Análise Comparativa de Estratégias de Mitigação de Alucinações em Modelos de Linguagem de Grande Escala

Salef Gabriel Gamberini Silva

SERVIÇO DE PÓS-GRADUAÇÃO FACOM-UFMS

Data de Depósito:

Assinatura:_____

Análise Comparativa de Estratégias de Mitigação de Alucinações em Modelos de Linguagem de Grande Escala

Salef Gabriel Gamberini Silva

Orientador: *Profº Drº Renato Porfirio Ishii*

Dissertação entregue a Faculdade de Computação da Universidade Federal de Mato Grosso do Sul - FACOM-UFMS como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

UFMS - Campo Grande
Julho/2025

À minha mãe,
Régia,
que acreditou em mim muito antes de eu mesmo acreditar.

Agradecimentos

A Deus, por me conceder o dom da vida e, com ela, a capacidade de aprender e compreender.

À minha mãe, Régia, ao meu pai, Francisco Airton, e à minha irmã, Lara Maria, por sempre estarem ao meu lado, oferecendo todo o suporte, compreensão e carinho possíveis. Eu precisaria de mil vidas para retribuir o amor que recebi durante esta jornada.

Ao Professor Fabio do Vale, pela oportunidade da publicação do meu primeiro artigo. Às Professoras Ana Claudia e Edilene, por me auxiliarem durante o processo de inscrição no mestrado. A paciência de todos foi inestimável.

Ao meu orientador, Professor Renato Ishii, por toda a paciência, disponibilidade e compreensão nos momentos de dúvida e em minhas falhas. Agradeço também pelo bom humor sempre presente e por ter me aceitado como seu orientando.

Aos professores e funcionários da FACOM, em especial ao Professor Diego Padilha, pela orientação impecável e grande apoio; ao Professor Awdren de Lima, por me tornar mais criterioso; e ao Professor Ronaldo Alves, pela confiança e por compartilhar seu vasto conhecimento.

Ao meu amigo Alex Koseki, por me inspirar a seguir a carreira na área de computação. Ao meu amigo Rodrigo Araujo, pelas perspectivas valiosas. E aos meus amigos Mariela Nicodemos e Marcos Estevão, por caminharem ao meu lado mesmo nos momentos mais difíceis.

Abstract

Large Language Models (LLMs) have advanced rapidly, demonstrating remarkable capabilities in natural language understanding and generation. However, their tendency to "hallucinate" information represents a critical challenge for their professional adoption. This work surveys contemporary hallucination mitigation methods in LLMs, classifying them into distinct categories through a systematic literature review. Based on code availability for replication, four representative techniques (KCA, ICD, Wiki-Chat, RepE) were selected for empirical analysis on a single base model, the Llama2 7b. Evaluations using question-answering benchmarks and textual similarity metrics revealed variable performance: KCA excelled in question-answering tasks, WikiChat achieved the best results in token-matching metrics like ROUGE, and ICD demonstrated superiority in semantic precision. It is concluded that the effectiveness of mitigation methods depends on the task and evaluation metric, underlining the absence of a universal solution and the importance of a multifaceted approach.

Keywords: Large language models, mitigation of hallucinations, natural language processing, comparative analysis, machine learning, systematic review of literature.

Resumo

Modelos de Linguagem de Grande Escala (LLMs) avançaram rapidamente, demonstrando capacidades notáveis em compreensão e geração de linguagem natural. Contudo, sua tendência a "alucinar" informações representa um desafio crítico para sua adoção profissional. Este trabalho investiga e separa métodos contemporâneos de mitigação de alucinações em LLMs por meio de uma revisão sistemática da literatura, classificando-os em categorias distintas. Com base na disponibilidade de código para replicação, quatro técnicas representativas (KCA, ICD, Wiki-Chat, RepE) foram selecionadas para uma análise empírica sobre uma base de modelo unificada, o Llama2 7b. As avaliações, utilizando benchmarks de perguntas e respostas e medidas de similaridade textual, revelaram um desempenho variável. KCA destacou-se em tarefas de perguntas e respostas, WikiChat obteve os melhores resultados em medidas de correspondência de tokens como ROUGE, e ICD demonstrou superioridade em precisão semântica. Conclui-se que a eficácia dos métodos de mitigação depende da tarefa e da medida de avaliação, sublinhando a ausência de uma solução universal e a importância de uma abordagem multifacetada.

Palavras-chave: Modelos de linguagem de grande escala, mitigação de alucinações, processamento de linguagem natural, análise comparativa, aprendizado de máquina, revisão sistemática da literatura.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AM Aprendizado de Máquina

BERT *Bidirectional Encoder Representations from Transformers*

BART *Bidirectional and Auto-Regressive Transformer*

IA Inteligência Artificial

PPO *Proximal Policy Optimization*

GPT *Generative Pre-trained Transformer*

NLP *Natural Language Processing*

LLM *Large Language Models*

RAG *Retrieval-augmented Generation*

seq2seq *Sequence to Sequence*

BLEU *Bilingual Evaluation Understudy*

RepE *Representation Engineering*

ARC *Abstraction and Reasoning Corpus*

KCA *Knowledge Consistent Alignment*

ICD *Induce-then-Contrast Decoding*

RLHF *Reinforcement Learning from Human Feedback*

CAPÍTULO

1

Introdução

O amplo acesso aos modelos de linguagem de grande escala (LLMs) iniciou-se com o lançamento público em 2019 do GPT-2 (*Generative Pre-trained Transformer*) da OpenAI, que demonstrou seu potencial para várias tarefas linguísticas. Modelos subsequentes viram avanços significativos em capacidade e utilidade [3, 4], culminando no GPT-3. Este modelo mais potente ajudou a popularizar as aplicações de LLMs em diversas áreas, como programação de computadores [5, 6], atendimento ao cliente [7, 8] e criação de conteúdo [9, 10].

Conforme a adoção desses modelos foi crescendo, sua aplicação deixou de ser puramente experimental para se tornar parte integrante de indústrias e serviços, desde áreas de relativo baixo risco como desenvolvimento de software e análise financeira [11] até áreas de extremo risco como assistência médica [12, 13] e treinamento cirúrgico [14]. Entretanto, existe um obstáculo titânico para a adesão de LLMs na área profissional: a tendência dos modelos "alucinarem", ou seja, produzir conteúdo incoerente ou factualmente incorreto. Isso ocorre devido à natureza do funcionamento intrínseco desses geradores, pois as respostas entregues são baseadas em padrões encontrados nos dados em que foram treinados, sem uma compreensão verdadeira do conteúdo ou da realidade. Isso pode resultar em erros que variam de simples inconvenientes a falhas críticas com consequências graves.

A literatura apresenta um leque de estratégias de mitigação, que frequentemente combinam técnicas para aprimorar a confiabilidade, a robustez e o alinhamento [15, 16, 17, 18]. Contudo, o problema permanece como um desafio em aberto, para o qual ainda não há uma solução universal. Esta realidade fundamenta a hipótese do presente estudo: a de que a eficácia de uma dada estratégia de mitigação é dependente do contexto, apresentando variações significativas em função da tarefa específica, do domínio de aplicação e das métricas de avaliação empregadas.

1.1 Objetivos

O objetivo principal desta pesquisa é analisar comparativa e taxonomicamente métodos de mitigação de alucinações em modelos de linguagem de grande escala (LLMs). Com isso, busca-se aprofundar a compreensão acerca da adequação e da eficácia de cada estratégia em variados contextos de aplicação.

1.1.1 Objetivos Secundários

Para alcançar o objetivo principal, os seguintes objetivos secundários são definidos:

1. **Realizar uma Revisão Sistemática da Literatura:** Identificar, categorizar taxonomicamente e analisar criticamente os métodos de mitigação de alucinações em LLMs publicados entre 2021 e 2025, culminando na seleção de um subconjunto representativo de abordagens para avaliação prática com base em critérios de relevância e reprodutibilidade;

2. **Conduzir uma análise comparativa de desempenho:** Avaliar experimentalmente a eficácia dos métodos de mitigação selecionados, por meio de sua implementação e execução em cenários controlados com *benchmarks* padronizados, a fim de interpretar os resultados, discutir suas implicações práticas e prover um discernimento sobre as potencialidades e limitações de cada abordagem.

A análise comparativa revela que a eficácia dos métodos de mitigação de alucinações é altamente dependente do cenário de avaliação. Observou-se que certas

estratégias, como a otimização de modelo, geram ganhos substanciais de acurácia em tarefas de raciocínio e resposta a perguntas. Em contrapartida, outras abordagens, como o uso de conhecimento externo ou o aprimoramento da inferência, mostram-se superiores em métricas de correspondência lexical e precisão semântica, respectivamente. Tais resultados indicam a inexistência de uma solução universal, reforçando a necessidade de alinhar a estratégia de mitigação à aplicação desejada.

As principais contribuições desta dissertação residem no mapeamento e na categorização sistemática das abordagens recentes para mitigação de alucinações, bem como em uma análise quantitativa que evidencia o desempenho diferencial dessas estratégias em cenários diversos. Tais achados oferecem um panorama atualizado do campo, fornecendo subsídios para a escolha fundamentada de técnicas e indicando direções promissoras para pesquisas futuras.

Este trabalho organiza-se da seguinte maneira: o Capítulo 2 apresenta a fundamentação teórica, enquanto o Capítulo 3 discute os trabalhos correlatos de avaliação. O Capítulo 4 descreve a metodologia comparativa adotada, sendo os resultados detalhados no Capítulo 5. Por fim, o Capítulo 6 expõe as conclusões, as contribuições deste estudo e os direcionamentos para pesquisas futuras.

CAPÍTULO

2

Fundamentação Teórica

Este capítulo apresenta os fundamentos teóricos necessários ao entendimento do fenômeno de alucinação em modelos de linguagem de grande escala (LLMs), assim como as metodologias para sua identificação, análise e mitigação.

A fundamentação teórica deste trabalho organiza-se da seguinte maneira: a Seção 2.1 esclarece a definição do tipo de modelo focalizado neste estudo; a Seção 2.2 delineia brevemente os tipos de alucinações observadas nesses modelos; a Seção 2.2.1 revisa a literatura sobre os principais fatores causais do fenômeno; a Seção 2.3 examina as opções para identificar tais adversidades; a Seção 2.4, que constitui o enfoque principal deste trabalho, explora métodos e estratégias de mitigação e suas origens; por fim, a Seção 2.5 apresenta um resumo dos pontos principais.

2.1 Introdução aos Modelos de Linguagem de Grande Escala e sua Arquitetura

Preliminarmente à especificação do que é um LLM, torna-se imprescindível compreender a arquitetura Transformer, que serve como alicerce para a vasta maioria desses modelos. O desenvolvimento desta tecnologia iniciou-se com o conceito de *soft-alignment*, introduzido por Bahdanau et al. em "Neural Machine

Translation by Jointly Learning to Align and Translate" [19]. Este mecanismo permite que um sistema se concentre em diferentes partes de uma sequência de entrada, superando assim as limitações dos vetores contextuais de dimensão fixa, especialmente em sentenças longas. Essa inovação foi um precursor direto do moderno mecanismo de atenção, que permite às redes neurais ponderar dinamicamente a importância de diferentes palavras na entrada. Ao se concentrar nos elementos mais relevantes durante o processamento da informação, esta abordagem evita a perda de informação vista em modelos mais antigos, que tentavam comprimir sentenças inteiras em um único vetor de tamanho fixo [20, 21].

A partir desses fundamentos, Vaswani et al. [22] introduziram o *transformer*, uma arquitetura baseada inteiramente em mecanismos de atenção. Um componente-chave é a *Scaled Dot-Product Attention*, que permite ao modelo processar todas as partes da entrada em paralelo, capturando eficientemente as dependências de longo alcance. Sua formulação matemática é:

$$\text{Atenção}(Q,K,V) = \texttt{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.1}$$

A entrada é representada por três matrizes: Consultas (*Queries*, Q), Chaves (*Keys*, K) e Valores (*Values*, V). O modelo calcula os escores de atenção realizando o produto escalar entre as matrizes de Consultas e Chaves, escalona o resultado pela raiz quadrada da dimensão da chave (d_k) para estabilizar o treinamento e aplica uma função `softmax` para obter os pesos. Esses pesos são então aplicados à matriz de Valores para produzir a saída. Isso permite ao modelo ponderar a importância de diferentes palavras ao processar uma sequência.

O design fundamental do *transformer* foi adaptado em três configurações arquitetônicas principais, cada uma projetada para categorias distintas de tarefas de NLP (*Natural Language Processing*):

- Modelos somente com codificador (*Encoder-only*): São ideais para tarefas que exigem uma compreensão profunda do texto de entrada, como classificação e análise de sentimentos. Processam toda a sequência de entrada de uma só vez para criar representações contextuais ricas [23].

- Modelos somente com decodificador (*Decoder-only*): São utilizados para a geração de texto. Operam de forma sequencial, gerando um *token* de cada

vez com base nos *tokens* que o precederam. Este processo autorregressivo é o que lhes permite escrever sentenças e parágrafos coerentes [24].

- Modelos codificador-decodificador (*Encoder-decoder*): São empregados em tarefas de sequência a sequência, como tradução automática ou sumarização. O codificador processa o texto de entrada, e o decodificador gera o texto de saída com base na compreensão do codificador [25].

O pilar fundamental destas configurações é um empilhamento de camadas idênticas, cada uma construída a partir de duas subcamadas primárias. A primeira é o mecanismo de auto-atenção de múltiplas cabeças (*multi-head self-attention*), que permite ao modelo ponderar a importância das palavras ao executar o processo de atenção em paralelo, onde cada "cabeça" (*head*) captura diferentes tipos de relações contextuais. A segunda parte é a rede neural de alimentação direta (*feed-forward network*), que adiciona complexidade e profundidade ao aplicar transformações não lineares adicionais de forma independente à representação de cada token, conforme ilustrado na Figura 2.1. De forma crucial, ambas as subcamadas são envoltas por uma conexão residual e uma normalização de camada; esta estrutura é vital para o treinamento estável de arquiteturas profundas, pois facilita o fluxo de gradiente e previne a degradação do sinal [22, 25].

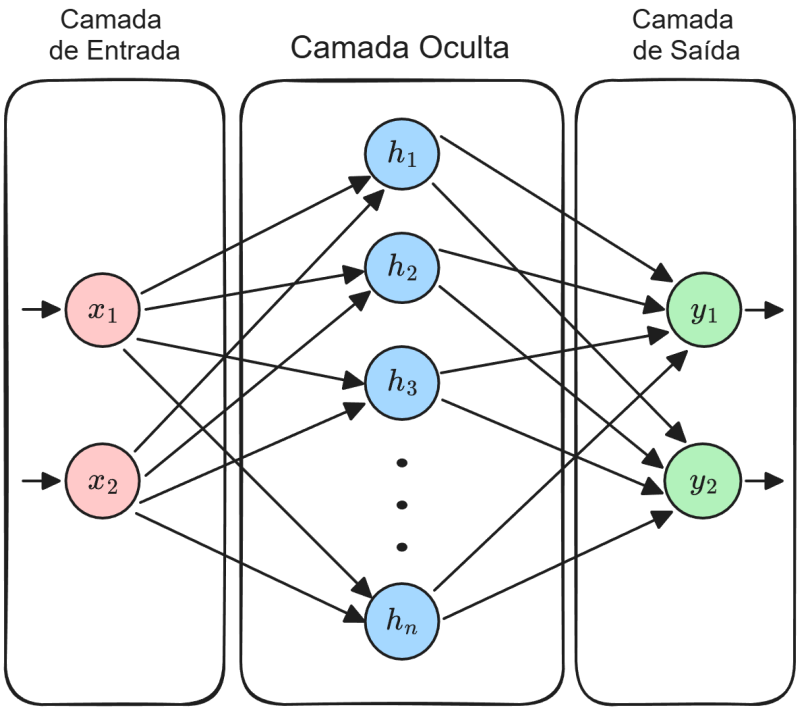

Figura 2.1: Representação visual da rede *feed-forward*.

A partir desses rudimentos, pode-se adentrar nos tipos de paradigmas de aprendizado, onde cada variação oferece uma nova forma de "ensinar". O aprendizado supervisionado, por exemplo, é caracterizado pelo uso de exemplos rotulados como dados de treinamento, permitindo que o modelo faça previsões para pontos não vistos anteriormente. Esta abordagem é comum em problemas de classificação, regressão e *ranking*, como o caso da detecção de *spam*. Já no aprendizado não supervisionado, o modelo é treinado apenas com dados não rotulados, o que pode dificultar a avaliação quantitativa do desempenho em tarefas como agrupamento e redução de dimensionalidade. O aprendizado semi-supervisionado, por sua vez, combina dados rotulados e não rotulados no treinamento [26], que visa aproveitar a distribuição dos dados não rotulados para melhorar o desempenho do modelo, especialmente em situações onde obter rotulações é custoso. Esse método é aplicável a uma variedade de problemas, incluindo classificação, regressão e *ranking*. Por fim, o aprendizado por reforço se distingue por sua interação ativa com o ambiente, onde o modelo, através de tentativa e erro, busca maximizar uma recompensa ao longo do tempo. Essa abordagem é marcada pelo dilema entre explorar novas ações para ganhar informação ou explorar o conhecimento já adquirido, sem *feedback* direto de recompensas a

longo prazo [27].

Adições metodológicas mais recentes como o aprendizado com poucos exemplos (*few-shot learning*), demonstram a capacidade destes de assimilar tarefas específicas com uma quantidade limitada de dados. A eficácia do aprendizado com poucos exemplos em modelos de linguagem avançados deriva da pré-aprendizagem abrangente sobre extensos conjuntos de dados, que equipam o modelo com um conhecimento profundo sobre linguagem, gramática e contextos variados. Essa base de conhecimento prévio, aliada à arquitetura de atenção, permite ao modelo focar em elementos-chave dos exemplos fornecidos para inferir e generalizar para novas tarefas com precisão; durante o processo, a capacidade de interpolação do modelo facilita a geração de respostas adequadas ao contexto a partir de um número limitado de dados, ao utilizar o entendimento contextual e da atenção seletiva para adaptar-se rapidamente a novos desafios sem a necessidade de extensivo re-treinamento [28].

Com os fundamentos em mente, podemos formular uma definição formal para um LLM, entendendo-os como modelos avançados de aprendizado de máquina, elaborados a partir de redes neurais profundas, com a capacidade de realizar atividades relacionadas a NLP [29]. E tendo em vista que, como dito anteriormente, grande parte se baseia na arquitetura *transformer*, pode-se refinar o conceito com base em alguns fatores que, de acordo com Thimira Amaratunga [30] são: número de parâmetros, escala de dados, capacidade computacional e adaptação de tarefas.

O número de parâmetros em um LLM serve como um indicador chave de sua capacidade de compreensão e geração de linguagem natural. Quanto maior o número de parâmetros, mais complexas são as relações e padrões que o modelo pode aprender, permitindo aumentar a performance em tarefas de NLP. Esta correlação entre tamanho e capacidade foi destacada por Brown et al [28] que mostram como aumentos significativos no número de parâmetros melhoram a fluência, a precisão e a relevância das respostas geradas por esses modelos [30].

Igualmente crítico é a escala de dados para treinamento, os modelos mais notórios como GPT, Falcon e Gemini são treinados com corpus textuais contendo centenas de bilhões a trilhões de tokens, abrangendo uma ampla gama de tópicos e estilos de linguagem [31, 32, 33]. Isso não só permite que o modelo aprenda uma diversidade de conceitos e contextos, mas também o habilita a gerar respos-

tas e conteúdo que são relevantes para uma variedade de entradas [34].

A capacidade computacional necessária para treinar e operar modelos LLM é substancial, o treinamento desses modelos exige acesso a recursos computacionais de alto desempenho, incluindo GPUs (Unidades de Processamento Gráfico) especializadas e TPUs (Unidade de Processamento de Tensor). A escalabilidade do poder computacional tem sido um desafio constante para a pesquisa e desenvolvimento de LLMs, com inovações na eficiência de treinamento e otimizações de hardware sendo cruciais para avanços na área. Esta necessidade de alto poder computacional é destacada na literatura e aponta para a crescente demanda por infraestruturas computacionais mais eficientes [35].

Por fim temos a adaptação de tarefas, ou como é melhor conhecido, *fine-tuning*, que permite que esses modelos sejam finamente ajustados para desempenhar tarefas específicas de NLP com elevada precisão. Após o treinamento inicial em grandes conjuntos de dados, os LLMs podem ser adaptados a tarefas particulares através de técnicas de *fine-tuning*, onde são expostos a um conjunto menor de dados específicos da tarefa. Essa abordagem tem mostrado sucesso em melhorar significativamente o desempenho do modelo em tarefas como compreensão de texto, geração de linguagem e tradução automática [33, 36].

Essas características predominam em uma parte considerável de Modelos de Linguagem de Grande Escala (LLMs) modernas, enquanto as abordagens anteriores utilizavam-se de modelos como redes neurais recorrentes (RNN), caracterizadas por sua estrutura projetada para o reconhecimento de sequências, como texto ou sinais de fala, através de conexões que se estendem ao longo do tempo, as quais alimentam o passo temporal seguinte em vez de uma camada concorrente [37].

2.2 Definição e Caracterização de Alucinação

Um dos maiores obstáculos para a adoção ampla, tanto em áreas de maior risco, quanto em uso pessoal de LLMs é indubitavelmente a presença de "alucinações", uma antonomásia para imprecisões do modelo [38], que possui diversas variações de como elas se manifestam. Entretanto, todas elas seguem o mesmo princípio fundamental: a falta de aderência à realidade [39], que, por consequência, traz diversas ramificações, desde operações matemáticas imprecisas [40] até

informações médicas inverídicas [41].

