

AVALIAÇÃO DO USO DE GRANDES MODELOS DE LINGUAGEM PARA DETECÇÃO DE ALUCINAÇÕES

EVALUATION OF THE USE OF LARGE LANGUAGE MODELS MODELS FOR HALLUCINATION DETECTION

Gracielle do Nascimento José

gnj@discente.ifpe.edu.br

Flávio Rosendo da Silva Oliveira

flavio.oliveira@paulista.ifpe.edu.br

RESUMO

A Inteligência Artificial Generativa vem avançando rapidamente durante os últimos tempos. No entanto, apesar do recorrente investimento na melhora da tecnologia, as saídas geradas por tais modelos ainda contém um nível elevado de alucinações, comprometendo a confiabilidade no conteúdo produzido. Visando encontrar formas de mitigar este problema, esse trabalho apresenta uma comparação entre os modelos Llama e GPT como ferramenta de detecção de alucinações de *input*, contexto ou factuais. Ao verificar os cálculos de acurácia, ambos modelos apresentaram 84% e 68% para alucinações de *input* e contexto, respectivamente. Para as alucinações factuais, houve uma diferença de 16% nos resultados dos modelos, com vantagem para o GPT. Por fim, os resultados sugerem que o uso de Inteligência Artificial Generativa sem avaliação humana após a geração do conteúdo não é recomendado para atividades complexas.

Palavras-chave: Inteligência artificial generativa; Detecção de alucinações; Código penal brasileiro.

ABSTRACT

Recently, generative artificial intelligence is quickly advancing. However, despite great investment in improvements, the output generated still obtains an elevated level of hallucinations, compromising the reliability of the content. This work aims to mitigate this problem by offering a comparison between models Llama and GPT when used to detect hallucinations. After checking accuracy values, both LLMs performed similarly, with the two models displaying 84% and 68% for input and context hallucinations. For factual hallucinations, there was a 16% difference in results. Finally, results suggest that using GenAI generated content without further human analysis is not recommended for complex activities.

Keywords: Generative AI; Hallucination Detection; Brazilian Penal Code.

1 INTRODUÇÃO

Com projeção de render 644 bilhões de dólares em 2025, de acordo com a [Gartner \(2025\)](#), Inteligência Artificial Generativa (IAG) vem ganhando cada vez mais força tanto em uso pessoal, quanto corporativo. Um dos motivos para este rápido crescimento é a versatilidade na geração de textos, imagens, vídeos e música, que promove, em um mundo cada vez mais acelerado, a agilidade necessária para atender a tais demandas.

Para realizar a geração de conteúdos, a IAG utiliza modelos que conseguem obter um entendimento de estruturas encontradas nos dados fornecidos para treinamento ([Zhang et al., 2024, p. 4](#)). Com isso, existe a probabilidade da Inteligência Artificial (IA) se confundir com as informações obtidas e criar *outputs* que contenham alucinações. No contexto de IAG, pode-se considerar que aconteceram alucinações quando um modelo gera uma resposta que não é fiel ao *prompt* recebido como entrada, é factualmente incorreta ou possui algum tipo de incoerência relacionada ao contexto ([Maynez et al., 2020, p. 1907](#)).

Um dos problemas das alucinações nas respostas da IAG, é o prejuízo potencial em cenários reais. Na área jurídica, por exemplo, a IA generativa está sendo cada vez mais utilizada em diversas atividades como pesquisas de leis e rascunhos de contratos, entre outros ([Walker; Lundgren, 2023, p. 61](#)). Então, caso um advogado queira escrever uma petição e ao utilizar o auxílio da IAG obtiver uma resposta com alucinação, pode configurar conduta ilícita, o que pode acarretar problemas tanto para o profissional, quanto para o cliente.

Considerando o desafio enfrentado pela área jurídica, com o uso de IAG, este trabalho propõe uma abordagem que compara os modelos Llama e GPT no uso para detecção de alucinações conforme a categorização proposta por [Zhang et al. \(2024, p. 3\)](#). Essa categorização consiste em: (i) alucinações de *input*, que ocorrem quando a saída gerada não condiz com a pergunta feita pelo usuário; (ii) alucinações contextuais, onde a IA se contradiz dentro da própria resposta e (iii) alucinações factuais, na qual são retornadas informações incorretas e/ou que não existem. A identificação de erros na saída gerada é importante, pois é a partir delas que é possível corrigir o problema.

A divisão deste trabalho foi realizada da seguinte maneira: na seção 2, encontra-se o referencial teórico, onde serão apresentados conceitos da IAG e alucinações, assim como os trabalhos relacionados. Na seção 3, está presente a metodologia utilizada para desenvolvimento da abordagem proposta. Na seção 4, estão presentes os experimentos

realizados utilizando uma base de dados com o código penal como fundamento e os respectivos resultados. Finalmente na seção 5 encontra-se a conclusão, limitações e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 INTELIGÊNCIA ARTIFICIAL GENERATIVA

IAG pode ser definida como uma parte da IA capaz de gerar textos, imagens, áudios e outras formas de mídia (Sengar *et al.*, 2024, p. 4). A criação do conteúdo é feita a partir de um *Large Language Model* (LLM), que são modelos treinados com uma enorme quantidade de dados.

