tarefa depende menos de factualidade direta e mais de raciocínio contextual

6 CONCLUSÃO

O presente trabalho teve como objetivo investigar alternativas para mitigar alucinações em modelos de linguagem de larga escala, com foco no modelo *Llama3.2 3B*. A motivação central decorre do fato de que, apesar dos avanços recentes dos modelos de linguagem, esses sistemas permanecem suscetíveis à geração de informações incorretas, não verificáveis ou completamente inventadas. Tal limitação compromete a confiabilidade dos modelos em aplicações sensíveis, especialmente em áreas que exigem rigor factual, como direito e saúde.

Para enfrentar esse desafio, foram avaliadas diferentes estratégias de mitigação de alucinações, abrangendo métodos baseados em recuperação de informações (*Retrieval-Augmented Generation*), mecanismos de reranqueamento (MMR e reranqueamento neural), técnicas de verificação interna (*Chain-of-Verification*) e um método de pós-edição via agente revisor (*Answer + Reviewer*). Cada técnica foi implementada sob condições experimentais padronizadas e avaliada nos benchmarks *TruthfulQA* e *ARC Challenge*, que medem, respectivamente, veracidade factual e capacidade de raciocínio em múltiplas escolhas.

Os resultados mostram que não há uma solução única que maximize o desempenho em todas as tarefas. Métodos baseados em recuperação, como o RAG e suas variações, apresentaram melhorias expressivas no *ARC Challenge*, demonstrando que a incorporação de conhecimento externo é particularmente eficaz em tarefas que exigem inferência ou raciocínio científico básico. Por outro lado, técnicas de verificação interna, como o CoVe obtiveram resultados superiores no *TruthfulQA*, evidenciando maior capacidade de reduzir alucinações factuais diretamente na resposta do modelo.

Esses achados sugerem que diferentes tipos de alucinação - intrínseca e extrínseca - requerem abordagens distintas. Métodos de recuperação tendem a mitigar alucinações por ausência de conhecimento, enquanto técnicas de verificação interna são mais eficazes em corrigir erros de raciocínio ou afirmações infundadas. Assim, uma combinação híbrida entre recuperação de informação e verificação estruturada surge como um caminho promissor para o desenvolvimento de sistemas mais confiáveis.

Finalmente, espera-se que os achados aqui apresentados sirvam de base para pesquisas futuras voltadas à confiabilidade de modelos de linguagem, bem como para o desenvolvimento de ferramentas práticas que reduzam riscos e ampliem o uso seguro de sistemas de IA em ambientes críticos, contribuindo para uma adoção mais responsável e tecnicamente fundamentada dessas tecnologias.

REFERÊNCIAS

BOMMASANI, Rishi. On the opportunities and risks of foundation models. **arXiv preprint arXiv:2108.07258**, 2021.

BROWN, Tom et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877–1901, 2020.

CLARK, Peter; COWHEY, Isaac; ETZIONI, Oren; KHOT, Tushar; SABHARWAL, Ashish; SCHOENICK, Carissa; TAFJORD, Oyvind. Think you have solved question answering? try arc, the ai2 reasoning challenge. **arXiv preprint arXiv:1803.05457**, 2018.

DHULIAWALA, Shehzaad; KOMEILI, Mojtaba; XU, Jing; RAILEANU, Roberta; LI, Xian; CELIKYILMAZ, Asli; WESTON, Jason. Chain-of-verification reduces hallucination in large language models. In*:* FINDINGS of the Association for Computational Linguistics: ACL 2024. [S.l.*:* s.n.], 2024. p. 3563–3578.

DSA, Equipe. **10 Casos de Uso da IA no Direito**. [S.l.*:* s.n.], 2024. Acesso em 23 nov. 2025. Disponível em: https://blog.dsacademy.com.br/10-casos-de-uso-da-ia-no-direito/.

FOUNDATION, Wikimedia. **Wikimedia Downloads**. Disponível em: https://dumps.wikimedia.org.

JI, Ziwei et al. Survey of hallucination in natural language generation. **ACM computing surveys**, ACM New York, NY, v. 55, n. 12, p. 1–38, 2023.

LEWIS, Patrick et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in neural information processing systems**, v. 33, p. 9459–9474, 2020.

LIN, Stephanie; HILTON, Jacob; EVANS, Owain. Truthfulqa: Measuring how models mimic human falsehoods. In*:* PROCEEDINGS of the 60th annual meeting of the association for computational linguistics (volume 1: long papers). [S.l.*:* s.n.], 2022. p. 3214–3252.

MELO, João Ozorio de. **Corte dos EUA multa advogado e pune até adversário por "fabricações"do ChatGPT**. [S.l.*:* s.n.], 2025. Acesso em 23 nov. 2025. Disponível em: https://www.conjur.com.br/2025-set-25/corte-dos-eua-multa-advogado-e-pune-ate-adversario-por-fabricacoes-do-chatgpt/.

OUYANG, Long et al. Training language models to follow instructions with human feedback. **Advances in neural information processing systems**, v. 35, p. 27730–27744, 2022.

PORTILHO, Luciana. **TIC Saúde 2024**. [S.l.*:* s.n.], 2024. Acesso em 23 nov. 2025. Disponível em: https://cetic.br/media/analises/tic_saude_principais_resultados_2024.pdf.

RADFORD, Alec; NARASIMHAN, Karthik; SALIMANS, Tim; SUTSKEVER, Ilya et al. Improving language understanding by generative pre-training. San Francisco, CA, USA, 2018.

SILVA, Salef Gabriel Gamberini. **Análise comparativa de estratégias de mitigação de alucinações em modelos de linguagem de grande escala**. 2025. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Mato Grosso do Sul, Faculdade de Computação, Campo Grande. Acesso em: 4 nov. 2025. Disponível em: https://posgraduacao.ufms.br/portal/trabalho-arquivos/download/15338.

TURING, Alan M. Computing machinery and intelligence. In*:* PARSING the Turing test: Philosophical and methodological issues in the quest for the thinking computer. [S.l.]: Springer, 1950. p. 23–65.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N; KAISER, Łukasz; POLOSUKHIN, Illia. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.