Baseado no estudo "*Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*" [42] é possível delinear três tipos distintos de alucinações de LLMs:

1. Alucinação Conflitante com a Entrada: Manifesta-se quando há divergência entre o conteúdo gerado pelos LLMs e o *input* (entrada) do usuário. Essa discrepância geralmente resulta de um mal-entendido das intenções do usuário ou de um desalinhamento entre as instruções da tarefa e a saída real. Um exemplo é a alteração de informações críticas em uma tarefa de resumo, na qual o LLM modifica detalhes essenciais (como nomes ou eventos) que divergem da entrada original;

2. Alucinação Conflitante com o Contexto: A segunda categoria surge em cenários onde os LLMs produzem conteúdo que é internamente inconsistente ou contraditório. Este problema é predominantemente devido às limitações do modelo em manter uma memória de longo prazo coerente, identificar e preservar com precisão o contexto ao longo de interações ou documentos extensos;

3. Alucinação Conflitante com Fatos: Ocorre quando os modelos de linguagem geram informações que contrariam o conhecimento factual estabelecido [38]. Este tipo de alucinação é particularmente perigoso, pois pode induzir os usuários com informações aparentemente plausíveis, mas incorretas.

Essa caracterização das diferentes manifestações de alucinações é crucial para desenvolver estratégias de mitigação específicas. Isso implica adaptar técnicas de detecção e correção de erros a cada categoria, em detrimento de abordagens genéricas. Por exemplo, a detecção de inconsistências entre *input* e *output* é mais adequada a alucinações conflitantes com a entrada [43], ao passo que a verificação de evidências externas pode ser mais eficaz contra alucinações conflitantes com fatos [44].

2.2.1 Causas e Mecanismos

A causa de tal fenômeno é uma miríade de fatores complexos, intricadamente interligados, que se estendem desde a qualidade e diversidade do conjunto de

dados até os intricados mecanismos internos dos modelos. Para compreender plenamente as causas subjacentes que levam à ocorrência de alucinações, é essencial desmembrar esta análise em componentes críticos: limitações na coleta de dados, limitações no treinamento de dados, vieses modelares, sobreajustes e a natureza intrinsecamente generativa do modelo, que, ao ser otimizado para produzir a resposta mais estatisticamente provável, é compelido a construir uma resposta mesmo quando não possui dados factuais para sustentá-la.

Limitações na Coleta de Dados

O treinamento de LLMs depende crucialmente da qualidade, diversidade e representatividade dos dados do *dataset*. Essas qualidades são frequentemente comprometidas em repositórios sem curadoria, como os provenientes de *Web-Crawlers*, ferramentas automatizadas que coletam informações de páginas web em grande escala. Embora eficientes na agregação, essas técnicas podem introduzir contaminação nos dados, resultando em grandes volumes de texto de baixa qualidade. Nos melhores cenários, esse texto é apenas ininteligível ou irreconhecível pelo vocabulário do modelo [45]; nos piores, pode incluir conteúdo ofensivo, explícito ou inapropriado [46], comprometendo a integridade do treinamento.

Esses mesmos *datasets* podem também possuir duplicatas de dados, como por exemplo um livro aparecer diversas vezes no mesmo banco de dados [47], contendo o mesmo conteúdo. Além dos obstáculos já mencionados, outra limitação significativa relatada na literatura são os problemas relacionados à congruência entre as informações coletadas e seus usos pretendidos, os quais podem afetar negativamente a acurácia das saídas geradas [48, 49].

Limitações durante o Treinamento

Mesmo com um bom *dataset* em mãos, isso pode não ser o suficiente, uma vez que a eficácia do processo de treinamento também desempenha um papel crítico na determinação da qualidade do modelo resultante. Durante essa fase, várias limitações podem afetar adversamente o desempenho e a confiabilidade dos LLMs devido ao processo de treinamento envolver várias etapas e técnicas, cada uma com seu próprio conjunto de desafios e possíveis fontes de problemas.

Fine-tuning é um início apropriado para esse tópico, uma vez que é uma prática de treino de modelos extremamente comum no campo de processamento de

linguagem natural e pode ser um desses pontos críticos no processo de treinamento que podem levar à geração de alucinações. Este processo consiste em que um modelo pré-treinado, adaptado a uma tarefa específica, é treinado em dados específicos da tarefa, envolvendo o ajuste de todos os parâmetros do modelo de ponta a ponta, consumindo muito menos recursos do que o pré-treino inicial [50]. Apesar da eficiência, métodos como *fine-tuning* supervisionado podem aumentar a probabilidade de alucinações, isto é atribuído a ligações causais inadequadas entre o conjunto de dados de treino e as respostas do modelo, o que resulta na geração de informações factualmente incorretas ou falsas [51].

Esse procedimento também pode levar a um subfenômeno da alucinação: o esquecimento catastrófico, no qual um modelo "esquece" informações previamente conhecidas. Tal falha compromete um aspecto crucial para "agentes inteligentes": a aprendizagem contínua, definida como a habilidade de aprender sem descartar dados anteriores [52]. A correlação entre *fine-tuning* e esquecimento é diretamente abordada no artigo "*Fine-tuning Deep Pretrained Language Models with Less Forgetting*" [53]. Nele, uma análise centrada na variação da distância euclidiana entre os pesos do modelo (antes e após o ajuste fino) evidencia uma fase inicial de recuperação de conhecimento, seguida por um aumento gradual dessa distância, o que reflete o esquecimento conforme o modelo se adapta à nova tarefa.

Além das questões de *fine-tuning* e esquecimento catastrófico, as estratégias de decodificação em sistemas de geração de linguagem apresentam outra camada de complexidade e limitação. O uso de KGs (grafos de conhecimento) em sistemas de geração de linguagem evidencia como a escolha entre a decodificação gulosa e a amostragem top-k impacta diretamente na ocorrência de alucinações [54]; a busca pela palavra mais provável na decodificação gulosa minimiza erros, enquanto a variabilidade da amostragem top-k pode levar a fabricações não fundamentadas no KG.

Outrossim, observações em pesquisas recentes, como no estudo *TruthfulQA* [55], indicam uma tendência intrigante: nem sempre o aumento do tamanho dos modelos de linguagem resulta em maior veracidade das respostas. Essa tendência, que contrasta com a melhoria geral de desempenho em outras tarefas de NLP associada a modelos maiores, sugere que em alguns casos, modelos de maior escala podem gerar respostas que imitam inverdades ou exploram de forma

adversária fraquezas específicas. Tal fenômeno aponta para uma complexidade intrincada no treinamento de LLMs, onde a correlação entre o tamanho do modelo e a fidelidade da informação não é linear, e escalonar o tamanho do modelo não é uma panaceia para todos os desafios associados à geração de linguagem natural [56].

Vieses Modelares

É comumente pressuposto que os vieses são exclusivamente decorrentes dos conjuntos de dados nos quais os modelos são treinados. No entanto, as inclinações incorporadas aos próprios algoritmos e as metodologias empregadas durante o desenvolvimento também desempenham um papel significativo. Essa realidade é particularmente evidente ao se considerar que estratégias de regularização, essenciais para controlar a complexidade do modelo e prevenir o excesso de ajuste, podem paradoxalmente levar a uma simplificação excessiva [57].

Tal simplificação manifesta-se quando o modelo carece da capacidade de decifrar adequadamente as nuances dos dados, tal cenário, denominado subajustamento, revela-se quando, por exemplo, uma abordagem linear para prever PIB (Produto Interno Bruto) de um país falha por não capturar a multifacetada realidade subjacente, e isso resulta em predições falhas até para as instâncias de treino [58]. Isso sublinha a importância de um equilíbrio na modelagem, onde nem a supercomplexidade que obscurece a generalização, nem a simplicidade excessiva que omite detalhes cruciais, são desejáveis.

Esse fenômeno é particularmente pronunciado quando se trata de resultados de grupos minoritários em categorias sensíveis, como o gênero ou a raça, em que a simplificação dos modelos através da regularização leva a uma subestimação significativa de resultados pouco frequentes, mas cruciais [59]. Essa sub-representação não é apenas uma questão de supervisão algorítmica, mas um reflexo da complexa interação entre o desejo de simplicidade do modelo e o imperativo de ajustamento e inclusão nas previsões do modelo.

Sobreajuste

Como já destacado, o fenômeno do sobreajuste, onde o modelo exibe desempenho excepcional nos dados de treino, mas apresenta baixo desempenho ao ser confrontado com novos dados, revela uma faceta crítica na modelagem de apren-

dizado de máquina. Esse excesso de domínio em detalhes específicos do conjunto de treinamento pode resultar em previsões pouco confiáveis quando aplicadas a situações reais, um reflexo do desafio em equilibrar a complexidade do modelo com a generalização. Por exemplo, um modelo de satisfação de vida baseado em um polinômio de alto grau pode se ajustar perfeitamente aos dados de treino, mas essa precisão pode ser ilusória e não se refletir em um conjunto de teste, como ilustrado na Figura 2.2, exigindo, portanto, maior cautela. A capacidade de modelos avançados de discernir padrões intrincados é notável, contudo, se esses "padrões" emergem do ruído ou de características irrelevantes, como nomes de países, eles podem falhar em prever corretamente fora da amostra original [58].

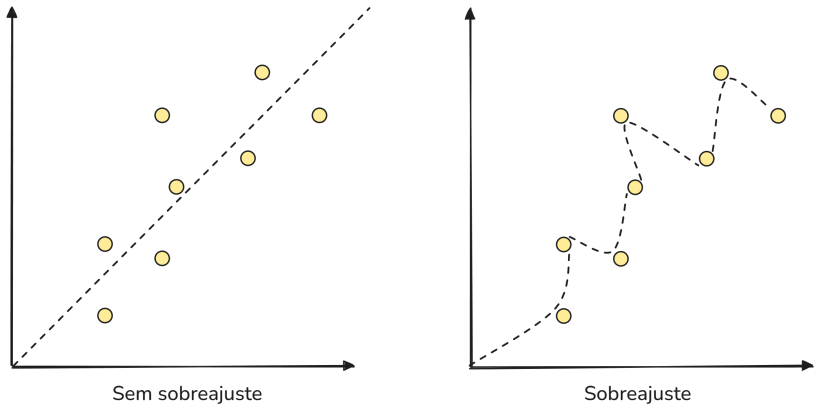

Figura 2.2: Representação de sobreajuste em um conjunto de dados.

Limites do Autoconhecimento

Esse arquétipo de limitação em LLMs origina-se, sobretudo, das metodologias de *design* e treinamento adotadas, que não os preparam para identificar adequadamente quando lhes faltam informações essenciais ou quando se baseiam em premissas equivocadas. Fundamentalmente, o problema reside na dificuldade dos modelos em distinguir entre o conhecimento que detêm e as lacunas em seu entendimento. Embora métodos e conjuntos de dados de treinamento convencionais foquem em garantir precisão em cenários com respostas definidas, falham em equipar os modelos para reconhecer e admitir suas próprias limitações perante perguntas sem respostas claras ou informações falsas. Esse desalinhamento leva a situações onde os modelos podem produzir respostas enganosas ou fabricadas, ao confundir áreas de sua programação carentes de precisão ou

fundamentação com aquelas onde seu conhecimento é sólido e confiável [60].

Além disso, os mecanismos de avaliação dos modelos, como os baseados em conjuntos de dados ou *benchmarks* existentes, geralmente não medem adequadamente a capacidade de um modelo de expressar incerteza ou se recusam a responder quando apropriado. Essa limitação não apenas dificulta o desenvolvimento de modelos capazes de autoconsciência genuína, mas também contribui para o excesso de confiança observado nos resultados do LLM, que pode se manifestar como alucinações [61].

2.3 Detecção e Avaliação de Alucinações

A detecção e a avaliação de alucinações requerem métodos e ferramentas especializadas. Os *benchmarks*, em particular, são essenciais para medir objetivamente o desempenho dos modelos e para orientar o desenvolvimento de estratégias eficazes de detecção e mitigação. Por meio dessas análises, estabelecem-se parâmetros essenciais para a avaliação da eficácia das técnicas de mitigação, fornecendo um referencial que permite não apenas identificar as limitações atuais dos modelos, mas também traçar caminhos para melhorias futuras [55, 62].

Para compreender a evolução das medidas de avaliação em NLP, é elucidativo considerar a métrica BLEU (*Bilingual Evaluation Understudy*) [63], um método pioneiro para a avaliação automática de tradução automática que antecede os modelos baseados em redes neurais. A inovação do BLEU foi sua capacidade de realizar avaliações rápidas, de baixo custo e independentes do idioma, que se correlacionavam fortemente com os julgamentos humanos sobre a qualidade da tradução. Ela opera comparando uma tradução gerada por máquina com uma ou mais traduções de referência humanas, calculando uma pontuação com base na precisão das correspondências de n-gramas[1] de diversos comprimentos. Para evitar a premiação de traduções excessivamente curtas, a medida incorpora uma "penalidade por brevidade", que ajusta a pontuação para baixo caso o texto candidato seja mais curto que as referências. A introdução desse *benchmark* marcou um ponto de virada ao estabelecer um padrão quantitativo para o campo, permitindo ciclos mais rápidos de pesquisa e desenvolvimento em tradução automática e abrindo caminho para as sofisticadas metodologias de detecção de alucinações

[1] n-gramas são sequências contíguas de n itens de uma dada amostra de texto ou fala.

utilizadas posteriormente.

As medidas também podem ser mais diretas e exordiais como a medição de precisão, utilizada para avaliar a capacidade do modelo de identificar informações relevantes. O cálculo para esta medida é a razão entre os positivos verdadeiros (instâncias nas quais o modelo prevê corretamente a classe positiva) e o número total de identificações positivas que o LLM faz [64]. Outra medida que utiliza precisão na sua equação e que é amplamente adotada em avaliações é a medida F1 [65] para uma análise mais balanceada, que consiste de uma média harmônica entre *recall* e precisão para evitar que um modelo pareça eficiente ao performar bem apenas em uma das duas [66].

Igualmente relevante é a acurácia, que fornece uma medida simples e intuitiva do desempenho geral de um modelo, que calcula a proporção de todas as previsões (tanto positivas quanto negativas) que o modelo acerta [67]. Essa métrica pode ser usada como um ponto de partida para avaliar modelos de aprendizado de máquina, devido à sua compreensão direta.

À medida que o desenvolvimento na área de processamento de linguagem natural avançava significativamente, outros métodos de avaliação ganharam destaque, transcendendo as limitações anteriores com inovações metodológicas e tecnológicas. Em particular, a atenção se voltou para abordagens mais sofisticadas que conseguem capturar nuances complexas da linguagem, refletindo um progresso notável na capacidade de avaliar e entender modelos de forma mais profunda e abrangente. Este avanço se manifesta na elaboração de *benchmarks* e *datasets* diagnósticos, que não apenas desafiam os modelos com uma gama mais ampla de tarefas de NLP, incluindo perguntas e respostas, inferência de linguagem natural e análise de sentimentos, para avaliar a capacidade dos modelos de generalizar conhecimento através de tarefas, mas também submetem a avaliações focadas em fenômenos linguísticos específicos. Tais análises são essenciais para uma compreensão aprofundada da disciplina focada na interação entre computadores e a linguagem humana [68].

Perguntas e respostas, mais especificamente, evoluíram para englobar uma gama variada de técnicas de avaliação, refletindo a necessidade crescente de modelos que compreendam profundamente os textos e sejam capazes de inferir respostas de maneira contextualizada. Isso se manifesta na criação de conjuntos de dados como o SQuAD (*Stanford Question Answering Dataset*), que desafia mode-

los de compreensão de texto com indagações baseadas em passagens de artigos da Wikipedia, exigindo não apenas a identificação de informações explícitas, mas também a inferência e a compreensão de nuances implícitas no texto. A diversidade de perguntas e a exigência de respostas específicas, derivadas diretamente dos textos, ilustram um passo significativo rumo a avaliações mais rigorosas e abrangentes de modelos de linguagem, que priorizam uma compreensão de leitura mais profunda e a capacidade de raciocínio lógico sobre o reconhecimento superficial de padrões [69].

Na progressão da linha temporal dos métodos de avaliação e *benchmarking* em NLP, a introdução da arquitetura do Transformador [22] marcou um uso estratégico da medida BLEU para testar sua eficiência em tarefas de tradução automática, o que evidencia a importância de *benchmarks* robustos na validação de novas arquiteturas. Este marco preparou o terreno para a adoção de avaliações ainda mais complexas, com o BERT (*Bidirectional Encoder Representations from Transformers*) [70], o qual incorpora GLUE [68] e SQuAD [69] para avaliar sua proficiência em compreensão de texto e inferência.

Essa progressão não apenas demonstrou a capacidade do BERT de navegar por nuances linguísticas e contextuais com precisão sem precedentes, mas também reiterou a evolução contínua dos *benchmarks*, que passaram a enfatizar não somente o reconhecimento de padrões, mas a verdadeira compreensão e raciocínio sobre a linguagem. Assim, a transição do uso de BLEU para a aplicação de GLUE (*General Language Understanding Evaluation*) que desafia modelos em tarefas de inferência e análise textual e SQuAD (*Stanford Question Answering Dataset*) que exige localização e extração de respostas a partir de um contexto, sublinha uma jornada em direção a métodos de avaliação que refletem a complexidade e a profundidade das tarefas de NLP, estabelecendo um novo padrão para a medição do avanço tecnológico no campo.

Todavia, esses métodos não são isentos de falhas, conforme revelam investigações sobre artefatos de anotação presentes em conjuntos de dados de Inferência de Linguagem Natural (NLI). Sublinha-se que simples modelos de classificação de texto conseguem prever a classe de uma hipótese sem acesso à premissa, o que demonstra que a precisão destes modelos pode ser artificialmente inflada por esses artefatos [71]. Estes são padrões linguísticos inadvertidamente inseridos durante o processo de anotação, que permitem a classificação correta baseada

em heurísticas superficiais, sem necessidade de compreensão profunda do texto. Esse fenômeno desafia a confiabilidade de *benchmarks*, e leva a sugerir uma superestimação da capacidade dos modelos. A descoberta destes artefatos ressalta a necessidade de desenvolver métodos de avaliação mais robustos e representativos que possam efetivamente medir a verdadeira capacidade de compreensão e inferência dos modelos, de modo a evitar a dependência de atalhos heurísticos que não refletem habilidades de raciocínio real.

Ao reconhecer as limitações dos métodos baseados em artefatos de anotação, uma abordagem significativa na avaliação de modelos de linguagem é o uso do BERTScore, que utiliza o entendimento contextual profundo proporcionado de BERT diferentemente de métodos baseados em correspondência exata de palavras ou frases, a técnica computa a similaridade semântica entre os *tokens* dos textos de referência e os gerados com representações vetoriais densas [72]. Isso permite uma avaliação mais minuciosa e significativa da qualidade textual, levando em consideração nuances linguísticas que métodos anteriores, focados em sobreposições de n-gramas ou distância de edição, poderiam negligenciar; n-gramas referindo-se a conjuntos contíguos de n itens (neste contexto, palavras) extraídos do texto, enquanto a distância de edição mede o número mínimo de operações necessárias para transformar uma *string* de texto em outra. Porém, apesar dos avanços feitos pelo método, ele possui limitações, principalmente na sua sensibilidade a erros menores, particularmente em contextos onde o candidato é lexical ou estilisticamente similar à referência [73]. Isso implica que, enquanto o BERTScore é eficaz em detectar divergências em elementos de conteúdo importantes, ele pode não ser tão preciso em identificar discrepâncias sutis, que não alteram significativamente a similaridade semântica ou o estilo das expressões.

Como pode ser observado, a evolução de processamento de linguagem destaca a necessidade de diversidade de metodologias cada vez mais sofisticadas; a adoção de estratégias adversárias emergiu como uma resposta aos desafios de avaliação precisa de IA. Essa abordagem, fundamentada na aplicação de filtros que intencionalmente procuram expor fraquezas, encoraja a criação de algoritmos capazes de decifrar e reagir adequadamente a contextos intricados e sutilezas linguísticas. O processo envolve a geração de instâncias de teste concebidas para questionar profundamente os modelos, forçando-os a navegar por situações além

das habituais. Ao submeter sistemas a esses testes rigorosos, impulsiona-se o aprimoramento de mecanismos de compreensão e adaptação, fundamentais para a superação de incongruências ou imprecisões [74]. Tal metodologia enfatiza a importância de desenvolver *benchmarks* que realmente testem a capacidade dos modelos de processar e interpretar a linguagem com um nível de discernimento comparável ao humano, movendo o campo para além da mera identificação de padrões para uma genuína análise e resposta contextual.

A emergência de técnicas avançadas para a detecção ressalta a evolução do campo para além das estratégias convencionais, e, com isso, abraça abordagens que incorporam verificações de veracidade e o emprego de bases de conhecimento externas [54]. Essas inovações refletem um movimento em direção a uma compreensão mais profunda da precisão e relevância do conteúdo gerado por modelos de linguagem, na importância de avaliar as respostas numa união de avaliações de coerência linguística com aderência a fatos. Este avanço metodológico enfatiza o uso de informações verificáveis para assegurar a precisão das gerações de texto [55], estabelecendo um marco na busca por sistemas de IA capazes de informar de maneira confiável e precisa. Com a integração de verificações de factualidade e análises baseadas em conhecimento externo, marca-se um progresso significativo na capacidade dos modelos de linguagem de lidar com complexidades e sutilezas, promovendo uma geração de linguagem que não apenas imita, mas também reflete e responde de forma informada ao mundo real.