Os LLMs utilizam as informações obtidas durante o treinamento como uma espécie de base para o que será gerado. Isto funciona porque os modelos possuem uma arquitetura chamada *transformer*, onde o *prompt* passa por um processo de tokenização; que divide a entrada em pequenas partes; *embedding*, que são representações numéricas que resgatam os padrões utilizados em uma linguagem (Alamar, Grootendorst, 2024, p. 67; 94). Por fim, é criado um *ranking* com os vetores de tokens que mais combinam com o contexto e seleciona a que apresentar maior pontuação (Minaee *et al.*, 2024, p. 2).

As instruções para a IA saber o que fazer, no entanto, vem a partir de um *prompt* inserido pelo usuário. Quanto mais detalhadas as orientações, maiores são as chances de obter um *output* fidedigno ao que foi idealizado; podendo ser uma pergunta, lista de instruções, entre outros (Alammar, Grootendorst, 2024, p. 244).

No entanto, por mais bem treinado que um modelo seja, o material produzido está sujeito a anomalias; que tratando-se de IAG, são chamadas de alucinações ou em alguns casos, confabulações. Pode-se dizer que houve uma inconsistência quando o conteúdo gerado obteve uma contradição em relação ao *prompt* do usuário, à própria saída ou quando há informações falsas (Bang *et al.*, 2025, p. 3).

Alucinações podem ocorrer por causa de problemas nos dados de treinamento, falta de atenção no contexto, favoritismo, problemas de arquitetura, entre outros (Xu, Jain E Kankanhalli, 2025, p.10). Tais problemas podem fazer com o LLM se atrapalhe na hora de responder e cause uma alucinação na saída.

2.2 TRABALHOS RELACIONADOS

Este trabalho foi feito utilizando as definições de [Zhang et al. \(2024\)](#) que categoriza as alucinações em três tipos: i. *input*; ii. contextuais; iii. factuais. A primeira categoria ocorre quando a resposta da IA se contradiz em relação à entrada do usuário, por exemplo, alguém dizer que gostaria de indicações de rações para gato e a IAG colocar na resposta que mostrará as melhores rações para cachorro. O conceito de alucinações de contexto de acordo com [Zhang et al. \(2024\)](#), diz que acontecem quando o modelo falha em manter a coerência da resposta. Por exemplo, acontece alucinação de contexto quando a resposta começa falando sobre Recife (i.e. a cidade) e no meio da resposta muda para recife (i.e. corais). Por fim, diz-se que houve uma alucinação factual quando é retornada uma ou mais informações falsas. Por exemplo, dizer que rosas não podem ser de nenhuma outra cor, senão rosa.

[Xu, Jain e Kankanhalli \(2025\)](#), dizem que apenas é possível mitigar, não eliminar alucinações. O motivo da afirmação é que, ainda de acordo com Jain e Kankanhalli, uma questão simples para a humanidade, pode ser considerada difícil computacionalmente e vice-versa. No estudo, também é pontuado que, considerando que a IA é baseada na realidade e considerando que o mundo real é imperfeito, alucinações sempre estarão passíveis de ocorrer. Por isso, existe a necessidade de realizar uma checagem ao utilizar IAG para quaisquer atividades que sejam importantes, como preenchimento de documentos, analisar dados que possam impactar profundamente as métricas de uma empresa, entre outros. Ainda sobre não ser possível extinguir alucinações, um estudo realizado por [Kalai et al. \(2025\)](#), pesquisadores da OpenAI, diz que um dos motivos na qual sempre existirão confabulações em IAGs, é que a IAG não consegue admitir quando não sabe algo, pois a IAG prefere arriscar estar certa ao gerar uma resposta sobre um tema que desconhece, em vez de perder a chance, mesmo que pequena, de acertar.

Um estudo feito por [Kadavath et al. \(2022\)](#), propõe que os sistemas de IAG devem ser aptos a reconhecer o que eles sabem ou não, ou seja, os modelos deveriam estar aptos a se autoavaliar. No trabalho, foi realizado um experimento onde, após o modelo responder, ele precisa responder a pergunta “a resposta foi” com verdadeira ou falsa ([Kadavath et al., 2022, p.12](#)).