A abordagem de testar as capacidades de conhecimento factual de um modelo, entretanto, apresenta-se como um desafio inerente à sua própria natureza: a dificuldade em determinar se o modelo realmente "sabe" algo ou se está apenas repetindo informações memorizadas dos dados de treinamento. Essa preocupação se torna especialmente relevante com a proliferação de *benchmarks* e conjuntos de dados públicos, frequentemente utilizados para avaliar LLMs. A ubiquidade desses recursos aumenta a probabilidade de que modelos, especialmente aqueles treinados em conjuntos de dados massivos e diversos, tenham encontrado partes significativas dessas avaliações durante o treinamento [75]; tal exposição prévia (denominada como contaminação) pode levar a uma superestimação do desempenho do modelo em testes subsequentes, pois o modelo pode simplesmente recuperar respostas memorizadas em vez de gerar respostas com base em uma compreensão real do conhecimento [76].

Contudo, a verificação baseada em referências não é a única forma de realizar *benchmarks* e avaliações. Detecções do tipo *reference-free* (livre de referências), por exemplo, utilizam análises baseadas em *tokens* individuais para identificar sutis incoerências sem depender de textos de referência pré-existentes. Esse método avança significativamente além da verificação de consistência em níveis mais amplos (como sentenças ou documentos), habilitando a detecção de erros factuais e lógicos em uma escala muito mais granular. Um exemplo dessa abordagem é a criação de conjuntos de dados anotados especificamente para tal finalidade, como o HADES (*Hallucination Detection Dataset*) [64]. O HADES simula alucinações em sistemas de geração de linguagem natural por meio de perturbações contextuais e emprega métodos de anotação iterativa para equilibrar as distribuições de rótulos em cenários de acentuado desequilíbrio de classes.

Outros métodos de análise de previsão de alucinação em nível de *token* envolvem a distinção entre alucinações extrínsecas e intrínsecas. Alucinações extrínsecas referem-se a conteúdo adicional gerado sem base clara na entrada, enquanto alucinações intrínsecas contêm informações incorretas geradas a partir do conteúdo presente na entrada. Este enfoque permite um tratamento com mais nuance dos erros de geração, pois distingue entre tipos de desvios da confiabilidade ao texto-fonte, para avaliar a presença e o tipo de alucinação, a análise se baseia na identificação de spans de texto na saída gerada que não encontram suporte no texto de entrada [77].

É possível também usar previsões em nível de *token* com mecanismos de atenção e penalidades ajustáveis para mitigar a propagação de alucinações. Este método visa corrigir desvios introduzidos por *tokens* historicamente problemáticos e aplica penalidades aos subsequentes que se baseiam em pesos de atenção [78]. Ajustes na probabilidade dos *tokens*, condicionados à sua tipologia e frequência, alinham mais estreitamente a distribuição de probabilidade com avaliações humanas e simultaneamente abordam questões de excesso e falta de confiança.

Vale ressaltar que avaliações não estão necessariamente “presas” com detecções em nível de *token*, estratégias com uma abordagem mais holística, considerando a coerência e factualidade do texto de maneira mais integrada e abrangente, em nível que pode-se dizer sentencial, podem ter êxito. Abordagens recentes [79] sugerem a unificação de uma diversidade de fontes de dados e tarefas de compreensão de linguagem para treinar uma função de alinhamento unificada.

Essa função é desenhada para avaliar se todas as informações em um segmento de texto estão presentes e alinhadas com outro, de modo a proporcionar uma medida para consistência factual que pode lidar com textos longos e acomodar os diferentes papéis de contexto e reivindicação. A agregação de pontuações de alinhamento entre fragmentos de contexto e sentenças de reivindicação resulta em uma pontuação final de consistência dos fatos, o que destaca a importância de considerar a interação complexa entre diferentes partes do texto gerado.

Outras adoções de estratégias abordam os desafios convencionais por meio da utilização de estratégias de amostragem-então-filtragem, onde amostras alucinadas são geradas de forma automatizada por outra LLM, a qual emprega métodos para amostragem diversas e filtragem. Essa depuração visa selecionar as amostras mais plausíveis e desafiadoras entre as geradas, empregando critérios como plausibilidade e dificuldade de identificação para escolher entre duas candidatas alucinadas. Este processo é complementado por anotações humanas, onde rotuladores avaliam as respostas de uma IA para identificar conteúdo alucinado. Para a detecção e a recuperação de conhecimento relevante, a implementação do raciocínio em cadeia demonstrou ser particularmente eficaz, pois facilita a introdução de etapas intermediárias de raciocínio, o que permite que os modelos processem e validem informações de maneira mais sequencial e lógica, aprimorando não apenas a capacidade de reconhecimento de alucinações, ao encorajar uma análise mais profunda do contexto e dos fatos apresentados, mas também contribui para uma maior compreensão e interpretação dos dados por parte de NLPs [80].

Ao concluir esta seção sobre detecção e avaliação de alucinações, é crucial reconhecer as limitações dos novos métodos empregados, a geração automática de amostras pode ser limitada pela capacidade da ferramenta utilizada em seguir instruções complexas, o que exige um controle de qualidade rigoroso e filtros de alta precisão para garantir a fidelidade dos dados gerados. A similaridade visual entre amostras fabricadas e dados autênticos acarreta riscos de uso indevido, e requer monitoramento contínuo e regulamentações apropriadas para mitigar potenciais abusos [80]. No âmbito da interpretabilidade, a dificuldade em decifrar a lógica subjacente às previsões de determinados modelos destaca a necessidade de métricas explicáveis que possam identificar e justificar erros de consistência factual de maneira compreensível. Por fim, a cobertura limitada a uma única

língua aponta para a importância de expandir avaliações de consistência factual para ambientes multilíngues e idiomas com menos recursos, a fim de garantir uma aplicabilidade mais ampla e inclusiva [79].

2.4 Estratégias de Mitigação

Uma classe principal de estratégias de mitigação aborda as alucinações ao fundamentar as saídas do modelo em conhecimento externo e verificável. Com raízes nos princípios fundamentais da Recuperação da Informação (IR) [81], esta abordagem enfrenta o problema de que o conhecimento interno de um LLM é estático e pode estar desatualizado ou incorreto. Ao aprimorar o modelo com informações em tempo real de fontes externas confiáveis, como a web, bancos de dados com curadoria ou grafos de conhecimento [82], esses métodos garantem que as respostas não sejam apenas plausíveis, mas factualmente exatas. Este princípio de fundamentação no conhecimento pode ser implementado por meio de vários métodos distintos.

Uma aplicação direta envolve o enriquecimento do conjunto de dados com justificativas extraídas de artigos ou de sites especializados em verificação de fatos. Esta abordagem não apenas captura a veracidade da alegação, mas também incorpora o raciocínio subjacente utilizado por humanos no processo de verificação, levando a uma melhoria significativa na tarefa de classificação de veracidade [83, 84].

Uma implementação mais potente e automatizada é a RAG (*Retrieval augmented Generation*). Este mecanismo aprimora um modelo de linguagem de base ao fornecer-lhe dinamicamente informações externas e pertinentes durante a inferência, garantindo que as respostas sejam fundamentadas em dados atuais. Conforme ilustrado na Figura 2.3, o processo se inicia com a conversão da consulta do usuário em um *embedding*, um vetor de alta dimensão. Esse *embedding* facilita uma busca semântica em uma base de dados vetorial previamente populada, que contém as incorporações de um corpus de documentos externos. O sistema recupera os trechos de documentos com a maior similaridade semântica com a consulta. Essa informação recuperada é então combinada com a consulta original para formar um *prompt* aumentado. Ao alimentar o modelo de linguagem com este novo e abrangente *prompt*, rico em dados externos pertinentes, a

saída final é diretamente informada pelos fatos recuperados, em vez de depender unicamente do conhecimento interno do modelo [85].

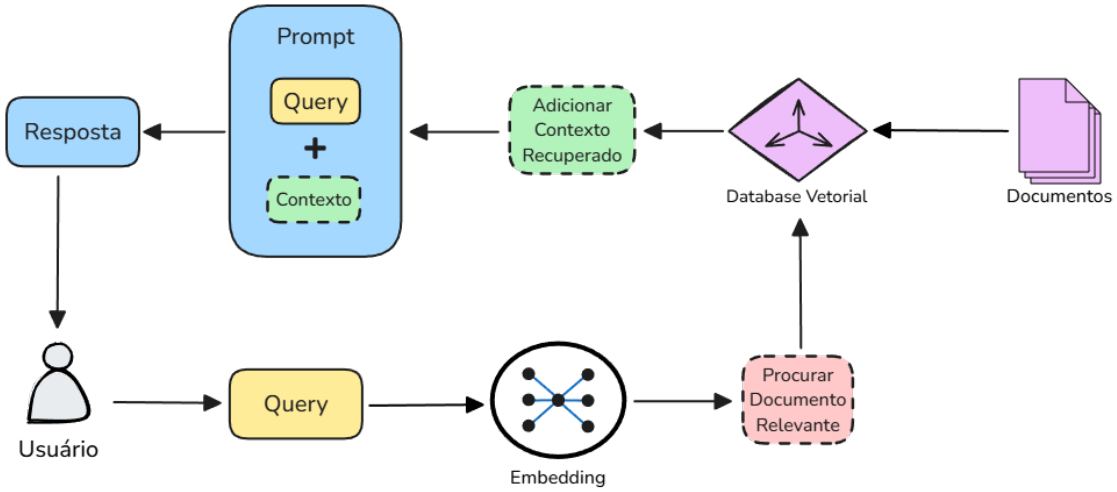

Figura 2.3: Diagrama simplificado de *RAG*, desenvolvido com base no diagrama apresentado por Gao et al. [1].

Contudo, a eficácia das técnicas de fundamentação é limitada pela qualidade da informação externa que utilizam. Em métodos como a RAG, a etapa de recuperação é uma vulnerabilidade crítica: a recuperação de documentos irrelevantes pode levar o modelo a sintetizar um contexto falho, resultando nas chamadas "alucinações externas" [86, 87]. Essa dependência estende-se às próprias fontes de dado, que podem conter vieses ou detalhes supérfluos que distorcem o desempenho do modelo. Por exemplo, sistemas treinados em conjuntos de dados como o FEVER (*Fact Extraction and VERification*) frequentemente apresentam uma queda significativa na acurácia ao serem avaliados em conjuntos de teste mais robustos, revelando uma dependência excessiva de artefatos exclusivos do conjunto de dados em vez de uma verdadeira generalização. Em última análise, essas estratégias são limitadas por sua dependência da qualidade dos corpora externos e pelo risco de incorporar seus vieses inerentes [88, 89].

Diante dessa adversidade distinta, uma potencial solução envolve métodos para avaliar a consistência da informação gerada em relação a fatos e contextos verificáveis. Tal abordagem utiliza estudos comparativos multilíngues para detectar anomalias ou incoerências na saída, facilitando a descoberta precoce de falhas no sistema de geração. Em paralelo, táticas direcionadas de recuperação da informação são implementadas para aprimorar a criação de conteúdo. Es-

sas táticas enriquecem a saída com dados externos confiáveis, o que minimiza a ocorrência de imprecisões e, em última análise, eleva a qualidade e a confiabilidade das respostas fornecidas. Não obstante seus benefícios inerentes, esta estratégia é limitada por sua dependência de correções post-hoc, restringindo a identificação de erros à etapa posterior à formação da sentença. Isso é agravado pela necessidade de múltiplas chamadas de API, o que afeta adversamente o desempenho geral do método [86].

Um processo distinto para incorporar conhecimento externo é o CoVe (*Chain of Verification*), que se concentra na autoverificação para garantir a fidelidade da resposta. Esta técnica opera em quatro fases distintas: primeiro, uma resposta base é elaborada. Segundo, o sistema cria um conjunto de perguntas pontuais para verificar os fatos de suas declarações iniciais. Terceiro, essas indagações são resolvidas por meio da consulta a informações externas confiáveis. Na última etapa, uma resposta final aprimorada é construída utilizando os dados validados. Esta abordagem, centrada na verificação, garante uma saída mais exata ao reduzir substancialmente a propagação de alucinações [90].

Além da fundamentação dos modelos em fatos externos, uma segunda linha defensiva fundamental envolve o aprimoramento da arquitetura interna e do regime de treinamento do modelo para promover uma melhor generalização e raciocínio. Para lidar com o desafio da generalização deficiente, conforme abordado anteriormente na Seção 2.2.1 e Seção 2.2.1, uma das estratégias mais populares é a regularização. Essa técnica atua ajustando o equilíbrio entre viés e variância, otimizando a capacidade do modelo de prever corretamente dados não vistos sem sacrificar a acurácia em dados conhecidos. Ao impor restrições ou adicionar penalidades à função objetivo, a regularização promove um modelo menos complexo, mas suficientemente flexível para capturar as nuances dos dados. Consequentemente, a regularização emerge não apenas como uma ferramenta para o ajuste de modelos, mas também como uma pedra angular na construção de sistemas de aprendizado de máquina que sejam simultaneamente potentes e generalizáveis [91].

Arquiteturas pioneiras de aprendizado profundo, como as CNNs (*Convolutional Neural Networks*), estabeleceram um princípio fundamental para a mitigação: o aprimoramento da generalização do modelo por meio da regularização. Uma técnica fundamental utilizada nesses modelos é o *Dropout*, que desativa aleatoria-

mente um subconjunto de neurônios durante o treinamento [92]. Esse processo previne a coadaptação de características e força a rede a aprender representações mais robustas e distribuídas. Ao aprimorar a generalização, a regularização reduz a tendência de um modelo ao sobreajuste aos dados de treinamento, mitigando, assim, a geração de saídas sem fundamento, ou alucinações. Este princípio de aprimoramento da generalização permanece como uma estratégia central para aprimorar a factualidade do modelo.

Com o avanço fornecido por CNNs, o campo de estudo se expandiu, um dos marcos iniciais foi o desenvolvimento do ELMo (*Embeddings from Language Models*) que não só representou um avanço no campo de redes neurais, mas também em métodos que aprimoram a compreensão de linguagem através de representações de palavras profundamente contextualizadas, derivadas de modelos de linguagem bidirecionais, os quais são modelos que utilizam tanto a previsão de palavras futuras a partir de um histórico de palavras anteriores, quanto a previsão de palavras anteriores a partir de um futuro conhecido, e que, dessa forma, capturam efetivamente o contexto de ambas as direções. Assim, as camadas bidirecionais de memória de curto prazo do ELMo geram representações de termos que variam dinamicamente em função do contexto sentencial, abrindo caminho para avanços significativos em tarefas que dependem fortemente da compreensão contextual, como a desambiguação de sentidos de palavras, reconhecimento de entidades nomeadas e inferência textual [93].

Alguns meses após a introdução de ELMo, surgiu o BERT, no qual se apresenta uma mudança paradigmática com o uso da arquitetura de Transformadores; este modelo aprimorou a capacidade de análise linguística ao empregar mecanismos de atenção que possibilitam a avaliação simultânea de todas as partes do texto, superando as limitações dos métodos sequenciais anteriores. Através de objetivos de pré-treinamento inovadores, como o Modelo de Linguagem de Mascaramento (MLM) e a Predição da Próxima Sentença, o BERT alcançou uma compreensão contextual profunda, o que o capacita a prever palavras ocultas dentro de uma frase e entender a relação entre duas frases consecutivas. Esses avanços permitiram uma interpretação mais acurada do significado e contexto das palavras, contribuindo significativamente para a redução de erros e “alucinações” em tarefas complexas de processamento de NLP, estabelecendo novos padrões de desempenho em classificação de textos, análise de sentimentos e questões de

compreensão de leitura [70].

À medida que modelos como o BERT estabelecem novos padrões em compreensão linguística, surgem metodologias adicionais para reduzir alucinações mais específicas, como a ocorrência de alucinações em modelos de sequência condicional, fenômenos onde o modelo gera informações não presentes nos dados de entrada. A necessidade de abordar essas limitações nos conduz ao desenvolvimento de metodologias mais sofisticadas para a identificação e correção dessas discrepâncias. Neste contexto, a criação de dados sintéticos apresenta-se como uma estratégia promissora. Ao utilizar modelos avançados como o BART (*Bidirectional and Auto-Regressive Transformer*), que combina técnicas de *auto-encoder* e auto-regressão para prever texto original a partir de versões corrompidas, torna-se possível gerar exemplos sintéticos enriquecidos com novos *tokens* alucinados. Essa abordagem não apenas facilita a geração de dados de treinamento personalizados, mas também oferece uma maneira eficaz de rotular alucinações de maneira detalhada, estabelecendo uma base sólida para treinar modelos capazes de distinguir entre conteúdo autêntico e inserções fictícias [94].

Ao aprofundar o uso do BART para enfrentar alucinações em modelos de linguagem, o método ConSeq (*Contrastive Sequence to Sequence learning*) destaca-se, por aproveitar especificamente a implementação Fairseq (*Facebook AI Research Sequence-to-Sequence Toolkit*), uma biblioteca de código aberto para processamento de linguagem natural baseada em PyTorch [95], do BART-large para estabelecer novos padrões em sumarização. Essa estratégia emprega a versão pré-treinada do modelo como base, e otimiza-o ainda mais com aprendizado contrastivo para aprimorar a consistência factual dos resumos gerados. A técnica ConSeq de aprendizado contrastivo aprimora a precisão factual de modelos de sumarização abstrativa ao utilizar uma abordagem seq2seq (*Sequence to Sequence*)[2], que são modelos treinados para converter sequências de entrada em sequências de saída (por exemplo, de texto para resumo). Inicialmente, um modelo seq2seq é treinado com um conjunto de dados rotulado empregando máxima verossimilhança, um método estatístico para estimar os parâmetros de um modelo, maximizando a probabilidade de observar os dados e os parâmetros. Após o treinamento inicial, sequências alvo verdadeiras e sequências amostradas são coletadas para formar um conjunto de candidatos, divididos em subconjuntos de

[2]Arquitetura de rede neural que mapeia sequências de entrada para sequências de saída, a qual utiliza um codificador para processar a entrada e um decodificador para gerar o resultado.

alta recompensa (sumários factuais) e baixa recompensa (sumários não factuais). A técnica minimiza uma função de perda contrastiva que promove a geração de sumários consistentes com os fatos (alta recompensa) e desencoraja a produção de sumários factuais incorretos (baixa recompensa). Esse processo direciona o modelo a preferir a geração de conteúdo factualmente correto, o que melhora a precisão e a confiabilidade dos sumários produzidos [96].

Embora as estratégias acima mencionadas criem modelos mais capazes, garantir que suas saídas se alinhem com valores e instruções humanas nuançadas requer paradigmas de ajuste especializados. O mais proeminente destes é o RLHF (*Reinforcement Learning from Human Feedback*), uma técnica cuja implementação completa em arquiteturas de linguagem, com foco na acurácia, ganhou proeminência após a modernização dos transformers [97]. O método se inicia com um ajuste fino supervisionado em um conjunto de dados com demonstrações humanas do comportamento desejado. Posteriormente, avaliadores humanos classificam um conjunto de saídas geradas por ordem de preferência. Essas classificações são usadas para treinar um modelo de recompensa, que aprende a pontuar as saídas com base nesse julgamento humano. Finalmente, o sistema de linguagem é otimizado em relação a esse modelo de recompensa utilizando o algoritmo de Otimização de Política Proximal (PPO), que aprimora a estabilidade e a eficiência do treinamento por meio de atualizações progressivas da política.

O resultado principal do processo de RLHF é o alinhamento do desempenho do sistema com as preferências humanas. Demonstrou-se que este alinhamento mitiga as alucinações e gera respostas mais exatas e menos "tóxicas". Ao empregar o *feedback* humano como um sinal de recompensa direto, a técnica melhora a aderência às instruções, frequentemente sem comprometer a eficácia em benchmarks gerais de NLP [98].

Formalmente, o processo de RLHF envolve duas etapas principais. Primeiramente, um modelo de recompensa (r_θ) é treinado com os dados de preferência humana. Seu objetivo é aprender um escore de recompensa escalar para qualquer par de instrução (*prompt*) e resposta. O processo de treinamento otimiza esse modelo ao maximizar a margem entre os escores atribuídos às respostas preferidas (y_w) e às respostas "não preferidas" (y_l) para uma dada instrução (x) [99, 98].

Em segundo lugar, esse modelo de recompensa treinado é utilizado para reali-

zar o ajuste fino da política do modelo de linguagem (π_ϕ^{RL}) por meio do aprendizado por reforço, tipicamente com um algoritmo como o PPO. A função objetivo nesta etapa é projetada para maximizar as recompensas do modelo de recompensa, ao mesmo tempo que inclui um termo de regularização. Este termo, uma penalidade baseada na divergência de Kullback-Leibler (KL) entre a política atual e a política original do ajuste fino supervisionado, é crucial; ele impede que o modelo se desvie drasticamente de sua base de conhecimento inicial na busca por recompensa, mitigando, assim, problemas como o esquecimento catastrófico e garantindo a coerência da resposta [98]. Um segundo termo também pode ser incluído para manter o desempenho do modelo na distribuição original dos dados de pré-treinamento.

Apesar disso, RLHF demonstra algumas limitações que residem primariamente na complexidade de capturar a diversidade de valores humanos através de *feedback* direto, onde a variabilidade e a subjetividade do *feedback* humano impõem desafios significativos para a modelagem precisa. Especificamente, a dificuldade de generalizar a partir do *feedback* coletado, a tendência para "*reward hacking*" (onde os modelos aprendem a maximizar indicadores imperfeitos de recompensas, destaca falhas fundamentais na abordagem [100]), e a incapacidade de garantir a robustez e segurança em políticas aprendidas são questões prementes, pois a modelagem de recompensas baseada em retroalimentação humana enfrenta a complexa tarefa de interpretar preferências intrinsecamente variadas e contextuais, levando a potenciais problemas de especificação incorreta e generalização inadequada. Além disso, as estratégias para mitigar essas limitações, como a avaliação cuidadosa e a introdução de robustez através do design de sistemas, são essenciais, mas apresentam seus próprios conjuntos de desafios, incluindo a avaliação precisa da qualidade dos modelos de recompensa e a implementação de soluções robustas que previnam exploração "adversarial" [101].

No quesito de vieses, em particular, vieses no *dataset* como já citado como um dos fatores que causam a problemática, um dos métodos de mitigação envolve a utilização do método de "disparo adversarial", que consiste em inserir gatilhos positivos nos *prompts* para influenciar o modelo a gerar saídas menos enviesadas. Esse método ajusta a forma como os *prompts* são apresentados ao modelo, que adiciona adjetivos positivos para alterar o contexto de forma favorável, o que demonstra a capacidade de ajustar dinamicamente o viés de um modelo através

de intervenções diretas nos dados de entrada. Essa técnica, ao promover uma alteração contextual positiva, permite que o modelo aprenda a normalizar as representações de forma mais equitativa, sem enfocar especificamente em vieses de nacionalidade ou preconceitos, mas sim em como a estruturação dos dados de entrada pode modular a saída do modelo [102].

Outras estratégias para mitigar os vieses no *dataset* envolvem a ampliação do *dataset* através da geração de dados contrafactuais para abordar estereótipos de gênero em línguas com morfologia rica. Esse método utiliza um campo aleatório de Markov (modelo estatístico usado para modelar informações espaciais ou temporais distribuídas) para inferir mudanças necessárias na morfologia das palavras, mantendo a concordância gramatical ao modificar o gênero gramatical de palavras, permitindo assim a transformação de sentenças para incluir representações femininas e masculinas equitativamente, que visa reduzir o viés de gênero sem comprometer a gramaticalidade das sentenças [103].

Entretanto, uma limitação notável nos estudos de mitigação de vieses, que permeia amplamente o campo de NLP, é o monolinguismo [104, 102]. Essa tendência não apenas restringe a aplicabilidade e a relevância das soluções desenvolvidas, mas também perpetua uma lacuna significativa entre os usuários da internet anglófonos (com inglês como língua nativa) e não anglófonos (com outros idiomas como língua nativa), o que é particularmente irônico ao levar em consideração que o objetivo primordial desses estudos é precisamente diminuir os vieses.

Não obstante, outros métodos de mitigação de vieses focam-se em aspectos distintos. Estratégias de treinamento equilibrado, por exemplo, buscam ajustar a representação demográfica por meio de reponderação e subamostragem, com o objetivo de garantir a equidade. Ao aplicar pesos inversamente proporcionais à frequência de grupos demográficos, tais métodos promovem um equilíbrio, no qual a reponderação ajusta a importância de cada exemplo, e a subamostragem modifica a composição do conjunto antes do treinamento [105]. Técnicas como o uso de modelos com mecanismos de adaptação baseada em atributos, oferecem um refinamento adicional ao balancear precisão e neutralidade [104]. Além dessas técnicas, a aprendizagem contrastiva de desvio emerge como uma solução direcionada à atenuação de características latentes enviesadas e à captura da influência dinâmica dos vieses, por meio da utilização de estratégias de amostra-

gem positiva de desvio para selecionar amostras positivas menos semelhantes e uma estratégia de amostragem negativa dinâmica para escolher amostras negativas mais semelhantes baseadas em vieses, o DCT aprimora a performance fora da distribuição, mantendo a eficácia dentro dela [106].

Outra limitação nas mitigações vem também do paradigma de aprendizado com poucos exemplos (*few-shot learning*), o qual, apesar do seu sucesso e grande difusão da técnica, possui a tendência de que modelos de linguagem pré-treinados, quando ajustados com poucos exemplos, adotem heurísticas de inferência baseadas em sobreposição lexical. Em outras palavras, isso significa que durante o processo de ajuste fino com poucos exemplos, os modelos podem começar a presumir, de forma equivocada, que pares de sentenças têm o mesmo significado simplesmente porque compartilham palavras semelhantes, ignorando o contexto mais amplo. Tal comportamento indica uma dependência excessiva de características superficiais, que podem prejudicar a capacidade do modelo de entender e processar a linguagem de maneira mais profunda e contextualizada. Para mitigar essa limitação, uma sugestão apresentada pela literatura é a adoção de técnicas de regularização que penalizam desvios significativos dos pesos pré-treinados, que ajudam a preservar o conhecimento linguístico geral adquirido durante o pré-treinamento e minimizam a adoção de heurísticas simplistas baseadas em coincidência de termos [107].

É importante frisar que a aplicação de conhecimento externo como forma de mitigação não é uma solução uniforme, um exemplo de tal abordagem que independe de tais externalidades é o *framework* CoNLI (*Chain of Natural Language Inference*) que emprega um processo de duas etapas envolvendo um agente de detecção e um agente de mitigação, no qual, inicialmente, o agente de detecção identifica alucinações (primariamente alucinações relacionadas a consistência e contradições) no texto gerado por meio de tarefas de inferência de linguagem natural hierárquica em níveis de sentença e entidade, e então, após a detecção, o agente de mitigação utiliza as percepções da fase de detecção para refinar a resposta, visando reduzir as alucinações enquanto preserva a essência do conteúdo original. Este método se distingue por não exigir acesso direto ou modificação do LLM subjacente, focando em vez disso na pós-edição do *output* gerado [108].

Um exemplo de método sequencial que se demonstrou particularmente popular em LLMs é CoT (*Chain of Thought*), utilizado em modelos de alta performance

como DeepSeek-R1 [109] e o3-mini [110]. Essa técnica possibilita que esses modelos abordem tarefas complexas de raciocínio através da simulação de um pensamento passo a passo [111]. Ao invés de simplesmente fornecer a solução, o modelo articula uma sequência coerente de raciocínio, quebrando o problema em etapas gerenciáveis [112]. Essa abordagem não só aprimora a capacidade do modelo de lidar com problemas intrincados, mas também oferece uma maior transparência no processo de tomada de decisão do modelo, tornando o raciocínio da IA mais compreensível e passível de análise [111].

Por outro lado, as técnicas não se limitam àquelas baseadas em *prompts* e geração aprimorada por recuperação. Um exemplo de uma abordagem alternativa é o *Epinet*, uma arquitetura baseada em ENNs (Epistemic Neural Network) que, diferente das redes neurais convencionais, as ENNs produzem distribuições preditivas conjuntas sobre várias entradas, permitindo a expressão de correlações e relacionamentos que influenciam a incerteza [113]. Esta atuação alternativa pode ser particularmente útil na redução de alucinações através da combinação com outros modelos, permitindo uma calibração mais precisa dos *logits* (logaritmo das chances de uma determinada palavra ou frase ser gerada) de saída do modelo [114].

2.5 Considerações Finais

Nesta fundamentação, abordaram-se as causas, as detecções e as estratégias de mitigação de alucinações em Modelos de Linguagem de Grande Escala, evidenciando avanços desde técnicas de aprendizado profundo até abordagens de aprendizado reforçado por *feedback* humano e verificação de fatos, mas também a complexidade subjacente às origens do fenômeno. Cada método contribui de maneira única para a compreensão e o combate às alucinações, refletindo a importância de uma abordagem multifacetada. Da mesma forma, cada análise fortaleceu o entendimento necessário para o desenvolvimento de métodos de intervenção eficazes.

CAPÍTULO

3

Trabalhos Relacionados

Neste capítulo, são discutidos estudos relevantes que investigam estratégias de mitigação de alucinações em LLMs. Esta seleção abrange avanços significativos e metodologias inovadoras que demonstraram eficácia na redução de alucinações, sendo cruciais para o embasamento teórico e prático do presente trabalho.

3.1 RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture

Uma divergência fundamental na mitigação de alucinações reside na forma como um modelo acessa informações: recuperando dinamicamente fatos externos ou baseando-se no conhecimento internalizado durante o treinamento. Conforme apontado no Capítulo 2, a abordagem RAG suplementa a entrada de dados com passagens de uma base de conhecimento pertinente, fornecendo um alicerce externo e verificável para reduzir respostas sem sentido [85]; sua força reside em seu dinamismo. Um estudo de caso no setor agrícola, por exemplo, demonstra que, ao fundamentar as respostas em dados contextualmente relevantes, a RAG produz resultados mais sucintos e factualmente exatos do que um modelo base [115]. Esse procedimento efetivamente terceiriza a verificação factual para uma base de conhecimento, tornando-a ideal para domínios onde as informações são

constantemente atualizadas.

Em contraste direto, o ajuste fino (*fine-tuning*) representa a segunda vertente do estudo, visando instilar conhecimento de um campo específico ou habilidades diretamente nos parâmetros do modelo. Tal abordagem permite um controle rigoroso sobre o estilo do conteúdo e o conhecimento especializado. Contudo, isso gera um compromisso fundamental. Embora o ajuste fino possa produzir modelos altamente eficientes e especializados, ele exige um investimento inicial significativo na curadoria de dados de treinamento abrangentes e na disponibilização dos recursos computacionais necessários. Adicionalmente, o conhecimento torna-se estático; uma vez concluído o treinamento, o modelo não consegue incorporar novas informações sem ser treinado novamente. Portanto, RAG e ajuste fino não são apenas duas técnicas, mas representam uma escolha estratégica central entre a fundamentação dinâmica e externa e a especialização estática e interna.

3.2 Retrieval Augmentation Reduces Hallucination in Conversation

A partir do paradigma RAG, pesquisas subsequentes buscaram adaptá-lo e especializá-lo para tarefas mais complexas, como diálogos multi-turno. O trabalho de Shuster et al. [116] exemplifica essa evolução ao reconhecer que uma única recuperação de dados para toda uma conversação é insuficiente. A RAG foi modificada por um processo sensível à conversação, que trata o diálogo como uma sequência de turnos e recupera documentos pertinentes para cada um deles. Este método de recuperação granular, aprimorado com mecanismos como os Poly-encoders para uma melhor pontuação de documentos [117], assegura que o contexto fornecido permaneça relevante à medida que o diálogo muda de foco, combatendo diretamente o desvio temático e a inconsistência.

Dentro desta mesma filosofia baseada na recuperação, arquiteturas alternativas também emergiram. A arquitetura FiD (Fusion-in-Decoder), por exemplo, difere de forma fundamental dos modelos RAG padrão em seu processamento. Enquanto um modelo RAG tipicamente processa cada documento recuperado de forma independente, o FiD concatena as codificações de todos os documentos, permitindo que o decodificador sintetize informações de múltiplas fontes simulta-

neamente. Esta abordagem constitui uma vantagem fundamental quando uma consulta exige a integração de diversas evidências. Avaliações desses métodos avançados em comparação com modelos tradicionais seq2seq, como BART e T5, confirmam que eles reduzem consistentemente as alucinações e melhoram a coerência, sublinhando que o futuro da fundamentação em dados não reside apenas na recuperação em si, mas na inteligência com que essa informação é integrada ao processo de geração.

3.3 Alleviating Hallucinations of Large Language Models through Induced Hallucinations

Em nítido contraste com os métodos que aumentam a entrada de dados do modelo, uma linha de pesquisa distinta investiga como restringir sua saída durante a geração. O trabalho de Zhang et al. [118] introduz uma abordagem inovadora e contraintuitiva com ICD (*Induce-then-Contrast Decoding*). Esta técnica representa uma mudança teórica significativa, através do aprimoramento da factualidade sem depender de qualquer base de conhecimento externa em tempo de inferência. A ideia central é executar dois processos de decodificação paralelos: um com o modelo base e outro com uma versão deliberadamente "mais fraca" do modelo, que é induzida a alucinar. Ao penalizar as saídas que o modelo propenso à alucinação favorece, a ICD guia o modelo base em direção a um território mais factual.

As implicações desta abordagem são profundas. Ao atingir um desempenho comparável ao do GPT-4 com modelos muito menores (como o Llama2 7B), a ICD demonstra um caminho de excelente custo-benefício para a confiabilidade. Ela desassocia a factualidade da necessidade de sistemas de recuperação em larga escala, abordando a sobrecarga computacional e os problemas de dependência inerentes às abordagens baseadas em aumentação por recuperação. Este método evidencia que o próprio processo de geração é um ponto de controle crítico para a mitigação, oferecendo uma alternativa promissora e econômica às estratégias centradas em dados.

3.4 Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback

A estratégia de mitigação final nesta seção avança em direção a sistemas híbridos e mais complexos, que combinam os princípios de fundamentação em dados com a verificação automatizada. O LLM-AUGMENTER [95] exemplifica essa abordagem em nível de sistema, criando um arcabouço que aprimora um LLM de "caixa-preta" sem qualquer necessidade de novo treinamento. Isso representa uma resposta direta aos desafios de escalabilidade impostos por modelos cada vez maiores, nos quais o ajuste fino é frequentemente inviável. O sistema opera em um ciclo iterativo no qual aumenta a instrução inicial (*prompt*) com evidências recuperadas, gera uma resposta e, em seguida, utiliza um módulo de utilidade automatizado para verificar a resposta e fornecer retorno (*feedback*) para o refinamento.

Esta arquitetura é mais do que um simples fluxo de processamento (*pipeline*); ela é modelada como um processo de decisão de Markov, com componentes como uma Memória de Trabalho e um executor de ações que lhe permitem rastrear dinamicamente os estados do diálogo e refinar as saídas. Ao fazer isso, o LLM-AUGMENTER sintetiza o conhecimento externo com uma camada adicional de verificação iterativa. Resultados empíricos demonstram melhorias significativas na factualidade e na coerência, aferidas por medidas como BLEU e BERTScore. Esta abordagem oferece um caminho pragmático e escalável para o futuro, alavancando o poder dos LLMs existentes ao envolvê-los em uma estrutura de suporte inteligente que ativamente verifica e corrige sua principal fraqueza: a tendência a alucinar.

3.5 Considerações Finais

As estratégias discutidas neste capítulo ilustram que a mitigação de alucinações não é um desafio monolítico, mas sim multifacetado, que pode ser abordado a partir de diversos ângulos. Os trabalhos apresentados revelam um espectro de abordagens: desde a fundamentação em dados centrada na entrada, própria

da RAG (Seções 3.1 e 3.2), até o controle de decodificação centrado na saída, característico da ICD (Seção 3.3), e, por fim, o refinamento holístico e iterativo de sistemas como o LLM-AUGMENTER (Seção 3.4). Cada abordagem apresenta um conjunto único de compromissos entre custo computacional, dependência de dados externos e complexidade arquitetônica. Por exemplo, a escolha entre RAG e ICD reflete uma decisão entre fornecer conhecimento externamente e direcionar a geração internamente. O surgimento de técnicas especializadas, como a RAG sensível ao diálogo, e de arcabouços em nível de sistema, como o LLM-AUGMENTER, demonstra um campo em amadurecimento que avança em direção a soluções mais nuançadas e específicas para cada tarefa.

Diferentemente das investigações mais específicas encontradas nesses estudos, este trabalho propõe uma contribuição distinta. A partir de uma revisão sistemática da literatura, estabeleceu-se uma taxonomia para classificar essas variadas filosofias de mitigação. Essa categorização fundamentou a seleção de técnicas representativas de diferentes abordagens para uma análise empírica comparativa. As técnicas selecionadas, escolhidas por sua reprodutibilidade, foram então submetidas a um espectro variado de cenários de avaliação, incluindo *benchmarks* de perguntas e respostas, medidas de sumarização e uma avaliação em um contexto jurídico específico da língua portuguesa. Este delineamento metodológico visa oferecer um panorama mais completo de como a eficácia das estratégias varia conforme a categoria do método, a tarefa e o critério de avaliação, permitindo, assim, um discernimento mais profundo de suas potencialidades e limitações.

CAPÍTULO

4

Material e Métodos

Este capítulo descreve a metodologia utilizada para identificar, avaliar e selecionar métodos de mitigação de alucinações em LLMs. A pesquisa foi conduzida como uma revisão sistemática da literatura, seguindo as diretrizes propostas por Kitchenham e Charters [119]. O objetivo principal é fornecer uma análise comparativa e reprodutível da eficácia, aplicabilidade e implementação desses métodos, que visa facilitar a identificação de metodologias eficazes e permitir uma avaliação detalhada de suas vantagens e limitações específicas.

4.1 Planejamento da Revisão

O processo de revisão, ilustrado na Figura 4.1, foi segmentado para assegurar uma análise aprofundada e precisa. O processo inicia-se pela coleta sistemática dos métodos de mitigação de alucinações, seguida pela categorização meticulosa de cada técnica em sua respectiva classe. Posteriormente, ocorre a inclusão e exclusão dos estudos, em que são selecionadas as estratégias de mitigação mais apropriadas, o que culmina na fase final: a reprodução dos experimentos e sua subsequente avaliação por meio de *benchmarks* padronizados.

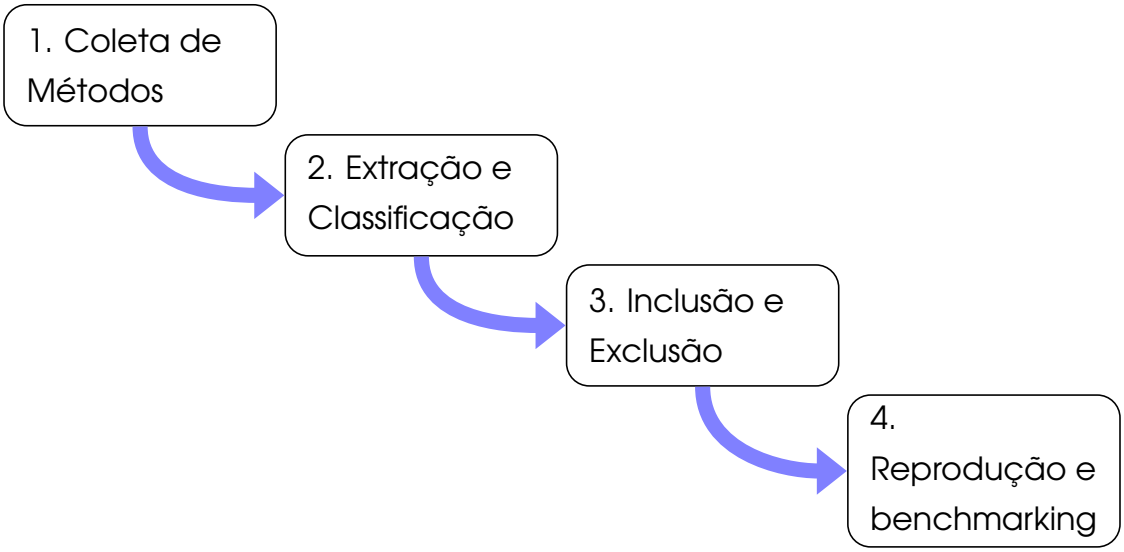

Figura 4.1: Segmentos necessários para obtenção dos resultados.

Para delimitação rigorosa do escopo da revisão, adotou-se o *framework* PICOC (*Population*, *Intervention*, *Comparison*, *Outcomes*, *Context*), conforme detalhado na Tabela 4.1.

Tabela 4.1: Definição do escopo com a utilização do *framework* PICOC.

Componente	**Especificação**
População	Métodos de mitigação de alucinações em LLMs (modelos com mais de 1 bilhão de parâmetros).
Intervenção	Técnicas de mitigação de alucinações.
Comparação	Comparação entre os diferentes métodos de mitigação de alucinações entre categorias.
Resultados	Eficácia na redução de alucinações (medida por *benchmarks*), reprodutibilidade (avaliada pela disponibilidade de código e documentação) e aplicabilidade (considerando o custo computacional).
Contexto	Estudos publicados, ou disponibilizados como *pre-prints* com data prevista de publicação, entre 2021 e o primeiro semestre de 2025. A busca e seleção dos estudos foram conduzidas até maio de 2025 .

Para coordenar a revisão, as seguintes perguntas de pesquisa foram definidas como principais:

- Quais métodos de mitigação de alucinações em LLMs estiveram disponíveis na literatura desde o ano de 2021 até o primeiro semestre de 2025?
- Em quais categorias principais esses métodos de mitigação podem ser classificados?
- Como a eficácia se compara entre essas categorias de métodos?
- Quais são as limitações e desafios para a reprodução desses métodos, com base na disponibilidade de código e na qualidade da documentação?

4.2 Estratégia de Busca

Para identificar métodos de mitigação de alucinações em modelos de linguagem de grande escala, foi adotada uma estratégia de busca direcionada. Essa estratégia consistiu na execução de buscas individuais, combinando o termo-chave "LLM" com cada um dos seguintes termos: "*hallucination*", "*mitigation hallucination*", "*factuality*", "*factual*" e "*accuracy*". Essa abordagem garante que os resultados sejam específicos para LLMs, abrangendo tanto pesquisas que mencionam explicitamente o termo "alucinação" quanto aquelas que focam na veracidade e precisão das informações geradas.

A estratégia de busca definida foi aplicada em diversas fontes de referência. No repositório arXiv, empregou-se a busca avançada com a categoria "*Computer Science* (cs)", a fim de filtrar os resultados para soluções técnicas e computacionais relevantes para a mitigação de alucinações em LLMs. Em plataformas como o `GitHub` e a `Hugging Face`, que se concentram no desenvolvimento e treinamento de modelos de linguagem, o foco no termo "alucinação" em combinação com "LLM" foi o suficiente para gerar resultados pertinentes, sem a necessidade de filtragem adicional por categoria. A escolha dessas plataformas se justifica por estarem consolidadas como os veículos primários e mais eficientes para a disseminação de pesquisa e código na comunidade de desenvolvimento em NLP.