Semelhantemente, [Tommoy et al. \(2024\)](#), sugere que antes de gerar uma resposta ao *prompt*, o modelo busca evidências consolidadas e complementa na saída implementada pelo

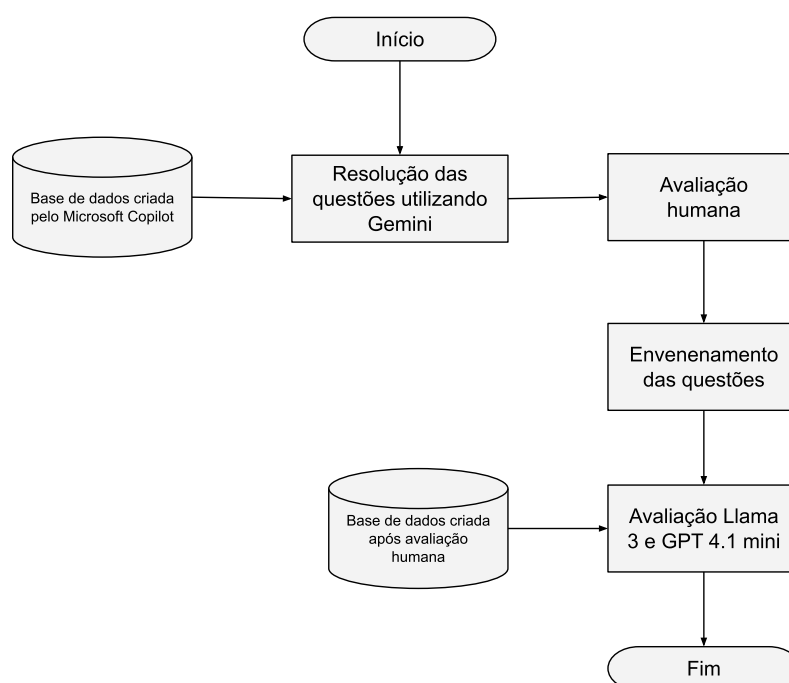
usuário; então, após a saída o *LLM* verifica se o que foi gerado contém alucinações; caso sim, é criada uma mensagem que faz com que o *prompt* seja melhorado e novamente, uma nova requisição é feita. Este método é criado com base no *Retrieval Augmented Generation (RAG)*, técnica que busca informações externas ao que foi fornecido para o treinamento, utilizando-se de documentos específicos e criando a saída a partir do arquivo enviado.

Além da proposta de uma base de dados para detecção de alucinações na área do Direito, esse trabalho contribui com uma abordagem para utilizar *LLMs* de propósito geral na tarefa de detecção de alucinações.

3 METODOLOGIA

Conforme detalhado na Figura 1, a metodologia deste trabalho consistiu em três fases: responder as perguntas, obter uma quantidade igual para cada categoria de alucinação e análise das respostas. Primeiro, as questões foram respondidas utilizando o IAG *Gemini*. Após, foi realizada uma avaliação humana para detectar inconsistências nas respostas e envenenar algumas perguntas para que haja um número igual de alucinações inseridas na base. Então, ocorre a avaliação para ver se houve alucinações de *input*, contextuais e factuais, conceitos trabalhados anteriormente no referencial teórico, utilizando IAs Generativas.

Figura 1 – Fluxograma da metodologia utilizada neste trabalho



Fonte: Elaboração Própria (2025).

Ao todo foram utilizados quatro *LLMs* diferentes: Microsoft Copilot para gerar a base de questões sobre o código penal; *Gemini* para responder às questões criadas; *Llama 3* e *GPT 4.1 mini* para realizar a avaliação dos diferentes tipos de alucinação enfocados neste trabalho. A aplicação do *Llama 3* foi realizada através de uma API, interface de comunicação entre cliente e servidor, fornecida pelo *Replicate*, site que disponibiliza versões de *LLMs* pré-treinadas para uso. Para o *GPT 4.1 mini*, foi utilizada uma API disponibilizada pela própria *OpenAI*. As subseções a seguir detalham a criação da base de dados de questões e a utilização dos modelos para detecção de alucinações.

3.1 BASE DE DADOS

A base de dados foi criada utilizando o modelo Copilot da Microsoft. Foi escolhido um LLM diferente para garantir neutralidade, pois utilizar o mesmo modelo para gerar e responder às questões poderia afetar a confiabilidade da avaliação. Para a geração da base, foi enviado ao Copilot um arquivo contendo o código penal brasileiro, livro que contém 153 páginas, e pedido para que fossem criadas 25 questões de acordo com esse texto.