O ano de "corte" é a partir de 2021. A escolha do ano de 2021 como marco inicial justifica-se pelo aumento da escala e complexidade dos LLMs a partir desse período. O aumento no número de parâmetros e no volume de dados de treinamento contribuiu para a maior frequência de alucinações, o que, por consequência, aumenta a demanda por metodologias de mitigação.

Em complemento à definição do limite inferior de 2021, o limite superior para a coleta de dados foi no mês de maio de 2025. A revisão sistemática, portanto, abrangeu o período desde o primeiro semestre de 2021 até o primeiro semestre de 2025.

O uso de ferramentas de mapeamento de citações, como o *Litmaps* e o *Influence Map*, complementou a estratégia, permitindo identificar conexões entre estudos e revelar referências cruciais que não seriam capturadas por palavras-chave, sendo esta abordagem responsável pela descoberta de sete métodos de mitigação. A Figura 4.2 exemplifica um mapa de citações gerado a partir do

artigo "*Few-shot Learning with Retrieval Augmented Language Models*" [2].

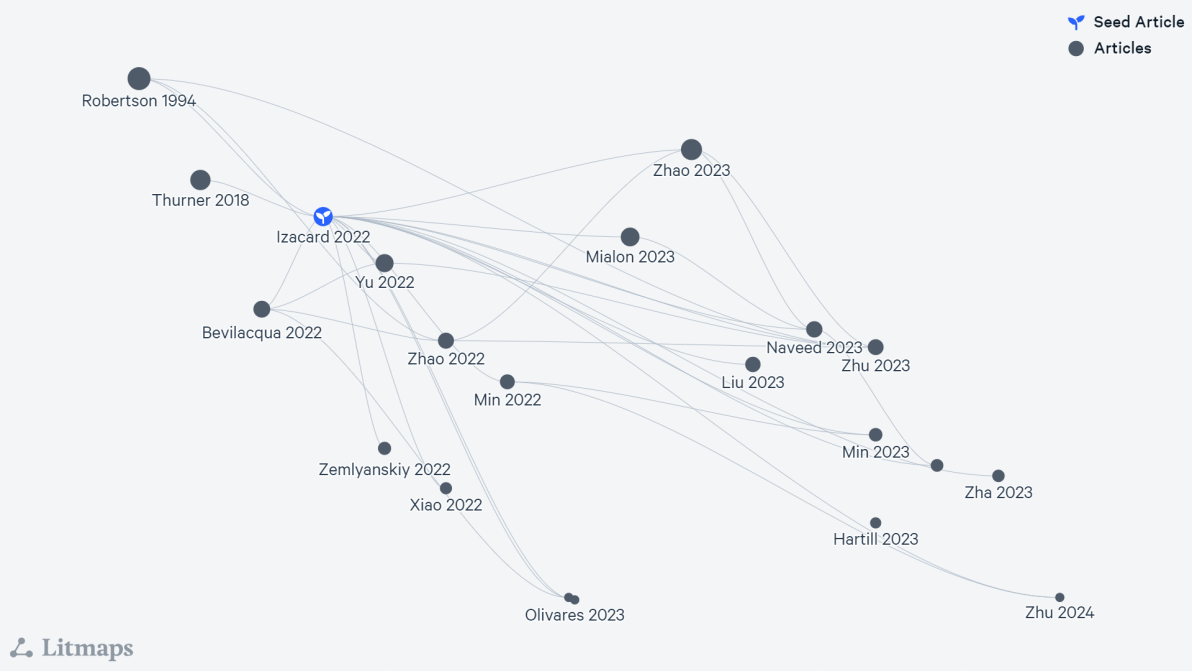

Figura 4.2: Mapa de citações gerado pelo Litmaps a partir do artigo *Few-shot Learning with Retrieval Augmented Language Models* [2].

A busca identificou 93 trabalhos relevantes sobre mitigação de alucinações. A distribuição temporal dos estudos selecionados (Figura 4.3) exibe um pico em 2023. Durante a análise dos trabalhos coletados, notou-se uma tendência crescente em publicações mais recentes (especialmente a partir do final de 2024 e início de 2025) que exploram a mitigação de alucinações em contextos multimodais, embora o escopo final desta revisão tenha-se mantido focado em LLMs primariamente textuais.

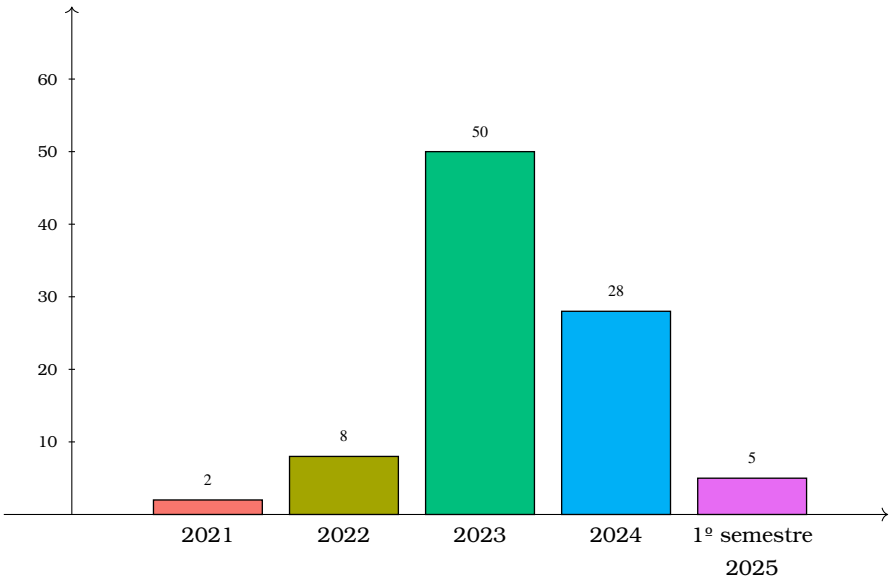

Figura 4.3: Distribuição dos métodos de mitigação coletados por ano (total de 93).

4.3 Extração e Classificação

Os métodos selecionados atuam de formas distintas e visam a diferentes tipos de alucinações. Uma comparação direta sem categorização seria inadequada. Portanto, os métodos foram categorizados para permitir uma análise precisa e contextualizada, agrupando as técnicas conforme suas características. Adicionalmente, a categorização oferece um panorama do campo, indicando a popularidade de cada abordagem. Isso permite investigar as razões para essa distribuição, como a eficácia ou a disponibilidade dos métodos.

Iniciando-se pela categoria **Otimização de Modelo**, a qual engloba técnicas que aprimoram a arquitetura do modelo e seu processo de aprendizado, e que inclui estratégias como o enriquecimento de dados, ajuste fino supervisionado, e a experimentação com diferentes configurações de hiper-parâmetros. Estas abordagens visam melhorar diretamente a capacidade do modelo de gerar respostas precisas e relevantes, de forma a minimizar a incidência de alucinações através de uma base de conhecimento mais sólida e representativa. Uma dessas técnicas é a previamente descrita DCT (*Debiasing Contrastive Learning*) que visa otimizar o processo de aprendizagem do modelo para atenuar o impacto de recursos tendenciosos, que resulta em melhor desempenho em diversos conjuntos

de dados.

Em seguida, o grupo **Aprimoramento de Inferência e Saída** abrange técnicas aplicadas no momento da inferência com o propósito de refinar as saídas geradas pelos modelos. Isso inclui métodos de pós-processamento, como a aplicação de filtros baseados em regras ou a correção de erros via *feedback* em tempo real, com o objetivo de melhorar a qualidade e a relevância das respostas do modelo, além de reduzir as discrepâncias e os erros. Técnicas como RLHF (*Reinforcement Learning from Human Feedback*) utilizam esse mecanismo de *feedback* para refinar continuamente a saída do modelo.

Penúltimo, o conjunto de métodos caracterizado como **Conhecimento Externo e Interação** refere-se aos métodos que utilizam fontes de dados externas ou mecanismos de *feedback* interativo para enriquecer ou corrigir as respostas geradas pelo modelo. Essa abordagem reconhece a importância de uma perspectiva mais ampla, além dos limites do conjunto de dados de treinamento, o que permite que o modelo se beneficie de informações atualizadas e contextualmente relevantes. Um exemplo de método já bem popular é o RAG (*Retrieval Augmented Generation*), o qual incorpora a capacidade de consultar bases de conhecimento externas, como a `Wikipedia`, e que, assim, fornece respostas mais abrangentes e precisas.

Para concluir, a categoria **Métodos Experimentais** trata-se de um conjunto de técnicas inovadoras e especulativas, como o uso de sistemas multi-agentes e a análise dos estados internos do modelo, que buscam entender e mitigar alucinações através de novas perspectivas de análise do modelo. Essa categoria reflete a vanguarda da pesquisa em mitigação de alucinações, e explora novas fronteiras com o objetivo de compreender e melhorar o comportamento dos modelos. Para melhor ilustrar os tipos de métodos nessa categoria, é necessário retornar ao último parágrafo da Seção 2.4, onde é demonstrado o uso de redes neurais epistêmicas para modelar a incerteza e aprimorar a detecção de alucinações em grandes modelos de linguagem. Ao treinar ENNs sobre os modelos existentes, são realizados experimentos que visam melhorar a precisão dos resultados e entender melhor os processos de tomada de decisão do modelo.

No gráfico apresentado na Figura 4.4, observa-se a distribuição dos métodos na fase traçada nesta seção.

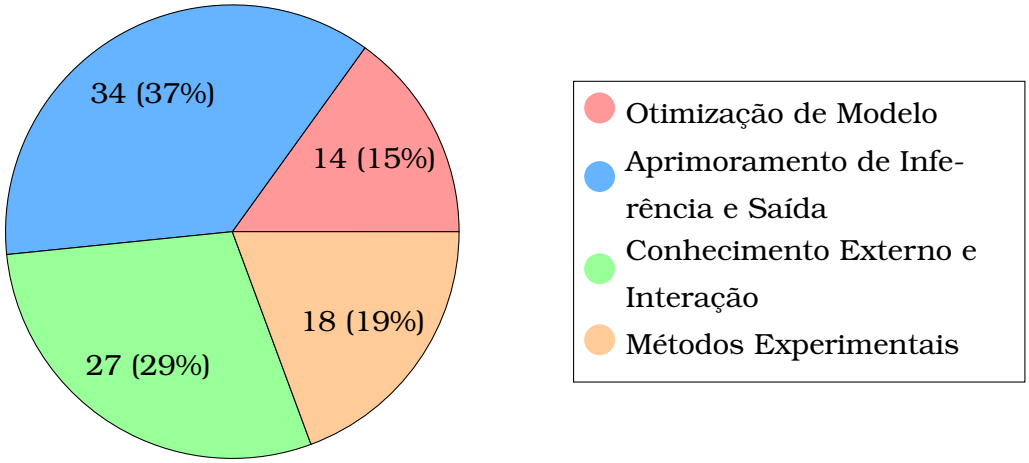

Figura 4.4: Visualização da distribuição dos métodos em suas categorias.

4.4 Inclusão e Exclusão dos Métodos

A definição dos critérios de inclusão e exclusão é uma etapa essencial na metodologia de uma revisão sistemática. Esses critérios asseguram a objetividade da seleção dos estudos, delimitam o escopo do trabalho e garantem sua reprodutibilidade. Os parâmetros que orientaram a seleção dos trabalhos nesta análise são detalhados a seguir.

4.4.1 Critérios de Inclusão

Esta revisão sistemática da literatura prioriza a inclusão de estudos com implementações de código prontamente disponíveis e acessíveis, o que levou à exclusão de vários materiais.

Foram considerados elegíveis para inclusão artigos científicos completos, *preprints* (disponíveis no repositório arXiv) e relatórios técnicos que descrevessem métodos de mitigação de alucinações em LLMs.

Os estudos devem apresentar um método novo, uma modificação substancial de um método existente ou uma maior exploração de abordagem pré-existente, com o objetivo principal de mitigar alucinações; estudos focados apenas na detecção de alucinações, sem propor uma solução, não foram incluídos. Além disso, foram também considerados apenas estudos publicados em inglês, devido à abrangência e uso predominante deste idioma em conjunto com a aplicação

apropriada das palavras-chave, visto que as mesmas estão em inglês.

4.4.2 Critério de Exclusão

A primeira etapa da exclusão vem de artigos sem código acompanhante, como aqueles com código que se tornou inacessível devido a mudanças no repositório ou outras circunstâncias, e estudos com bases de código ainda não disponibilizadas. A motivação de tal escolha foi que a reprodutibilidade é um pilar de rigorosidade que oferece a capacidade de executar e verificar independentemente a implementação de um método de mitigação proposto, sendo crucial para validar sua eficácia e compará-lo a abordagens alternativas.

Outra exclusão feita refere-se a técnicas que, apesar de interesse teórico, revelam-se economicamente inviáveis, como por exemplo uma replicação que dependeria de acesso a modelos com custos proibitivos. Com base nesta decisão, que se fundamenta na premissa de que a investigação científica deve ser acessível e replicável, enfatiza-se a importância de métodos que possam ser adotados por outros pesquisadores sem incorrer em custos excessivos. Dessa maneira, ao priorizar abordagens tanto inovadoras quanto acessíveis, promove-se a validação e replicação independentes, resultando no estabelecimento de um equilíbrio entre avanço teórico e aplicabilidade prática.

O terceiro critério de exclusão filtra artigos que, apesar de fornecerem o código-fonte, não possuem documentação adequada, definida minimamente pela presença de instruções para a preparação do ambiente de execução. Entende-se que o código isolado, embora útil, é insuficiente para garantir a replicabilidade e a compreensão integral dos métodos. A documentação é essencial para replicar experimentos, aplicar metodologias em novos contextos e identificar limitações, facilitando a validação, a inovação e o avanço do conhecimento coletivo.

A exclusão final concentra-se integralmente na comparabilidade das técnicas entre as diferentes categorias atribuídas a cada método, conforme descrito na Seção 4.3. O processo consiste em uma análise direta que visa identificar o modelo mais utilizado entre os métodos restantes, onde verifica-se se cada categoria possui uma técnica que emprega o referido modelo, a fim de possibilitar a comparação. Consequentemente, esse processo de filtragem resultará em um número de técnicas que será múltiplo do número de categorias, garantindo a uniformidade e a equidade da análise.

A Tabela 4.2 contém os métodos restantes após a terceira etapa de exclusão, onde cada citação corresponde a uma técnica situada em sua respectiva categoria.

Tabela 4.2: Lista dos métodos com documentação.

Otimização de Modelo	**Aprimoramento de Inferência e Saída**	**Conhecimento Externo e Interação**	**Métodos Experimentais**
[120] [17] [121] [122] [31]	[123] [124] [125] [126] [127] [128] [129] [130] [90] [131] [132] [133] [134]	[135] [136] [44] [137] [138] [139] [140] [141] [142] [143] [144]	[145] [146] [147] [114] [148] [149]

O plano inicial previa a seleção de múltiplos métodos por categoria (oito ou mais no total). Contudo, para garantir que as diferenças de desempenho fossem atribuíveis exclusivamente aos métodos de mitigação, tornou-se essencial que todas as técnicas fossem avaliadas sobre o mesmo modelo LLM base. O modelo Llama 2 com 7 bilhões de parâmetros foi o mais prevalente entre os trinta e cinco estudos da seleção preliminar. A adoção deste critério de uniformidade levou à seleção final de quatro métodos, um representante para cada categoria. Essa decisão, embora tenha reduzido o escopo quantitativo da análise, foi fundamental para a validade e a comparabilidade dos resultados.

A Figura 4.5 ilustra o processo sistemático de seleção concluído após todas as exclusões feitas.

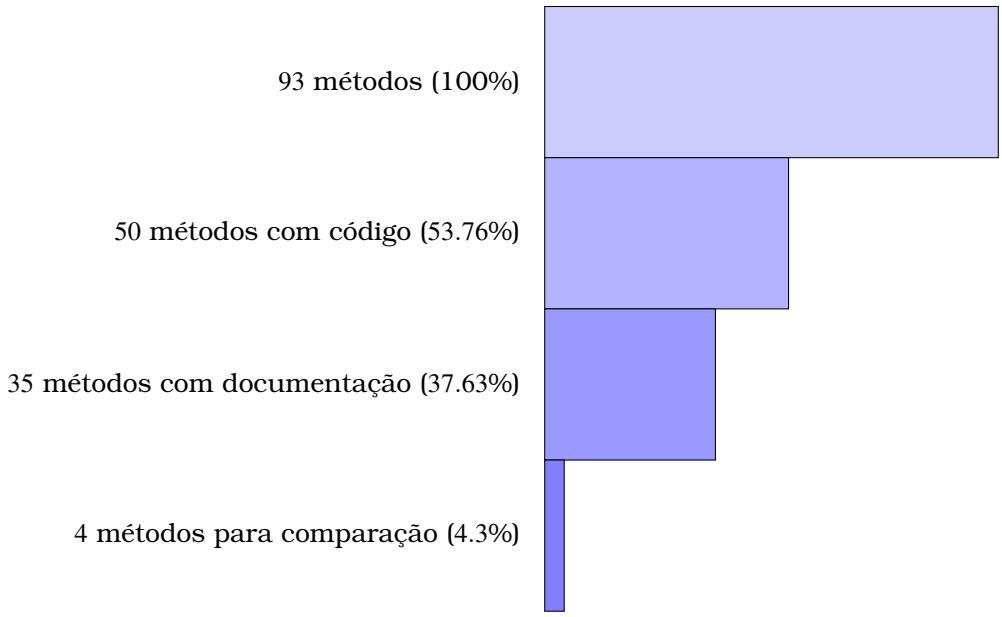

Figura 4.5: Visualização da filtragem.

4.4.3 Métodos Selecionados para Análise

Nesta seção, apresentam-se os métodos utilizados na condução das avaliações e *benchmarks*, sendo que todas as técnicas utilizam como base Llama2-7b.

Mitigating Hallucination in Large Language Models via Knowledge Consistent Alignment

O conflito fundamental entre o conhecimento externo, proveniente dos dados de treinamento de alinhamento, e o conhecimento intrínseco, memorizado durante o pré-treinamento, é abordado pelo método *Knowledge Consistent Alignment* (KCA), pois a inconsistência entre essas duas fontes de conhecimento frequentemente gera alucinações. Essa técnica, categorizada como **Otimização de Modelo**, já que a mesma visa mitigar os erros de geração através de estratégias de ajuste fino, incluindo ajuste com contexto adicional (*Open-Book Tuning*), descarte de dados inconsistentes (*Discarding Tuning*) e treinamento com recusa (*Refusal Tuning*). Conjuntamente é empregada a detecção de inconsistências no conhecimento por meio de três etapas: determinar se as instruções requerem conhecimento externo ou intrínseco, gerar conhecimento de referência suplementar para instruções que necessitem desse contexto adicional e formular testes de

múltipla escolha baseados no conhecimento gerado para avaliar a consistência das respostas do modelo [121].

Alleviating Hallucinations of Large Language Models through Induced Hallucinations

O *Induce-then-Contrast Decoding* (ICD) [118], já apresentado na Seção 3.3, se destaca como uma técnica de **Aprimoramento de Inferência e Saída** ao intervir diretamente no cálculo das probabilidades dos *tokens* durante a geração. A primeira etapa estabelece um modelo LLM auxiliar, intencionalmente treinado ou instruído para internalizar e reproduzir padrões de informação não factual, servindo como um "especialista em alucinações". Este modelo complementar não substitui o LLM base, mas atua como uma referência negativa cujas tendências de geração serão confrontadas.

A essência do aprimoramento ocorre na segunda etapa, através de um processo de decodificação por diferenciação em tempo de inferência. Para cada unidade textual a ser gerada, o ICD avalia as projeções tanto do LLM primário quanto do modelo auxiliar propenso a erros. As "sugestões" de *tokens* provenientes da instância geradora de inverdades têm sua influência ativamente reduzida na distribuição de chances final, enquanto as indicações consideradas verídicas pelo modelo principal são amplificadas. Crucialmente, para preservar a qualidade e coerência da linguagem, esse desincentivo é aplicado seletivamente: apenas os elementos que o modelo primário já considera com alguma plausibilidade, mas que também são fortemente sugeridos pela instância falível, são efetivamente atenuados. Esse mecanismo de contraste direcionado refina o resultado final, guiando o LLM principal para respostas mais acuradas e corretas.

WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia

Na categoria de **Conhecimento Externo e Interação**, o método WikiChat aplica uma abordagem de recuperação estruturada em sete etapas para mitigar alucinações factuais em LLMs. Diferentemente das abordagens convencionais de recuperação e geração, que frequentemente alucinam em caso de insuficiência de material recuperado, o WikiChat sintetiza sumários dos trechos extraídos da Wikipedia e utiliza um modelo para gerar respostas, das quais extrai reivindica-

ções específicas. Cada reivindicação é verificada em relação ao corpus original por meio de *prompts* baseados em raciocínio encadeado, garantindo que apenas as informações suportadas sejam mantidas. Essa técnica assegura a produção de respostas fundamentadas e visa reduzir o risco de alucinações [44].

Representation Engineering: A Top-Down Approach to AI Transparency

O último método da última categoria, **Métodos Experimentais**, trata-se do *Representation Engineering* (RepE) [149], visa aumentar a transparência e o controle de redes neurais, especialmente LLMs, através da manipulação de suas representações internas pela técnica LAT (*Linear Artificial Tomography*) para extrair vetores de leitura que representam conceitos de alto nível, como a honestidade, que será o foco dos testes.

O LAT opera ao apresentar estímulos contrastantes (por exemplo, *prompts* para gerar respostas honestas vs desonestas), ao coletar as ativações correspondentes em camadas intermediárias da rede e ao aplicar *Principal Component Analysis* (PCA) para identificar a direção no espaço de ativação que melhor separa os dois estímulos. Esse vetor de leitura pode então ser usado para monitorar o nível de "honestidade" das ativações durante a geração de texto e, mais importante, para ajustar o comportamento do modelo, estimulando ou suprimindo as ativações relacionadas à honestidade de forma que possa reduzir alucinações.

4.5 Avaliações e Benchmarks

Para a avaliação quantitativa da performance dos métodos selecionados, emprega-se um conjunto abrangente de instrumentos de avaliação. Este conjunto inclui três medidas baseadas em perguntas e respostas (TruthfulQA, ARC e OAB), um *benchmark* estabelecido para mensurar correspondência lexical (ROUGE) e outro para similaridade semântica (BERTScore), os quais se destacam por utilizarem abordagens metodológicas distintas. A descrição a seguir detalha cada um desses componentes, explicitando os critérios que fundamentaram sua escolha.

4.5.1 TruthfulQA

O primeiro método de avaliação selecionado foi o *benchmark* denominado *TruthfulQA* que se trata de perguntas e respostas projetadas para avaliar a capa-

cidade dos modelos de linguagem natural de distinguir entre informações verídicas e erros comuns replicados entre seres humanos, as denominadas “falsidades imitativas”. O *dataset* é composto por 817 perguntas distribuídas por 38 categorias que abrangem uma diversidade de tópicos, incluindo saúde, direito, finanças e política. Toda pergunta é acompanhada de respostas de referência, classificadas como verdadeiras ou falsas, com base em evidências científicas. A avaliação é realizada em um cenário de *zero-shot*, ou seja, sem treinamento prévio específico, de modo a testar a habilidade dos modelos de evitar afirmações incorretas e fornecer respostas fundamentadas cientificamente [55].

A escolha é justificada pelos resultados, que elucidam um aspecto crucial do comportamento dos modelos de linguagem de grande escala: o fenômeno denominado “escalabilidade inversa” [150]. Conforme discutido anteriormente na Seção 2.2.1, esse fenômeno demonstra que modelos maiores tendem a reproduzir com maior frequência falácias amplamente difundidas, o que contraria a suposição de que a ampliação da escala do modelo levaria necessariamente a uma melhoria na precisão. Dessa forma, este *benchmark* se torna um recurso relevante para a avaliação de alucinações que envolvem conflitos com informações factuais.

4.5.2 ARC

Em sequência, tem-se como *benchmarks* de perguntas e respostas ARC, designado para avaliar a capacidade dos modelos de linguagem natural em realizar raciocínios complexos e aplicar conhecimento de senso comum em contextos científicos de nível fundamental. O *dataset* é composto por 7.787 questões de ciências naturais, divididas em dois subconjuntos: o conjunto *Challenge* (2.590 questões), que compreende perguntas que algoritmos convencionais de recuperação de informações não foram capazes de resolver, e o conjunto *Easy* (5.197 questões) que engloba o restante [151].

A escolha do ARC como *benchmark* se justifica por desafiar os modelos a demonstrar compreensão real dos conceitos mais abstratos e lógicos, sendo um recurso importante para avaliar o potencial dos modelos de IA em avançar além da correspondência superficial de padrões textuais, que em decorrência disso é amplamente utilizado como parâmetro medidor de excelência em LLMs mais recentes [152, 33].

4.5.3 OAB

Para completar o conjunto de avaliações Q&A, foi selecionado um *dataset* para não só avaliar se as técnicas de mitigação de alucinações contribuem para uma melhora em geração de texto em idiomas diferentes do inglês, mas também para testar a capacidade do modelo de lidar com contextos específicos, como o da prova da OAB (Ordem dos Advogados do Brasil). Esse conjunto de dados é composto por 2.210 perguntas de múltipla escolha, abrangendo provas aplicadas entre 2010 e 2018.

Essa prova representa um desafio significativo, uma vez que envolve a compreensão e geração de conteúdo complexo em português, além de exigir uma compreensão profunda das nuances legais e da lógica jurídica.

O uso do *dataset* baseado em questões da OAB é particularmente relevante por se tratar de um exemplo representativo da linguagem formal e especializada. Dessa forma, ele oferece uma métrica objetiva para examinar a proficiência do modelo na produção de respostas coesas e adequadas em contextos onde o rigor conceitual é essencial.

4.5.4 ROUGE

A medida de avaliação ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*), originalmente introduzida por Lin et. al [153], é uma das mais antigas não somente no campo de NLP mas também em tradução automatizada, antecedendo até mesmo a criação das primeiras LLMs. Entretanto, ROUGE continua sendo relevante por dois motivos: sua evolução constante ao longo dos anos e seus *insights* valiosos, graças à sua forma peculiar de estimar corretude.

ROUGE não é apenas um método específico de avaliação, mas sim um conjunto de medidas para avaliar a qualidade de resumos, as quais os comparam com resumos de referência por meio de diferentes abordagens. No contexto deste trabalho, serão utilizados quatro tipos de ROUGE. Inicia-se com os métodos conhecidos atualmente como ROUGE-1 e ROUGE-2, que calculam a quantidade de n-gramas sobrepostos entre o resumo gerado automaticamente e o resumo de referência. Essas medidas são denominadas de ROUGE-N, onde “N” representa o número de palavras em cada n-grama. A fórmula utilizada para calcular o ROUGE-N está definida na Equação 4.1.

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{Reference Summaries}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \text{Reference Summaries}} \sum_{gram_n \in S} Count(gram_n)} \tag{4.1}$$

Onde:

- S representa um resumo de referência;
- $gram_n$ se trata de um n-grama de tamanho n presente no resumo de referência; [1]
- $Count_{match}(gram_n)$ é o número de n-gramas coincidentes entre o resumo gerado e o de referência;
- $Count(gram_n)$ é o número total de n-gramas no resumo de referência.

Ainda no texto original se utiliza também o método ROUGE-L, que baseia-se na LCS (*Longest Common Subsequence*) entre o resumo gerado automaticamente e o resumo de referência, a técnica considera tanto a precisão quanto o *recall* das correspondências de LCS, e combina essas medidas em uma única medida F, onde β ajusta a importância relativa entre elas. Esta abordagem permite capturar a similaridade estrutural e a ordem das palavras nos resumos, para proporcionar uma avaliação mais robusta da qualidade do resumo gerado em comparação com medidas que consideram apenas a sobreposição de n-gramas.

Além disso, o ROUGE-L não requer que as correspondências de palavras sejam consecutivas, o que permite identificar correspondências em sequência que refletem melhor a fluidez e a coerência do texto. Ao integrar LCS ao ROUGE-L automaticamente consideram-se os n-gramas mais longos em sequência comum, o que elimina a necessidade de definir previamente o tamanho dos n-gramas [154]. A fórmula dessa medida de ROUGE está definida na Equação 4.2.

$$\text{ROUGE-L} = \frac{(1+\beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}} \tag{4.2}$$

Onde:

- R_{LCS} (recall) é calculado como:

$$R_{LCS} = \frac{\text{LCS}(\text{Resumo Gerado}, \text{Resumo de Referência})}{|\text{Resumo de Referência}|} \tag{4.3}$$

[1] Nesse contexto, n denota o tamanho da n-grama especificada pelo N de ROUGE-N.

ou seja, é a razão entre o comprimento da LCS e o comprimento total do resumo de referência;

- P_{LCS} (precisão) é dado por:

$$P_{LCS} = \frac{\text{LCS}(\text{Resumo Gerado}, \text{Resumo de Referência})}{|\text{Resumo Gerado}|} \tag{4.4}$$

que é a razão entre o comprimento da LCS e o comprimento do resumo gerado;

- β é um parâmetro que ajusta a importância relativa entre a precisão e o recall (por exemplo, $\beta = 1$ balanceia igualmente ambos).

Por último, será utilizado ROUGE-LSum, que se trata de uma modificação feita de ROUGE-L, que visa mitigar as deficiências desta medida na avaliação de sumários multi-sentença; enquanto a técnica tradicional calcula a maior subsequência comum entre os sumários candidato e referência como um todo, a variação introduz uma granularidade a nível de sentença. A ideia central é que a ordem das sentenças não deve influenciar significativamente a pontuação, desde que o conteúdo semântico seja preservado; para isso, ambos os sumários são segmentados em sentenças e a maior subsequência comum é calculada para cada par de sentenças (candidato vs referência). A pontuação final do ROUGE-LSum é então derivada da soma dessas subsequências comuns, normalizada pelo comprimento dos sumários [155]. A fórmula utilizada para calcular o ROUGE-LSum está definida na Equação 4.5.

$$\text{ROUGE-LSum} = \frac{(1+\beta^2) \cdot R_{lcssum} \cdot P_{lcssum}}{R_{lcssum} + \beta^2 \cdot P_{lcssum}} \tag{4.5}$$

Onde:

- R_{lcssum} (recall) é calculado como:

$$R_{lcssum} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, s)}{m} \tag{4.6}$$

isto é, a soma das maiores LCS em união (entre cada sentença r_i do resumo de referência e o resumo candidato s) normalizada pelo total de m palavras do resumo de referência;

- P_{lcssum} (precisão) é dado por:

$$P_{lcssum} = \frac{\sum_{i=1}^{u} LCS_{\cup}(r_i, s)}{n} \tag{4.7}$$

  ou seja, a mesma soma das LCS em união, dividida pelo total de n palavras do resumo candidato;

- O parâmetro β possui a mesma definição e função apresentada na Equação 4.2.

4.5.5 BERTScore

Como já mencionado no Capítulo 2 na Seção 2.3, BERTScore (mais precisamente a versão apresentada no artigo de Zhang et al [72]) se assemelha ao ROUGE em sua busca de similaridade, mas difere no método, o qual, em vez de utilizar a medida por n-gramas, empregam-se representações contextualizadas de palavras geradas por modelos pré-treinados de *transformers* (como BERT), que calculam a semelhança semântica entre as sentenças ao medir o alinhamento entre os *embeddings* das palavras das duas sequências comparadas. Através dessa técnica, são utilizadas três fórmulas: uma para precisão, outra para *recall* e uma terceira para o F1-*score*.

A análise começa com a precisão, uma medida focada na confiabilidade das predições afirmativas. Em seu sentido clássico, ela representa a fração de previsões positivas que são genuinamente corretas, calculada pela razão entre Verdadeiros Positivos (TP) e a soma de TP com Falsos Positivos (FP) conforme a equação $Precisão = \frac{TP}{TP+FP}$, sendo crucial para garantir que um modelo não cometa muitos erros de falsos positivos [156]. Ao adaptar este princípio, o BERTScore avalia a qualidade do texto gerado ($\hat{x}$) em relação a um texto de referência (x). A precisão do BERTScore, portanto, mede o quão bem cada *token* do texto gerado corresponde a um *token* semanticamente similar no texto de referência. Essa correspondência é otimizada por um algoritmo de correspondência gulosa, conforme detalhado na Equação 4.8, permitindo que a medida lide com variações lexicais e sintáticas inerentes à linguagem.

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \left(x_i^{\top} \hat{x}_j \right) \tag{4.8}$$

Onde:

- x e $\hat{x}$ representam os vetores de referência e candidatos respectivamente;
- $\max_{x_i \in x} \left(x_i^\top \hat{x}_j \right)$ corresponde ao *greedy matching*.

O *Recall*, por sua vez, avalia a cobertura dos dados de referência. Diferente da precisão, seu foco é medir a fração de todas as instâncias verdadeiramente positivas que foram capturadas pelo modelo, utilizando a razão entre Verdadeiros Positivos (TP) e a soma de TP com Falsos Negativos (FN), conforme a equação $Recall = \frac{\text{TP}}{\text{TP+FN}}$. O BERTScore traduz este princípio para a análise textual ao inverter a perspectiva da sua medida de precisão: em vez de partir do texto gerado ($\hat{x}$), ele verifica o quão bem cada *token* do texto de referência (x) está semanticamente representado na predição. Para tal, a média é calculada sobre os elementos de referência x, onde para cada *token* x_i busca-se a correspondência mais semelhante no conjunto de *tokens* preditos $\hat{x}$. A fórmula que define essa avaliação é apresentada na Equação 4.9.

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{X}_j \in \hat{x}} \left(x_i^\top \hat{x}_j \right) \tag{4.9}$$

Por último, a medida F1, como já mencionado na Seção 2.3, que, diferente da média aritmética, dá mais peso aos valores menores. Isso significa que o F1 só será alto se tanto a precisão quanto o *recall* forem altas. A representação é definida na Equação 4.10.

$$F1 = 2 \times \frac{(P_{\text{BERT}} \times R_{\text{BERT}})}{(P_{\text{BERT}} + R_{\text{BERT}})} \tag{4.10}$$

4.5.6 CNN/Daily Mail

Em determinadas circunstâncias, como nos casos de TruthfulQA e ARC, onde o *dataset* constitui o próprio método de avaliação ou parte fundamental deste, não há necessidade de recorrer a dados externos, uma vez que o conjunto de dados em si já provê uma medida abrangente e robusta do desempenho do modelo. Contudo, em outras situações, a avaliação de modelos através de medidas como BERTScore e ROUGE exige a utilização de *datasets* externos, com o objetivo de testar a eficácia das técnicas de mitigação implementadas.

Para avaliar a qualidade das sumarizações geradas pelos modelos utilizando as medidas BERTScore e ROUGE, empregou-se o *dataset CNN/Daily Mail*. Este consiste em 312.000 artigos jornalísticos com sumarizações de referência humanas, divididos nos conjuntos (*subsets*) de treino (*train*), validação (*validation*) e teste (*test*). Deste último, composto por aproximadamente 11.500 itens, utilizou-se uma amostra de 1.000 exemplos, um tamanho consistente com práticas de avaliação em tarefas similares, e que é considerado robusto, permitindo que o Teorema do Limite Central assegure a estabilidade e a normalidade aproximada das médias das métricas calculadas, independentemente da distribuição das pontuações individuais [157, 158].

A medida ROUGE foi utilizada para quantificar a sobreposição entre as sumarizações automáticas e as de referência, enquanto o BERTScore foi aplicado para verificar a similaridade semântica, o que proporciona uma medida mais refinada da qualidade do significado transmitido.

4.6 Abordagem Estatística

Para que seja possível analisar os resultados, é fundamental esclarecer a abordagem estatística empregada para avaliar a performance dos métodos de mitigação em comparação ao modelo base (LLama-2 7b). Neste estudo, foram adotados testes estatísticos paramétricos, cuja aplicação é suportada pelas características dos dados e pelos tamanhos amostrais dos instrumentos de medição utilizados, conforme detalhado nas subseções seguintes e na Seção 4.5.6. Procedimentos diferentes foram utilizados para as avaliações de Q&A e para *benchmarks* de similaridade textual.

4.6.1 Abordagem Estatística Para Perguntas e Respostas

A performance dos métodos de mitigação foi avaliada individualmente para cada *benchmark*, tendo como referência o modelo base Llama-2 7b. A acurácia observada de um modelo em um determinado conjunto de avaliação, denotada por $\hat{p}$, foi calculada como a proporção de respostas corretas (x) em relação ao número total de questões (N) nesse teste ($\hat{p} = x/N$). Esta medida amostral $\hat{p}$ serve como uma estimativa da verdadeira, porém desconhecida, capacidade (p) do modelo para aquele tipo de tarefa.

O cerne da análise estatística consistiu em examinar a diferença entre a acurácia amostral de um método de mitigação ($\hat{p}_M$) e a do Llama-2 ($\hat{p}_B$) no mesmo *benchmark*. Para averiguar se esta disparidade observada, $\hat{p}_M - \hat{p}_B$, reflete uma distinção estatisticamente significativa entre as capacidades reais dos modelos ($p_M - p_B$), construiu-se um intervalo de confiança (IC) de 95%. Este intervalo delimita uma faixa de valores plausíveis para a genuína diferença $p_M - p_B$. A fórmula para o cálculo deste intervalo, fundamentada na aproximação normal para a diferença de duas proporções independentes, é apresentada na Equação 4.11.

$$\text{IC}_{95\%}(p_M - p_B) = (\hat{p}_M - \hat{p}_B) \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_M(1-\hat{p}_M)}{N_M} + \frac{\hat{p}_B(1-\hat{p}_B)}{N_B}} \tag{4.11}$$

Onde:

- $(\hat{p}_M - \hat{p}_B)$ representa a diferença calculada a partir das acurácias amostrais;
- $\hat{p}_M$ e $\hat{p}_B$ são as acurácias observadas (estimativas) do método de mitigação M e do modelo base B, respectivamente;
- N_M e N_B indicam o número total de questões (tamanho da amostra) no *benchmark* para o método M e para o modelo base B. No presente estudo, $N_M = N_B = N$ para cada avaliação;
- $z_{\alpha/2}$ é o valor crítico da distribuição normal padrão. Para um IC de 95%, com $\alpha = 0{,}05$, este valor é aproximadamente $1{,}96$.

A significância estatística (correspondente a um valor-p inferior a 0,05) foi inferida ao verificar se o intervalo de confiança de 95% para a verdadeira diferença $p_M - p_B$, obtido pela Equação 4.11, continha o valor zero. Se o referido intervalo não incluía o zero, a disparidade observada era considerada estatisticamente significativa, indicando que a diferença de desempenho entre os modelos provavelmente não é resultado do acaso.

A magnitude do efeito é indicada pela diferença observada nas acurácias amostrais $(\hat{p}_M - \hat{p}_B)$. O intervalo de confiança, por sua vez, complementa essa informação ao quantificar a precisão da estimativa da verdadeira diferença $p_M - p_B$.

4.6.2 Abordagem Estatística para medidas de Similaridade Textual

Para medidas como ROUGE (e suas variantes) e BERTScore, que atribuem uma pontuação contínua a cada par de texto gerado e referência (tipicamente por documento), a análise estatística da diferença de performance entre um método de mitigação e o modelo base utilizou-se do teste t pareado. Esta abordagem é apropriada pois cada documento de entrada resulta em uma pontuação da medida para ambos os modelos, formando pares de observações. Cada medida foi analisada individualmente.

O teste t pareado avalia se a diferença média nas pontuações entre os dois modelos é estatisticamente diferente de zero. Para um conjunto de n documentos comuns, calcula-se a diferença d_i entre a pontuação do método de mitigação (x_{M_i}) e a pontuação do modelo base (x_{B_i}) para cada i-ésimo documento. A estatística do teste t é então calculada como:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{4.12}$$

Onde:

- $\bar{d}$ é a média amostral das diferenças $d_i = x_{M_i} - x_{B_i}$;
- s_d é o desvio padrão amostral dessas diferenças d_i;
- n é o número de pares de observações (documentos) válidos, após alinhamento dos resultados e remoção de dados ausentes.

Esta estatística t segue uma distribuição t de *Student* com $n-1$ graus de liberdade. A significância estatística da diferença média é determinada pelo valor-p associado; um valor-p inferior a 0,05 é considerado indicativo de uma diferença estatisticamente significativa.

É importante notar que, especialmente com amostras grandes (frequentemente $n > 1000$ para avaliações por documento), valores p muito pequenos podem ser obtidos mesmo para diferenças médias $\bar{d}$ que possuem pouca relevância prática. Por esta razão, a análise de significância foi complementada pela estimativa da magnitude do efeito utilizando o d de Cohen para amostras pareadas:

$$d_{\text{Cohen}} = \frac{\bar{d}}{s_d} \tag{4.13}$$

Esta medida padroniza a diferença média, permitindo uma interpretação da magnitude do efeito que é independente da significância estatística baseada no valor-p. Para a interpretação do valor absoluto de d_{Cohen} ($|d|$), são adotados os seguintes limiares: $|d| < 0,01$ como efeito insignificante; $0,01 \leq |d| < 0,2$ como muito pequeno; $0,2 \leq |d| < 0,5$ como pequeno; $0,5 \leq |d| < 0,8$ como médio; e $|d| \geq 0,8$ como grande.

Adicionalmente, para avaliar a precisão da estimativa da diferença média $\bar{d}$, foi calculado um intervalo de confiança (IC) de 95%. De forma análoga à interpretação do IC na seção anterior (ver Equação 4.11), se o IC de 95% para a diferença média das pontuações não incluir zero, isso também corrobora a significância estatística. A combinação da diferença média observada $\bar{d}$, seu IC de 95%, o valor-p e o d de Cohen proporciona uma avaliação abrangente do desempenho relativo dos métodos para estas medidas de similaridade textual.

A título de ilustração, suponha a comparação do "Modelo de Mitigação Alfa" com o "Modelo Base" na medida BERTScore. Após o alinhamento dos dados e remoção de valores ausentes, foram obtidos $n = 4$ pares de observações. As diferenças d_i (pontuação do Modelo Alfa menos pontuação do Modelo Base) para estes pares foram: $0,03, 0,01, 0,03,$ e $0,01$.

Com base nessas diferenças:

1. Calcula-se a média das diferenças: $\bar{d} = 0,02$;
2. Calcula-se o desvio padrão amostral das diferenças: $s_d \approx 0,0115$;
3. Aplicando o teste t pareado (Equação 4.12), obtém-se:
   - Estatística $t \approx 3,46$ com $df = n - 1 = 3$ graus de liberdade;
   - Consultando a distribuição t o valor-p associado é $\approx 0,041$.
4. O d de Cohen (Equação 4.13) é $\frac{0,02}{0,0115} \approx 1,73$;
5. O intervalo de confiança de 95% para a diferença média $\bar{d}$ é calculado como $[0,0016, 0,0384]$.