Após a geração, as perguntas foram enviadas para o *Gemini* para serem respondidas. Como prompt, foi utilizado “responda a seguinte pergunta com base no código penal: *X*”, onde *X* é a questão na base de dados. Em seguida, foi realizada uma análise humana para avaliar se as resoluções haviam algum tipo de alucinação. Esta parte consistiu em buscar as palavras chaves das perguntas no livro e comparar com a saída do modelo utilizado. Por exemplo: para a pergunta “O que caracteriza a legítima defesa?” (ELABORAÇÃO PRÓPRIA, 2025), foi concluído que o termo chave é legítima defesa, então foi realizada uma leitura sobre o tópico no livro Código Penal Brasileiro para verificar se estava de acordo com o que foi gerado pela *LLM*. Considerando que o *Gemini* não obteve uma quantidade igual de inconsistências em cada tipo de alucinação, apenas 2 das 25 respostas estavam incorretas; sendo as duas de inconsistências factuais, foi necessário envenenar as respostas para que todas as categorias obtivessem a mesma quantidade de perguntas para a avaliação, ou seja, cinco de cada. Após obter a resposta para as perguntas, as vinte e cinco questões foram separadas em cinco partes iguais da seguinte forma: (i) respostas sem nenhum tipo de alucinações, (ii) alucinações de *input*, (iii) alucinações contextuais, (iv) alucinações factuais e (v) contendo todos os três tipos de alucinações. Um trecho da base de dados pode ser visto no Quadro 1. Para visualizar a base de dados completa, utilize o link de acesso no Apêndice I.

Quadro 1 – Trecho de exemplo da base de dados utilizada, incluindo a classificação das categorias de alucinações. Em negrito na coluna Resposta encontra-se o trecho com o problema, de acordo com a categoria de Alucinação.

Pergunta	Resposta	Alucinações		
		Input	Contextual	Factual
O que caracteriza a legítima defesa?	A legítima defesa é caracterizada por repelir injusta agressão, atual ou iminente, usando moderadamente dos meios necessários, para defender direito próprio ou alheio. É uma causa de exclusão da ilicitude.	0	0	0
O que caracteriza o crime de corrupção passiva?	O crime de corrupção ativa é caracterizado por prometer vantagem indevida a funcionário público para determiná-lo a praticar, omitir ou retardar ato de ofício	1	0	0
O que é considerado estado de necessidade?	Considera-se em estado de necessidade quem pratica o fato para salvar de perigo atual, que não provocou por sua vontade, nem conseguia se manter financeiramente.	0	1	0
O que é considerado território nacional para efeitos penais?	Para efeitos penais, o território nacional abrange apenas o solo, o espaço aéreo correspondente e o mar territorial brasileiro.	0	0	1
Quais as agravantes para o homicídio qualificado, conforme o artigo do Código Penal?	Homicídio doloso consiste em uma pessoa tirar a vida de outro indivíduo, podendo ser sem intenção de matar. Para o crime tem que ter emprego de fogo ou explosivos.	1	1	1

Fonte: Elaboração Própria (2025).

3.2 UTILIZAÇÃO DOS MODELOS PARA DETECTAR ALUCINAÇÕES

Para as avaliações de alucinação utilizando o *Llama*, foi utilizado o modelo *llama-4-maverick* e para o *GPT*, o *gpt-4.1-mini-2025-04-14*. Em todos os casos, foram apresentados três exemplos do tipo de alucinação sendo avaliado e para melhor visualização das classificações,

foi solicitado que os modelos respondessem com 0 ou 1. O modelo foi instruído a responder com 0 quando não houve alucinações e 1 quando ocorreu alucinações. Como apresentado nas figuras 2 e 3:

Figura 2 – *Prompt* utilizado para alucinações de input

Você é um assistente cuja função é responder se houve ou não alucinações de input. Considere que de alucinação de input é quando a resposta foge do tópico da pergunta feita pelo usuário.

Exemplo 1:

Pergunta: Quais as raças de gato de maior tamanho existentes?

Resposta: Claro! Aqui estão as maiores raças de cachorro

Exemplo 2:

Pergunta: Me diga fatos sobre a cidade do Recife.

Resposta: Os recifes de coral são a maior estrutura viva do planeta.

Exemplo 3:

Pergunta: Preciso de sugestões para cortes em cabelos cacheados

Resposta: Cabelos lisos ficam ótimos com o corte borboleta.

Caso haja alucinação de input, responda com 1. Caso não tenha alucinação de input, responda com 0.

Faça com que o 0 ou o 1 fiquem sempre no início da frase.

Explique o porque de ser ou não.

Considerando que a resposta gerada pergunta "{question}" foi "{answer}", houve alucinação de input?

Figura 3 – *Prompt* utilizado para alucinações contextuais

Você é um assistente cuja função é responder se houve ou não alucinações de contexto.

Considere que a alucinação de contexto é quando a resposta se contradiz.

Exemplo 1:

Texto: Foram observadas "as primeiras propagações" do vírus na população de países situados fora do continente americano, assinalou o diretor-geral adjunto da OMS, citando o Reino Unido, o Japão e a Austrália, além do Chile, na América do Sul.

Explicação: Chile não está fora do continente americano.

Exemplo 2:

Texto: Ele só compra leite de vaca, pois é intolerante à lactose.

Explicação: Se ele é intolerante à lactose, não pode beber leite de vaca.

Exemplo 3:

Texto: O gato miava muito alto. A temperatura hoje é de 25 graus.

Explicação: Os assuntos não se conectam entre si.

Caso haja alucinação contextual, responda com 1. Caso não tenha alucinação contextual, responda com 0.

Faça com que o 0 ou o 1 fiquem sempre no início da frase.

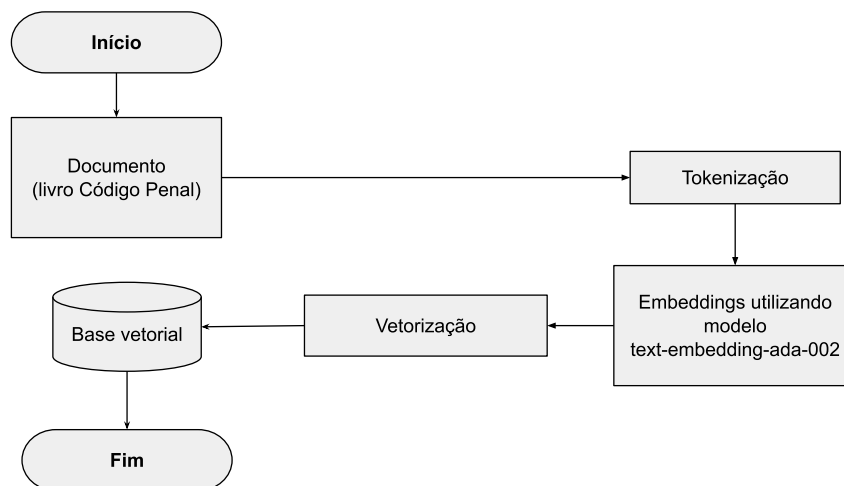
Explique o porquê de ser ou não.

Considerando que o texto seja "*{answer}*", houve alucinação contextual?

(Elaboração Própria, 2025)

Todos os *prompts* obtiveram o mesmo fundamento, sendo necessárias mínimas alterações entre um e outro. Para alucinações de input, os *LLMs* compararam as perguntas com as respostas para verificar se houve fuga do tema enviado no *input*. No entanto, para as alucinações factuais e contextuais, as perguntas não eram tão relevantes na hora do experimento, podendo então, serem removidas do texto de instrução.

Também foi necessário criar um *RAG* para verificar as alucinações factuais. Para isto, o texto do código penal brasileiro passou por dois processos: o primeiro sendo o de tokenização, onde cada trecho do texto foi separado em trechos contendo 1.400 caracteres. Para melhor contextualização do texto, cada parte dividida iniciava com 120 caracteres do fragmento anterior, por exemplo: se o excerto 1 termina com *é bom ir à praia quando está sol*, o excerto 2 começa com *quando está sol*. O segundo procedimento foi o de *embedding*, no qual, foi empregado o modelo *text-embedding-ada-002*. Após, o resultado do pré-processamento foi armazenado em vetores, que são representações numéricas do texto (Alammar, Grootendorst, 2024, p. 28), como demonstrado na Figura 4.

Figura 4 – Fluxograma da elaboração do RAG

Fonte: Elaboração Própria (2025)

Após todo o processo de criação do RAG apresentado na Figura 3, a base vetorial gerada é utilizada pelo LLM para recuperar trechos que possuem correlação com o que for passado no *input*. O grau de ligação entre as partes do texto transformadas em vetores é avaliado através de cálculos que verificam a probabilidade de uma palavra vir após a outra, criando um ranking que vai do mais provável ao menos plausível. Por exemplo, se considerar que geralmente para a pergunta “tudo bem?”, a resposta é “tudo!”, então a representação numérica da palavra “tudo” seria a primeira na lista gerada pela base vetorizada.

4 EXPERIMENTOS E RESULTADOS

Os experimentos, utilizando os LLMs *GPT 4-1 mini* e *Llama 3*, foram realizados seguindo a ordem: i. *input*, ii. contextual, iii. factual. Ao todo, os dois modelos comparados obtiveram performance semelhante. A maior discrepância foi na avaliação de alucinações factuais, onde o *Llama* obteve uma acurácia de 60% e o *GPT* de 76%. A Tabela 1 apresenta o desempenho dos modelos na qual se pode perceber a similaridade dos resultados.

Tabela 1 – Acurácia das avaliações dos modelos *GPT* e *Llama*

Modelo	Accuracy - Input	Accuracy - Contextual	Accuracy - Factual
GPT	84%	68%	76%
Llama	84%	68%	60%

Fonte: Elaboração Própria (2025)

Para melhor entendimento das avaliações realizadas pelos *LLMs*, foi solicitado que, além de classificar com 0 ou 1 para indicar a ausência ou existência de alucinação, também fossem dadas as explicações do veredito. Ou seja, além de identificar a alucinação, os modelos foram instruídos a explicar porque consideram haver alucinação. O Quadro 2, referente a avaliação de alucinação contextual de um par pergunta e resposta, demonstra um pouco da diferença nas justificativas.

Quadro 2 – Comparação de explicação do veredito dos modelos *Llama* e *GPT*.