Neste cenário, a diferença média de $0,02$ em favor do "Modelo Alfa" é estatisticamente significativa (valor-p $\approx 0,041$, $p < 0,05$). Esta conclusão é sustentada pelo intervalo de confiança de 95% para a referida diferença, $[0,0016, 0,0384]$, que, por não incluir o valor zero, estima que a superioridade do "Modelo Alfa" situa-se, em média, entre $0,0016$ e $0,0384$ pontos de BERTScore. Adicionalmente, o d de Cohen

(≈ 1.73) aponta para um efeito de grande magnitude, sublinhando a relevância prática destes achados. Assim, a análise conjunta indica um desempenho superior e de importância prática do "Modelo Alfa" sobre o "Modelo Base" para esta métrica e conjunto de dados.

4.7 Considerações Finais

A metodologia descrita neste capítulo é estruturada para prover uma análise sistemática e reprodutível dos métodos de mitigação de alucinações. A aplicação de diretrizes consolidadas da revisão sistemática da literatura por Kitchenham e Charters, juntamente com critérios de seleção rigorosos, incluindo disponibilidade de código, documentação e comparabilidade sobre o modelo base (LLama-2), resultou na identificação de quatro técnicas representativas para avaliação aprofundada.

Este arcabouço metodológico também compreendeu a definição de um conjunto diversificado de *benchmarks* e medidas de avaliação, além de uma abordagem estatística para a análise dos dados. Com isso, estabelecem-se as bases para a aferição quantitativa e a discussão dos resultados que serão apresentados no capítulo seguinte, permitindo uma investigação fundamentada sobre a eficácia das estratégias de mitigação selecionadas.

O código utilizado para a análise estatística deste capítulo e os respectivos arquivos CSV com os resultados detalhados estão publicamente disponíveis no repositório: Repositório de Código e Resultados da Análise Estatística. O arquivo `README.md` no local oferece um guia sobre o conteúdo e a função de cada arquivo. Por sua vez, os *notebooks* Jupyter contêm as instruções para executar os *scripts*.

CAPÍTULO

5

Resultados

Conforme delineado no Capítulo 4, a etapa subsequente refere-se à aferição prática das abordagens de mitigação de alucinações utilizando os *benchmarks* e medidas previamente estabelecidos. Este capítulo apresenta os achados decorrentes dessa análise, cujo propósito central é fornecer dados concretos que permitam cotejar a performance. As informações detalhadas a seguir atendem diretamente à necessidade de um exame quantitativo e comparativo, oferecendo a base para as conclusões sobre as potencialidades e limitações de cada alternativa no contexto da redução de alucinações em LLMs.

5.1 Ambiente de Execução

Os resultados apresentados na Seção 5.2.1, Seção 5.2.2 e Seção 5.2.3 são inteiramente executados no ambiente Google Colab. A escolha se dá devido aos planos com custo-benefício que permitem o uso da unidade de processamento gráfico NVIDIA A100, necessária por sua alta capacidade de memória VRAM para executar o modelo Llama2 7b.

Com o objetivo de padronizar os testes, é utilizado o recurso `pipeline` da biblioteca *Hugging Face Transformers*, não apenas por oferecer uma interface de alto nível que abstrai a complexidade da inferência do modelo, mas também

por automatizar as etapas de “tokenização”, inferência e decodificação, e por ser o formato preferencial de interação para um dos métodos avaliados (mais especificamente, a técnica RepE).

5.2 Resultado das Avaliações e Benchmarks

Nessa seção, os resultados das avaliações e benchmarks são apresentados e discutidos. A abordagem estatística paramétrica, delineada na Seção 4.6, é aplicada para analisar os dados. As tabelas com os resultados brutos e as análises comparativas são introduzidas em suas respectivas subseções.

Em virtude das restrições inerentes a estudos desta natureza, notadamente as limitações de tempo e recursos financeiros tipicamente encontradas em pesquisas acadêmicas (conforme detalhado na Seção 6.1 sobre os desafios específicos enfrentados), a avaliação comparativa do desempenho dos métodos é realizada por meio de uma única execução de cada avaliação e *benchmark*. Para assegurar a validade e replicabilidade dos resultados, mesmo com esta abordagem de execução única, é empregada a configuração de geração `do_sample=false` da biblioteca *transformers* do Hugging Face. Esta configuração garante a utilização da decodificação gulosa, um processo determinístico que assegura a obtenção de *outputs* idênticos em execuções repetidas, conferindo robustez e comparabilidade aos resultados apresentados, apesar da amostragem limitada.

5.2.1 Perguntas e Respostas

A análise da capacidade dos modelos em fornecer respostas corretas e factuais revela um desempenho heterogêneo entre os métodos de mitigação quando comparados à base. No conjunto TruthfulQA MC1, caracterizado por apenas uma resposta correta, os métodos KCA e WikiChat demonstraram uma melhoria com confiança estatística. Em contrapartida, os ganhos observados para ICD e RepE neste mesmo *benchmark* não alcançaram o mesmo nível de certeza estatística.

Já no TruthfulQA MC2, que permite múltiplas respostas corretas, todas as quatro estratégias avaliadas superaram o modelo base de forma expressiva. Os intervalos de confiança para a diferença de acurácia indicaram um avanço claro para todos os métodos neste cenário.

Os valores de acurácia brutos para cada método e avaliação podem ser con-

sultados na Tabela 5.1. As diferenças de acurácia em relação à base (representadas pelo símbolo Δ), juntamente com seus intervalos de confiança de 95% e a significância estatística, estão detalhadas na Tabela 5.2, fornecendo a base quantitativa para essas observações.

Tabela 5.1: Acurácia dos Métodos

LLM	TruthfulQA MC1	TruthfulQA MC2	ARC Easy	ARC Challenge	OAB
LLama-2	0,1468	0,2447	0,3144	0,3298	0,2163
KCA	**0,3904**	0,5226	**0,6088**	**0,4047**	**0,2964**
ICD	0,1774	0,4785	0,5632	0,4013	0,2525
WikiChat	0,2692	**0,5299**	0,5895	0,4013	0,2738
RepE	0,1713	0,3121	0,3526	0,3177	0,2371

Nota: LLama-2 mostrado como linha de base de referência (fundo cinza).

Tabela 5.2: Diferenças de Acurácia vs Llama-2 Base

	KCA		ICD		WikiChat		RepE	
Benchmark	Δ	IC 95%	Δ	IC 95%	Δ	IC 95%	Δ	IC 95%
TruthfulQA MC1	**0,2436**	[0,2022, 0,2849]	0,0306	[–0,0051, 0,0663]	**0,1224**	[0,0835, 0,1613]	0,0245	[–0,0110, 0,0599]
TruthfulQA MC2	**0,2779**	[0,2327, 0,3230]	**0,2338**	[0,1886, 0,2790]	**0,2852**	[0,2400, 0,3304]	**0,0674**	[0,0240, 0,1107]
ARC Easy	**0,2944**	[0,2761, 0,3127]	**0,2488**	[0,2303, 0,2673]	**0,2751**	[0,2568, 0,2935]	**0,0382**	[0,0200, 0,0562]
ARC Challenge	**0,0749**	[0,0487, 0,1011]	**0,0715**	[0,0453, 0,0976]	**0,0715**	[0,0453, 0,0976]	–0,0121	[–0,0375, 0,0135]
OAB	**0,0801**	[0,0545, 0,1057]	**0,0362**	[0,0112, 0,0612]	**0,0575**	[0,0322, 0,0828]	0,0208	[–0,0039, 0,0455]

Nota: Valores de Δ em negrito são estatisticamente significativos ($p < 0,05$).

Em relação aos dois *benchmarks* ARC, a magnitude da diferença de acurácia e seus intervalos de confiança oferecem uma perspectiva valiosa. No ARC Easy, todas as abordagens, inclusive RepE, ultrapassam a performance do Llama-2, com os dados corroborando uma melhoria substancial; KCA, por exemplo, evidencia um incremento em acurácia que, com 95% de confiança, situa-se entre aproximadamente 28% e 31% (baseado no IC: $[0,2761; 0,3127]$). No ARC Challenge, esses avanços, embora ainda confirmados estatisticamente para KCA, ICD e WikiChat, são mais comedidos, com ganhos de acurácia em torno de 7% a 7,5%. Neste cenário, ICD e WikiChat apresentam taxas de acerto virtualmente idênticas. RepE, por sua vez, não se distingue do modelo base em termos estatísticos, ficando ligeiramente aquém.

No conjunto OAB, que explora características multilíngues, as abordagens KCA, ICD e WikiChat demonstram um ganho sobre o Llama-2, ainda que os aumentos de acurácia sejam mais discretos, variando de aproximadamente 3,6% (ICD) a 8% (KCA). Para RepE, a diferença observada neste benchmark não se mostrou estatisticamente robusta.

Considerando o panorama completo das avaliações, KCA e WikiChat mantêm um desempenho consistentemente superior, ultrapassando o Llama-2 em todos os benchmarks com um grau de confiança estatística elevado. ICD também apresenta avanços na maioria dos cenários, à exceção do TruthfulQA MC1, onde a melhoria não foi conclusiva. Em contrapartida, RepE demonstra um perfil de desempenho mais irregular: enquanto em TruthfulQA MC2 e ARC Easy seus resultados são notavelmente superiores ao Llama-2, nos demais benchmarks não se observam incrementos consideráveis ou, como no ARC Challenge, sua performance fica ligeiramente abaixo da referência, sem que essa diferença seja decisiva.

Antecipava-se que a incorporação de conhecimento externo via mecanismos de recuperação de informação, como no WikiChat, conferiria uma vantagem de performance proeminente sobre outras classes de métodos de mitigação em todas as avaliações. Contudo, a análise quantitativa revela a notável solidez da estratégia KCA, baseada primariamente em técnicas de *fine-tuning*, que também se mostrou consistentemente eficaz em todos os cenários. Essa observação sugere que, embora a recuperação explícita de conhecimento seja potente, um regime de ajustes finos especializado pode ser igualmente ou mais efetivo para as medidas avaliadas. De fato, a performance de WikiChat e KCA foi altamente competitiva, com WikiChat obtendo uma ligeira vantagem no subconjunto de múltipla escolha TruthfulQA MC2, onde a abordagem de recuperação externa efetivamente prevaleceu marginalmente sobre o *fine-tuning* do KCA.

5.2.2 Correspondência ao Nível de Tokens

As medidas ROUGE (Tabela 5.3) e BERTScore (Tabela 5.5), utilizadas nesta seção e na subsequente (Seção 5.2.3), são calculadas para cada item individual do conjunto de avaliação. Isso significa que, se o conjunto de avaliação contém N itens (por exemplo, N sentenças resumidas a serem comparadas com suas respectivas referências), cada medida fornecerá N pontuações distintas.

Os resultados consolidados para as diversas variantes da medida ROUGE, apresentando as pontuações agregadas (Média, Mediana e Média Harmônica) para cada método, estão compilados na Tabela 5.3. Para uma análise comparativa focada nas melhorias em relação ao modelo base, a Tabela 5.4 detalha as diferenças, o tamanho do efeito e os intervalos de confiança.

Tabela 5.3: ROUGE Benchmark

LLM	**ROUGE-1**			**ROUGE-2**			**ROUGE-L**			**ROUGE-LSum**		
	Média	**Mediana**	**MédiaH**	**Média**	**Mediana**	**MédiaH**	**Média**	**Mediana**	**MédiaH**	**Média**	**Mediana**	**MédiaH**
LLama-2	0,2184	0,2086	0,1625	0,0719	0,0612	0,0370	0,1586	0,1485	0,1247	0,1901	0,1777	0,1452
KCA	0,2774	0,2745	0,2631	0,1050	0,0898	0,0428	0,1972	0,1885	0,1565	0,2360	0,2174	0,2003
ICD	0,2986	0,2872	0,2519	0,0979	0,0896	0,0639	0,2038	0,1939	0,1660	0,2400	0,2288	0,2003
WikiChat	**0,3134**	**0,3076**	**0,2780**	**0,1206**	**0,1091**	**0,0754**	**0,2128**	**0,2033**	**0,1723**	**0,2562**	**0,2462**	**0,2154**
RepE	0,2658	0,2593	0,2387	0,0952	0,0866	0,0708	0,1810	0,1771	0,1585	0,2168	0,2128	0,1945

Tabela 5.4: Diferenças de Rouge vs Llama-2 Base

	KCA			**ICD**			**WikiChat**			**RepE**		
medida	Δ	*d* (P/M/G)	IC 95%	Δ	*d* (P/M/G)	IC 95%	Δ	*d* (P/M/G)	IC 95%	Δ	*d* (P/M/G)	IC 95%
rouge1	0,059	0,489 (P)	[0,0516, 0,0665]	0,080	0,535 (M)	[0,0709, 0,0895]	**0,095**	0,652 (M)	[0,0860, 0,1041]	0,047	0,361 (P)	[0,0392, 0,0555]
rouge2	0,033	0,383 (P)	[0,0277, 0,0385]	0,026	0,355 (P)	[0,0215, 0,0306]	**0,049**	0,567 (M)	[0,0434, 0,0541]	0,023	0,344 (P)	[0,0191, 0,0275]
rougeL	0,039	0,350 (P)	[0,0318, 0,0455]	0,045	0,414 (P)	[0,0384, 0,0520]	**0,054**	0,477 (P)	[0,0472, 0,0613]	0,022	0,238 (P)	[0,0166, 0,0282]
rougeLsum	0,047	0,366 (P)	[0,0390, 0,0548]	0,050	0,398 (P)	[0,0421, 0,0577]	**0,066**	0,501 (M)	[0,0579, 0,0743]	0,027	0,240 (P)	[0,0198, 0,0337]

Nota: d (P/M/G) categoriza d de Cohen como efeito pequeno <0,5, impacto médio 0,5–0,8, impacto grande >= 0,8.

Em consonância com a metodologia empregada nas Subseções 5.2.1 e 5.2.3, a presente seção se vale de testes estatísticos para a avaliação da significância dos resultados obtidos. Para os *benchmarks* dessa seção, utiliza-se também teste t para amostras pareadas complementada pelo cálculo do d de Cohen e intervalo de confiança de 95%.

Similarmente aos testes anteriores, a técnica RepE apresenta o desempenho menos expressivo entre as abordadas. Entretanto, de forma distinta dos resultados precedentes, observa-se uma melhora estatisticamente significante (embora com magnitude de efeito pequena, $d < 0,2$) em todas as medidas ROUGE. Essa melhora, inclusive, permite que o RepE atinja um desempenho comparável ao de técnicas que demonstram resultados mais robustos na avaliação anterior.

De maneira notável, o modelo WikiChat, que na avaliação da Seção 5.2.1, embora tenha apresentado resultados consistentemente superiores à base, não

se estabeleceu como uniformemente dominante sobre todas as outras abordagens de mitigação, como se poderia antecipar, dada sua arquitetura focada em recuperação de informação.

Confirmando a tendência contraintuitiva observada anteriormente, o WikiChat apresenta desempenho superior em todas as medidas ROUGE deste trabalho, em que a explicação mais plausível para este resultado reside na sinergia entre a arquitetura da técnica e a própria natureza da avaliação ROUGE. Pois ao recuperar explicitamente trechos de informação relevantes de sua base de conhecimento antes da geração, o WikiChat incorpora material que frequentemente compartilha uma sobreposição lexical significativa com os textos de referência usados no *benchmark*.

Pode-se concluir que medidas como ROUGE ou similares (como BLEU) recompensam diretamente essa sobreposição literal de n-gramas e subsequências, estratégias de recuperação acabam favorecendo pontuações altas. Em contrapartida, técnicas de otimização de modelo ou aprimoramento de inferência podem gerar resumos mais abstratos, fluentes ou semanticamente ricos, mas que, ao parafrasear ou reestruturar a informação, tendem a divergir lexicalmente das referências.

5.2.3 Correspondência Semântica Contextual

Seguindo a estrutura de apresentação adotada na seção anterior, as pontuações agregadas para o BERTScore são apresentadas na Tabela 5.5. Da mesma forma, a Tabela 5.6 detalha as diferenças em relação ao modelo base.

Tabela 5.5: BERTScore Benchmark

LLM	**BERTScore Precision**			**BERTScore Recall**			**BERTScore F1**		
	Média	**Mediana**	**MédiaH**	**Média**	**Mediana**	**MédiaH**	**Média**	**Mediana**	**MédiaH**
LLama-2	0,8374	0,8423	0,8323	0,8535	0,8539	0,8531	0,8453	0,8466	0,8440
KCA	0,8667	0,8675	0,8663	0,8619	0,8622	0,8616	0,8636	0,8641	0,8634
ICD	**0,8717**	**0,8722**	**0,8709**	0,8679	0,8684	0,8676	0,8694	0,8699	0,8691
WikiChat	0,8667	0,8673	0,8665	0,8750	0,8751	0,8747	**0,8700**	**0,8704**	**0,8697**
RepE	0,8521	0,8524	0,8518	**0,8768**	**0,8776**	**0,8766**	0,8644	0,8646	0,8641

Tabela 5.6: Diferenças de BERTScore vs Llama-2 Base

	KCA			ICD			WikiChat			RepE		
medida	Δ	*d* (P/M/G)	IC 95%	Δ	*d* (P/M/G)	IC 95%	Δ	*d* (P/M/G)	IC 95%	Δ	*d* (P/M/G)	IC 95%
F1	0,018	0,498 (M)	[0,0160, 0,0206]	0,024	0,662 (M)	[0,0218, 0,0263]	**0,025**	0,673 (M)	[0,0224, 0,0269]	0,019	0,533 (M)	[0,0169, 0,0213]
Precision	0,029	0,447 (P)	[0,0252, 0,0334]	**0,034**	0,507 (M)	[0,0301, 0,0385]	0,029	0,456 (P)	[0,0254, 0,0333]	0,015	0,225 (P)	[0,0107, 0,0188]
Recall	0,008	0,354 (P)	[0,0070, 0,0099]	0,014	0,609 (M)	[0,0130, 0,0159]	0,022	0,906 (G)	[0,0200, 0,0230]	**0,023**	1.006 (G)	[0,0219, 0,0248]

Ao aplicar os mesmos critérios de avaliação de significância estatística descritos na Seção 5.2.2 (teste t, d de Cohen e intervalo de confiança de 95%), observa-se que a técnica RepE, em termos de precisão medida pelo BertScore, retorna ao domínio da insignificância estatística. Contudo, as medidas de *recall* e F1 apresentam melhoras modestas, porém estatisticamente significativas. Notavelmente, esta é a única instância, dentre as três subseções analisadas, em que a técnica RepE não figura como a de menor desempenho. Nesta avaliação, o modelo KCA, que se destacou expressivamente nas avaliações de perguntas e respostas (Seção 5.2.1), apresenta a performance menos robusta.

O desempenho superior do ICD em *precision* pode ser explicado por seu mecanismo de decodificação contrastiva. Ao penalizar ativamente as probabilidades de unidades lexicais associadas a respostas não factuais (induzidas pelo modelo "fraco"), o ICD refina a seleção de cada termo gerado. Esse processo de filtragem probabilística aumenta a chance de que os elementos na saída final possuam alta similaridade semântica com os da fonte de comparação, alinhando-se diretamente com a avaliação de precisão fornecida pelo *BERTScore*.

Por sua vez, o resultado elevado do RepE em *recall* sugere que a manipulação direcionada das representações internas do modelo promove eficazmente a inclusão do conteúdo essencial da fonte. Ao ajustar as ativações da rede para alinhar-se com um vetor conceitual (como "honestidade"), o RepE estimula um estado interno favorável à geração de informações relevantes, o que pode levar a uma cobertura mais abrangente dos elementos do texto base, mesmo sem uma triagem explícita em nível de *token*.

A maior pontuação do F1 é alcançada pelo WikiChat, o que parece ser uma consequência do equilíbrio em sua arquitetura multifacetada, em que a fase de recuperação inicial fornece material factual pertinente, potencialmente ampliando a cobertura do texto de referência, enquanto a etapa subsequente de verificação atua como um filtro. Essa combinação resulta em um desempenho balanceado, que é bem capturado pela média harmônica desta medida.

Finalmente, o KCA, apesar de seu destaque nas avaliações de perguntas e respostas (Seção 5.2.1), apresentou aqui seu desempenho mais discreto. Isso sugere que seu processo de *fine-tuning*, focado na otimização de respostas factuais e concisas, pode não ter priorizado igualmente a riqueza semântica ou a diversidade parafrástica que o BERTScore valoriza ao comparar com os textos de referência. Tal otimização pode resultar em um estilo de resposta que, embora correto, diverge lexical e estruturalmente dos textos de referência mais do que métodos que se beneficiam de recuperação explícita de informação ou manipulação de representações internas para maior cobertura.

CAPÍTULO

6

Conclusão

Este trabalho que combina revisão sistemática da literatura com exames quantitativos analisou estratégias de mitigação de alucinações em modelos de linguagem de grande escala, buscando responder a quatro questões de pesquisa centrais sobre os métodos disponíveis entre 2021 e o primeiro semestre de 2025, sua classificação, eficácia comparativa e desafios de reprodutibilidade.

Em resposta às duas primeiras questões, identificou-se um panorama das abordagens de mitigação recentes, classificando-as em quatro categorias principais: aprimoramento de inferência e saída, conhecimento externo e interação, otimização de modelo e métodos experimentais. Notou-se uma predominância das duas primeiras categorias (66% do total), sugerindo um foco em técnicas aplicáveis de forma modular ou que evitam o retreinamento extensivo do modelo base.

Abordando a terceira questão de pesquisa, a avaliação comparativa da eficácia indicou que a otimização de modelo apresentou desempenho superior em tarefas de perguntas e respostas, enquanto a abordagem de conhecimento externo e interação se destacou em *benchmarks* de correspondência em nível de *token*. Os *benchmarks* de correspondência semântica contextual, contudo, mostraram resultados heterogêneos, apontando que a adequação de uma categoria depende da medida e da tarefa específica.