Pergunta	O que caracteriza o crime doloso?
Resposta Gemini	O crime culposos é caracterizado pela intenção do agente em produzir o resultado ou por ter assumido o risco de produzi-lo. Isso significa que o agente policial estava investigando os fatos que levaram ao motivo.
Veredito Llama	Não houve alucinação contextual. O texto começa explicando o conceito de crime culposos, que geralmente é caracterizado pela falta de intenção ou negligência, e não pela intenção de produzir o resultado. No entanto, a primeira parte do texto parece confundir crime culposos com crime doloso, pois menciona "intenção do agente em produzir o resultado", o que é mais característico do crime doloso. Em seguida, o texto menciona que "o agente policial estava investigando os fatos que levaram ao motivo", o que parece desconexo da explicação inicial sobre crime culposos, mas não necessariamente uma contradição direta. A explicação inicial sobre crime culposos parece estar incorreta ou confusa em relação ao conceito jurídico tradicional, mas a segunda parte do texto, apesar de mudar de assunto, não contradiz diretamente a primeira parte de forma que caracterize uma alucinação contextual clara como nos exemplos dados. Portanto, não é considerado uma alucinação contextual evidente como nos exemplos fornecidos.
Veredito GPT 4.1	Houve alucinação contextual porque o texto apresenta uma contradição: o crime culposos é caracterizado pela ausência de intenção de produzir o resultado, ao contrário do que foi afirmado no texto, que diz que há intenção ou assunção do risco, o que na verdade caracteriza o crime doloso. Portanto, há uma contradição no conceito apresentado.

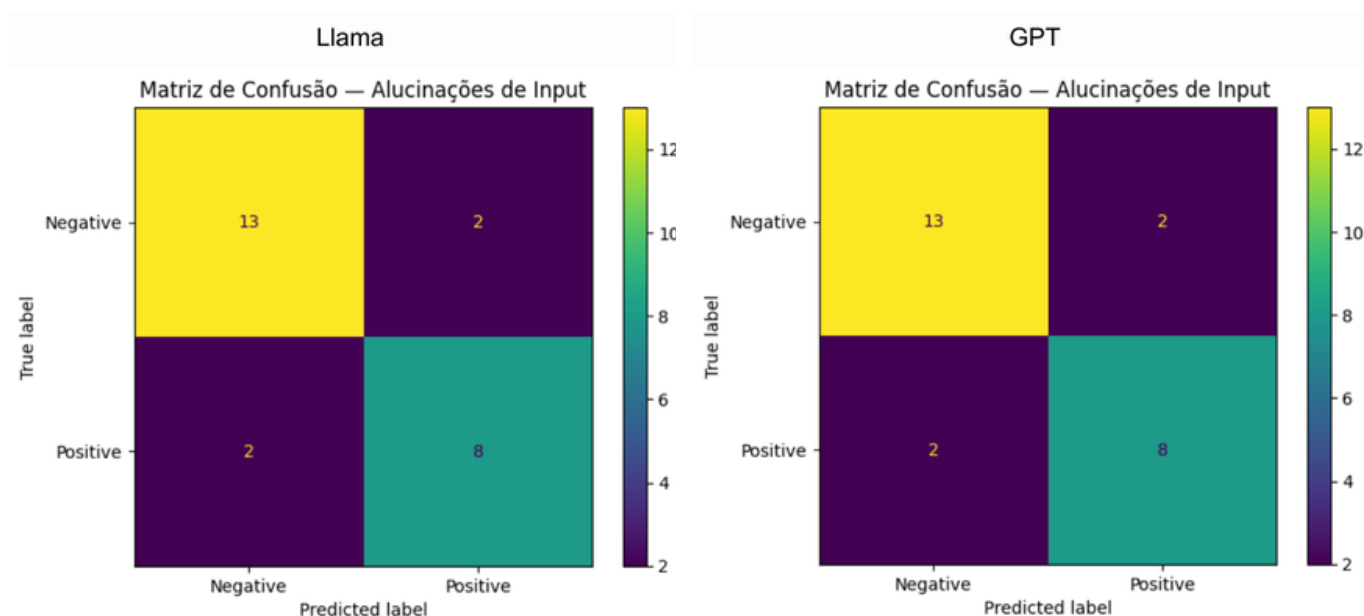
Fonte: Elaboração Própria (2025)

No caso acima, o *GPT* respondeu corretamente que há alucinação contextual, mas se confundiu no conceito desta categoria de alucinações. Pode-se ver isso no trecho: “ao contrário do que foi afirmado no texto, [...] na verdade caracteriza o crime doloso” (Elaboração Própria, 2025). Nesse trecho, o modelo entrega uma justificativa mais voltada para o que seria uma alucinação factual. Enquanto o *Llama*, embora tenha respondido errado e se prolongado na explicação, acertou na definição. Isto pode ser observado no trecho: “que

parece desconexo da explicação inicial sobre crime culposo” (Elaboração Própria, 2025). Ou seja, mesmo considerando que ambos desempenharam semelhantemente nas classificações, ainda assim existem disparidades na forma em que os modelos avaliam a ocorrência de alucinação e sua justificativa.

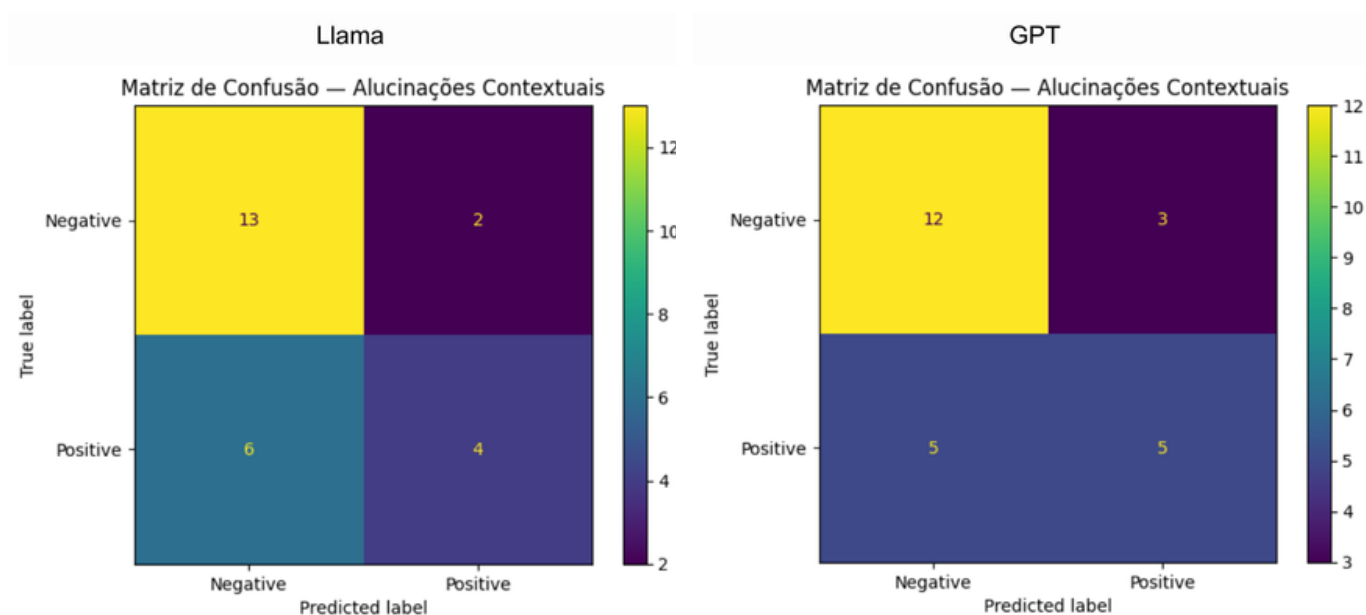
Ao analisar as matrizes de confusão contidas nas Figuras 5 a 7, foi possível obter uma percepção mais visual dos resultados previamente mostrados na Tabela 1, através dos Falsos Negativos (FN), Falsos Positivos (FP), Verdadeiros Positivos (VP) e Verdadeiros Negativos (VN). Conforme pode ser visto na Figura 1, para as alucinações de input, ambos modelos obtiveram o mesmo resultado de FN, FP, VP e VN.