Os aspectos levantados pela quarta questão de pesquisa, relativos aos desafios de reprodução dos métodos que incluem a análise da disponibilidade de código, presença de documentação e as dificuldades encontradas na realização das avaliações quantitativas deste trabalho serão discutidos em detalhe na Seção 6.1.

Em seguida, as principais contribuições deste estudo serão consolidadas na Seção 6.2. Por fim, a Seção 6.3 apresenta sugestões para trabalhos futuros que visem aprofundar o entendimento dos resultados e superar os desafios identificados.

6.1 Dificuldades

A condução da fase experimental enfrentou alguns desafios técnicos e de infraestrutura significativos. A principal limitação foi a disponibilidade restrita e imprevisível de GPUs NVIDIA A100 na plataforma Google Colab. Esse gargalo de recursos computacionais exigiu que a execução dos experimentos fosse concentrada em períodos de baixa demanda global, o que, por sua vez, gerou atrasos e impactou o cronograma geral.

Adicionalmente, surgiram dificuldades relacionadas às ferramentas e à replicação de métodos específicos. O uso da abstração `pipeline` (parte da biblioteca *Hugging Face*), embora conveniente, resultou em erros de falta de memória (`OutOfMemory`), em que a contramedida necessária foi a redução do tamanho dos lotes (*batch size*) para adequar o uso de memória à capacidade da GPU, uma solução comum, porém que pode impactar o tempo total de processamento.

Problemas de compatibilidade de dependências também emergiram durante a configuração do ambiente para certos métodos. Notavelmente, para a técnica RepE (representante da categoria de métodos experimentais), foi preciso realizar um *downgrade* da biblioteca `accelerate` para a versão 1.0.1, a fim de alinhar o ambiente com as especificações do trabalho original e garantir a funcionalidade da implementação disponibilizada.

Outra complexidade adicional se deve à avaliação das tarefas de Perguntas e Respostas (Q&A) no pós-processamento, dado que as respostas geradas pelos modelos e pelas diferentes técnicas de mitigação nem sempre aderiam a um formato estrito (como, por exemplo, apenas a letra da alternativa), foi necessário desenvolver e aplicar rotinas de sanitização e normalização para extrair a

resposta canônica antes do cálculo das medidas de acurácia, o que demandou esforço adicional de implementação e validação.

Finalmente, o número de técnicas incluídas na comparação justa foi significativamente menor do que o esperado, conforme os critérios de exclusão detalhados na Seção 4.4.2. Essa limitação deve-se a dois fatores principais: a exigência de garantir a comparabilidade entre os métodos (por exemplo, utilizando o mesmo modelo base) e a indisponibilidade de algumas técnicas para teste prático. É crucial ressaltar que, embora a escolha de um modelo LLM base comum (Llama2 7b) tenha sido fundamental para a validade da comparação direta entre as categorias de mitigação, essa restrição implica que os quatro métodos selecionados, apesar de representativos, podem não configurar o "estado da arte" absoluto em suas respectivas categorias.

6.2 Contribuições

Como principal contribuição, este trabalho propõe uma taxonomia que organiza os métodos contemporâneos de mitigação de alucinações em LLMs em quatro categorias distintas. Essa classificação foi desenvolvida a partir de uma revisão sistemática de 93 estudos. Essa estruturação e categorização não apenas permitiram a identificação de certas tendências na área, incluindo picos de interesse e preferências por determinadas abordagens, como também se mostrou fundamental para subsidiar a avaliação quantitativa e para visualizar as dinâmicas e focos da pesquisa no campo.

A análise comparativa, que utilizou um conjunto diversificado de *benchmarks*, forneceu valiosos *insights*. Ela não apenas destacou as categorias com desempenho superior em cenários específicos, como também demonstrou que as tendências de pesquisa nem sempre se traduzem diretamente na melhor abordagem em termos de eficácia pura. Em vez disso, sugere-se que tais tendências podem, muitas vezes, refletir uma otimização de custo-benefício, privilegiando métodos que requerem menor investimento em retreinamento extensivo do modelo base, são mais facilmente integráveis de forma modular ou demandam menos recursos computacionais para implementação e execução.

Para além dessa perspectiva sobre as dinâmicas de pesquisa, a própria análise comparativa elucidou de forma crucial o desempenho diferencial das abor-

dagens. Observou-se, por exemplo, a superioridade da otimização de modelo em perguntas e respostas e a eficácia do conhecimento externo e interação em correspondência lexical. Tais achados fornecem evidências sobre as respectivas potencialidades e restrições de cada técnica. Esta clareza é relevante pois informa a seleção de métodos para aplicações específicas, serve como ponto de referência e auxilia na identificação de lacunas, como o desempenho em contextos multilíngues.

Em suma, ao combinar a identificação de tendências de pesquisa com uma avaliação empírica detalhada do desempenho das técnicas de mitigação, este trabalho oferece uma compreensão mais nuançada do campo. As contribuições aqui apresentadas não apenas mapeiam o cenário atual, mas também fornecem subsídios práticos e direcionamentos para futuras investigações e aplicações voltadas à redução de alucinações em modelos de linguagem de grande escala.

6.3 Trabalhos Futuros

Os resultados e análises apresentados neste trabalho abrem caminhos prósperos para investigações futuras na mitigação de alucinações em LLMs. Sugere-se as seguintes direções:

- Investigar a eficácia da otimização de modelo versus conhecimento externo em Q&A devido a inesperada colocação do KCA (otimização via *fine-tuning*) sobre o WikiChat (recuperação externa). Estudos futuros podem analisar como o fine-tuning pode permitir ao modelo internalizar ou acessar conhecimento factual de forma mais eficaz para essas tarefas do que a recuperação explícita;

- Estender a análise para modelos multimodais (MLLMs) devido a crescente proeminência de modelos estado-da-arte que integram informações de diferentes fontes como texto e imagem. Trabalhos futuros poderiam avaliar a aplicabilidade e a eficácia das categorias de mitigação identificadas neste estudo (otimização, conhecimento externo, etc) no contexto multimodal, o que pode exigir a adaptação de técnicas existentes ou o desenvolvimento de novas estratégias e *benchmarks* específicos para avaliar a consistência e a veracidade;

- Validação com múltiplas execuções que devido a limitação de recursos (conforme destacado na Seção 6.1), a replicação dos experimentos com múltiplas *random seeds*[1] seria valiosa para confirmar a robustez estatística dos achados e quantificar a variância dos resultados;
- Desenvolvimento de mitigação em contextos multilinguísticos dada a fragilidade observada na avaliação da prova da OAB, o que destaca uma lacuna significativa. Torna-se crucial direcionar esforços para a criação ou adaptação de estratégias especificamente projetadas para LLMs que operam com múltiplos idiomas.

[1] Valores que controlam a aleatoriedade em etapas como inicialização de parâmetros do modelo ou embaralhamento de dados.

Referências Bibliográficas

[1] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. vi, 23

[2] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *arXiv*, 2022. vi, 41

[3] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv*, 2020. 1

[4] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv*, 2020. 1

[5] Immanuel Trummer. Codexdb: Generating code for processing sql queries using gpt-3 codex. *arXiv*, 2022. 1

[6] Aishwarya Narasimhan, Krishna Prasad Agara Venkatesha Rao, and Veena M B. Cgems: A metric model for automatic code generation using gpt-3. *arXiv*, 2021. 1

[7] Jonas Thiergart, Stefan Huber, and Thomas Übellacker. Understanding emails and drafting responses – an approach using gpt-3. *arXiv*, 2021. 1

[8] Simon Boman. Improving customer support efficiency through decision support powered by machine learning. 2023. 1

[9] Shruti Saravanan and K. Sudha. Gpt-3 powered system for content generation and transformation. In *2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT)*, pages 514–519, 2022. 1

[10] Raul Salles de Padua, Imran Qureshi, and Mustafa U. Karakaplan. Gpt-3 models are few-shot financial reasoners. *arXiv*, 2023. 1

[11] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023. 1

[12] Erica Koranteng, Arya Rao, Efren J Flores, Michael H. Lev, Adam Landman, Keith Dreyer, and Marc D. Succi. Empathy and equity: Key considerations for large language model adoption in health care. *JMIR Medical Education*, 9, 2023. 1

[13] Himani R Naik, Andrew D. Prather, and Grzegorz T. Gurda. Synchronous bilateral breast cancer: A case report piloting and evaluating the implementation of the ai-powered large language model (llm) chatgpt. *Cureus*, 15, 2023. 1

[14] JULIAN VARAS, BRANDON VALENCIA CORONEL, IGNACIO VILLAGRÁN, GABRIEL ESCALONA, ROCIO HERNANDEZ, GREGORY SCHUIT, VALENTINA DURÁN, ANTONIA LAGOS-VILLASECA, CRISTIAN JARRY, ANDRES NEYEM, and et al. Innovations in surgical training: exploring the role of artificial intelligence and large language models (llm). *Revista do Colégio Brasileiro de Cirurgiões*, 50:e20233605, 2023. 1

[15] Chao Feng, Xinyu Zhang, and Zichu Fei. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *arXiv 2309.03118*, 2023. 2

[16] Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. Automatic hallucination assessment for aligned large language models via transferable adversarial attacks. *arXiv*, 2023. 2

[17] Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. *arXiv*, 2023. 2, 46

[18] Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. Instruction mining: When data mining meets large language model finetuning. *arXiv 2307.06290*, 2023. 2

[19] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 5

[20] Alana de Santana Correia and Esther Luna Colombini. Attention, please! a survey of neural attention models in deep learning, 2021. 5

[21] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation, 2015. 5

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv 1706.03762*, 2023. 5, 6, 17

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 5

[24] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. 6

[25] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J. Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. Technical report, Google, 2019. 6

[26] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. *Proceedings of the Annual ACM Conference on Computational Learning Theory*, 10 2000. 7

[27] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2012. 8

[28] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 8

[29] Vlad Riscutia. *Large Language Models at Work: Enhancing Software Systems with Language Models*. 2023. 8

[30] Thimira Amaratunga. *What Makes LLMs Large?*, pages 81–117. Apress, Berkeley, CA, 2023. 8

[31] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36, 2024. 8, 46

[32] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 8

[33] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8, 9, 50

[34] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *arXiv 2212.13138*, 2022. 9

[35] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019. 9

[36] Savas Yildirim. Fine-tuning transformer-based encoder for turkish language understanding tasks. *arXiv 2401.17396*, 2024. 9

[37] Oludare Isaac Abiodun, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. State-of-the-art in artificial neural network applications: A survey. *Heliyon*, 4(11), 2018. 9

[38] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*, 2023. 9, 10

[39] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023. 9

[40] Shih-Wen Chen and Hsien-Jung Hsu. articlealtral: Reducing numeric hallucinations of mistral with precision numeric calculation. 12 2023. 9

[41] Mauro Giuffrè, Kisung You, and Dennis L Shung. Evaluating chatgpt in medical contexts: The imperative to guard against hallucinations and partial accuracies. *Clinical Gastroenterology and Hepatology*, 2023. 10

[42] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023. 10

[43] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models, 2023. 10

[44] Sina J. Semnani, Violet Z. Yao, Heidi C. Zhang, and Monica S. Lam. Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia. *arXiv 2305.14292*, 2023. 10, 46, 49

[45] Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning. *arXiv 1806.02847*, 2019. 11

[46] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022. 11

[47] Jack Bandy and Nicholas Vincent. Addressing"documentation debt"in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021. 11

[48] Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W Cohen. Handling divergent reference texts when evaluating table-to-text generation. *arXiv preprint arXiv:1906.01081*, 2019. 11

[49] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. 11

[50] Hobson Lane, Cole Howard, and Hannes Hapke. *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Manning Publications, 2019. 12

[51] Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, Dinesh Manocha, et al. A closer look at the

limitations of instruction tuning. *arXiv preprint arXiv:2402.05119*, 2024. 12

[52] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. 12

[53] Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. *arXiv preprint arXiv:2004.12651*, 2020. 12

[54] Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*, 2021. 12, 19

[55] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv 2109.07958*, 2022. 12, 15, 19, 50

[56] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv 2401.11817*, 2024. 13

[57] Lior Gazit and Meysam Ghaffari. *Mastering NLP from Foundations to LLMs: Apply advanced rule-based techniques to LLMs and solve real-world business problems using Python*. Packt Publishing Ltd, 2024. 13

[58] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2019. 13, 14

[59] Pádraig Cunningham and Sarah Jane Delany. *Underestimation Bias and Underfitting in Machine Learning*, page 20–31. Springer International Publishing, 2021. 13

[60] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023. 15

[61] Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. Prudent silence or foolish babble? examining large language models' responses to the unknown. *arXiv preprint arXiv:2311.09731*, 2023. 15

[62] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv 2005.04118*, 2020. 15

[63] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 15

[64] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv 2104.08704*, 2022. 16, 20

[65] Nancy Chinchor. Muc-4 evaluation metrics. In *Proceedings of the 4th Conference on Message Understanding*, MUC4 '92, page 22–29, USA, 1992. Association for Computational Linguistics. 16

[66] Letong Zhou. Loan defaults prediction based on stacked models trained by personalized features. *Highlights in Business, Economics and Management*, 40:422–428, 09 2024. 16

[67] International Organization for Standardization. *ISO 5725-1: 1994: Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 1: General Principles and Definitions*. International Organization for Standardization, 1994. 16

[68] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv 1804.07461*, 2019. 16, 17

[69] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 17

[70] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 17, 26

[71] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*, 2018. 17

[72] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv 1904.09675*, 2020. 18, 54

[73] Michael Hanna and Ondřej Bojar. A fine-grained analysis of BERTScore. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online, November 2021. Association for Computational Linguistics. 18

[74] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *International conference on machine learning*, pages 1078–1088. PMLR, 2020. 19

[75] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms, 2024. 19

[76] Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. How much are large language models contaminated? a comprehensive survey and the llmsanitize library, 2024. 19

[77] Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. Detecting hallucinated content in conditional neural sequence generation. *arXiv 2011.02593*, 2021. 20

[78] Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. *arXiv preprint arXiv:2311.13230*, 2023. 20

[79] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023. 20, 22

[80] Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 21

[81] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. 22

[82] Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610, 2014. Evgeniy GabrilovichWilko HornNi LaoKevin MurphyThomas StrohmannShaohua SunWei ZhangGeremy Heitz. 22

[83] Gautam Kishore Shahi and Durgesh Nandini. *FakeCovid- A Multilingual Cross-domain Fact Check News Dataset for COVID-19*. ICWSM, Jun 2020. 22

[84] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. Where is your evidence: Improving fact-checking by justification modeling. In James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos,

and Arpit Mittal, editors, *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium, November 2018. Association for Computational Linguistics. 22

[85] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. 23, 32

[86] Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *arXiv preprint arXiv:2402.10612*, 2024. 23, 24

[87] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR, 2023. 23

[88] Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*, 2019. 23

[89] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018. 23

[90] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023. 24, 46

[91] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 24

[92] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks

from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 06 2014. 25

[93] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv 1802.05365*, 2018. 25

[94] Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. Detecting hallucinated content in conditional neural sequence generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online, August 2021. Association for Computational Linguistics. 26

[95] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019. 26, 35

[96] Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O Arnold, and Bing Xiang. Improving factual consistency of abstractive summarization via question answering. *arXiv preprint arXiv:2105.04623*, 2021. 27

[97] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. 27

[98] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv. org/abs/2203.02155*, 13, 2022. 27, 28

[99] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc., 2020. 27

[100] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. Inverse reward design. *arXiv 1711.02827*, 2020. 28

[101] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023. 28

[102] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. Nationality bias in text generation. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. 29

[103] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy, July 2019. Association for Computational Linguistics. 29

[104] Xudong Han, Timothy Baldwin, and Trevor Cohn. Balancing out bias: Achieving fairness through balanced training. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11335–11350, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. 29

[105] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, 2018. 29

[106] Yougang Lyu, Piji Li, Yechang Yang, M. de Rijke, Pengjie Ren, Yukun Zhao, Dawei Yin, and Zhaochun Ren. Feature-level debiased natural language understanding. In *AAAI Conference on Artificial Intelligence*, 2022. 30

[107] Prasetya Ajie Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. Avoiding inference heuristics in few-shot prompt-based finetuning. In *Conference on Empirical Methods in Natural Language Processing*, 2021. 30

[108] Deren Lei, Yaxi Li, Mingyu Wang, Vincent Yun, Emily Ching, Eslam Kamal, et al. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*, 2023. 30

[109] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang,

Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 31

[110] OpenAI. Openai o3-mini system card. Technical report, OpenAI, 2025. 31

[111] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 31

[112] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. 31

[113] Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *Advances in Neural Information Processing Systems*, 36, 2024. 31

[114] Shreyas Verma, Kien Tran, Yusuf Ali, and Guangyu Min. Reducing llm hallucinations using epistemic neural networks. *arXiv preprint arXiv:2312.15576*, 2023. 31, 46

[115] Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O Nunes, et al. Rag vs fine-tuning:

Pipelines, tradeoffs, and a case study on agriculture. *arXiv e-prints*, pages arXiv–2401, 2024. 32

[116] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021. 33

[117] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv 1905.01969*, 2020. 33

[118] Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. Alleviating hallucinations of large language models through induced hallucinations. *ArXiv*, abs/2312.15710, 2023. 34, 48

[119] BA Kitchenham and S Charters. *Guidelines for performing systematic literature reviews in software engineering*. EBSE Technical Report, 2007. KerkoCite.ItemAlsoKnownAs: 2129771:7RP54LK8 2129771:G45N5C6G 2405685:ETPE564Y 2486141:KVCIQUQU. 37

[120] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better alpaca with fewer data. *arXiv*, 2024. 46

[121] Fanqi Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. Knowledge verification to nip hallucination in the bud. *arXiv 2401.10768*, 2024. 46, 48

[122] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. Factuality enhanced language models for open-ended text generation. *arXiv 2206.04624*, 2023. 46

[123] Junyu Luo, Cao Xiao, and Fenglong Ma. Zero-resource hallucination prevention for large language models. *arXiv 2309.02654*, 2023. 46

[124] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv*, 2023. 46

[125] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf. *arXiv*, 2023. 46

[126] Lei Li, Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, and Hua Wu. Tool-augmented reward modeling. *arXiv*, 2024. 46

[127] Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *arXiv*, 2023. 46

[128] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv*, 2023. 46

[129] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv*, 2024. 46

[130] Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. *arXiv*, 2024. 46

[131] Wei-Lin Chen, Cheng-Kuang Wu, Hsin-Hsi Chen, and Chung-Chi Chen. Fidelity-enriched contrastive search: Reconciling the faithfulness-diversity trade-off in text generation. *arXiv*, 2023. 46

[132] Sehyun Choi, Tianqing Fang, Zhaowei Wang, and Yangqiu Song. Kcts: Knowledge-constrained tree search decoding with token-level hallucination detection. *arXiv*, 2023. 46

[133] Zhichao Xu. Context-aware decoding reduces hallucination in query-focused summarization. *arXiv*, 2023. 46

[134] Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. Alleviating hallucinations of large language models through induced hallucinations. *arXiv 2312.15710*, 2024. 46

[135] Yifan Feng, Hao Hu, Xingliang Hou, Shiquan Liu, Shihui Ying, Shaoyi Du, Han Hu, and Yue Gao. Hyper-rag: Combating llm hallucinations using hypergraph-driven retrieval-augmented generation, 2025. 46

[136] Yuqiao Tan, Shizhu He, Huanxuan Liao, Jun Zhao, and Kang Liu. Dynamic parametric retrieval augmented generation for test-time knowledge enhancement, 2025. 46

[137] Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. "merge conflicts!"exploring the impacts of external distractors to parametric knowledge graphs. *arXiv 2309.08594*, 2023. 46

[138] Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. Freshllms: Refreshing large language models with search engine augmentation. *arXiv 2310.03214*, 2023. 46

[139] Hanseok Oh, Haebin Shin, Miyoung Ko, Hyunji Lee, and Minjoon Seo. Ktrl+f: Knowledge-augmented in-document search. *arXiv 2311.08329*, 2023. 46

[140] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv 2310.11511*, 2023. 46

[141] Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Lm-polygraph: Uncertainty estimation for language models. *arXiv 2311.07383*, 2023. 46

[142] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv 2303.08896*, 2023. 46

[143] Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv 2305.15852*, 2024. 46

[144] Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *arXiv 2305.03268*, 2023. 46

[145] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv 2305.14325*, 2023. 46

[146] Aleksander Buszydlik, Karol Dobiczek, Michał Teodor Okoń, Konrad Skublicki, Philip Lippmann, and Jie Yang. Red teaming for large language models at scale: Tackling hallucinations on mathematics tasks. *arXiv 2401.00290*, 2023. 46

[147] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 2023. 46

[148] Miriam Rateike, Celia Cintas, John Wamburu, Tanya Akumu, and Skyler Speakman. Weakly supervised detection of hallucinations in llm activations. *arXiv 2312.02798*, 2023. 46

[149] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv 2310.01405*, 2023. 46, 49

[150] Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*, 2023. 50

[151] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. 50

[152] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain

Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Ga-

briella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen,

Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. 50

[153] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 51

[154] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain, July 2004. 52

[155] Elsa Andersson. Methods for increasing cohesion in automatically extracted summaries of swedish news articles. 2022. 53

[156] Juri Opitz. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 12:820–836, 06 2024. 54

[157] Thomas Lumley, Paula Diehr, Scott Emerson, and Lu Chen. The importance of the normality assumption in large public health data sets. *Annual review of public health*, 23(1):151–169, 2002. 56

[158] Marvin M Kilgo III. *The statistical sleuth: a course in methods of data analysis*. Taylor & Francis, 1998. 56