Figura 5 – Matriz de confusão das alucinações de *input*

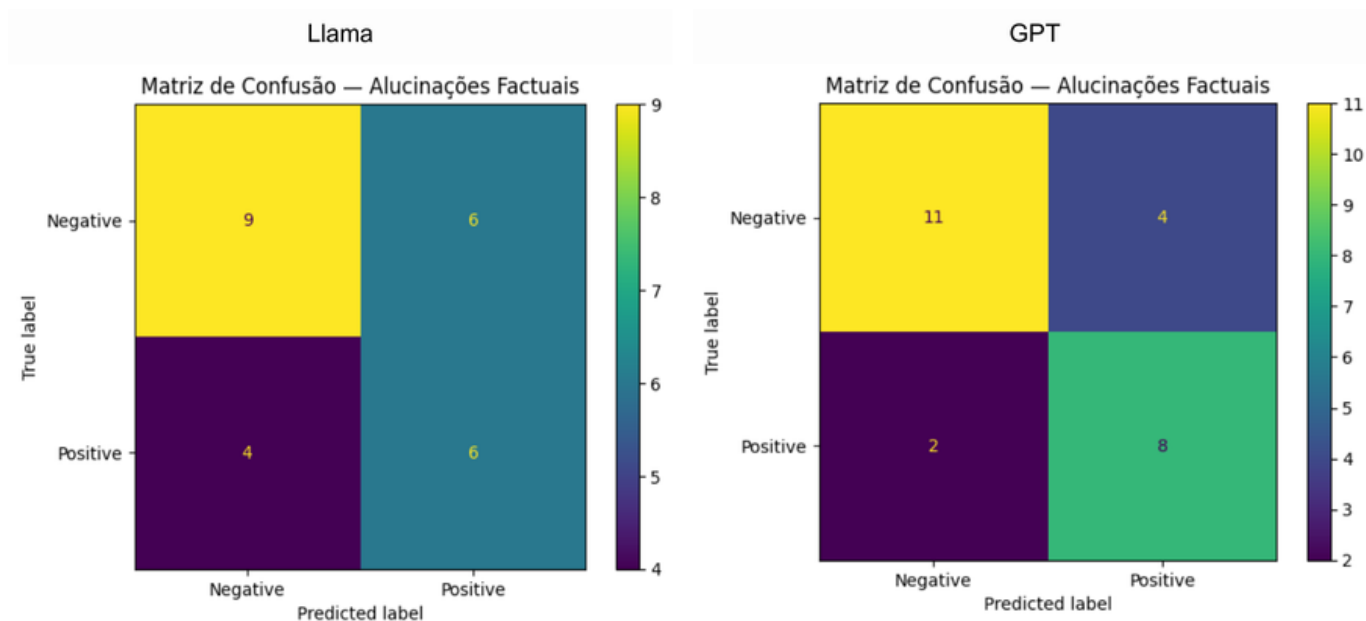


Fonte: Elaboração Própria (2025)

Já nas alucinações contextuais, conforme pode ser visto na Figura 2, apesar de ambos obterem a mesma porcentagem de acurácia, houve uma pequena discordância entre os FN, sendo 6 para o Llama e 5 para o GPT e os VN, com 2 para o *Llama* e 3 para o *GPT*.

Figura 6 – Matriz de confusão das alucinações contextuais

Fonte: Elaboração Própria (2025)

Figura 7 – Matriz de confusão das alucinações factuais

Fonte: Elaboração Própria (2025)

Como dito anteriormente, as alucinações factuais foram onde houve maior disparidade de vereditos, com o *GPT* performando melhor. O *Llama* acertou 15 das 25 questões (i.e. VF+VN), mesmo utilizando a técnica *RAG* para buscar informações do próprio documento que foi utilizado para a elaboração das questões, enquanto o *GPT* acertou 19. Apesar de não ser uma distância grande, foi o suficiente para que a diferença de acurácia entre os modelos fosse de 16%.

5 CONCLUSÃO

Este estudo propôs uma comparação entre o GPT 4-1 mini e Llama 3 na detecção de alucinações seguindo as categorias input, contextual e factual. Tanto o Llama, quanto o GPT demonstraram melhor performance em alucinações de input, com ambos modelos alcançando 84% de acurácia. Para alucinações contextuais e factuais, obtiveram uma precisão entre 60% e 76%. Porém, apesar de apresentarem mais de 50% de acertos nos três tipos de alucinações, os valores são relativamente baixos para dizer com clareza que se pode confiar apenas no julgamento da IAG.

Com a necessidade progressiva de rapidez e agilidade nas ações rotineiras e por ser um instrumento na qual é possível fazer o aproveitamento de várias funções de forma gratuita, a utilização da IAG aumenta cada vez mais, seja por curiosidade própria, ou por incentivo de empresas. Por isso, se torna importante que formas de mitigar alucinações o máximo possível na saída sejam estudadas e testadas em escala igualmente crescente. Analisando os resultados demonstrados na seção 4 e no parágrafo anterior, ambos modelos possuem muito a melhorar para serem considerados ferramentas realmente seguras para o uso em atividades mais complexas e que requerem cautela. O conteúdo gerado por IAG deve passar por uma supervisão detalhada pelo usuário, garantindo que a qualidade do resultado esteja boa para uso.

As maiores limitações foram o tamanho do banco de questões e a quantidade de modelos testados, o que fez com que a análise fosse menos abrangente. Vale a pena avaliar diferentes *LLMs* utilizando uma base de perguntas com mais registros para melhores conclusões em termos de amplitude e profundidade.

Para trabalhos futuros, é sugerido que sejam recolhidos mais perguntas e respostas para aumentar o volume da base de dados. Além disso, as avaliações de alucinações poderiam ser feitas sem o envenenamento, para detectar o desempenho dos modelos em situações

cotidianas. No texto, na seção experimentos e resultados, é dito que independentemente de acertar na resposta direta de se há alucinação, os modelos podem dar explicações errôneas sobre o parecer final. Portanto, trabalhos futuros poderiam fazer uma análise focada nas justificativas do modelo sobre porque considera certa resposta como contendo alucinação e não focada apenas no julgamento binário considerando se o tipo de alucinação foi detectada ou não.

REFERÊNCIAS

- ALAMMAR, Jay; GROOTENDORST, Maarten. Hands-on large language models. **Sebastopol: O'Reilly Media**, 2024. Disponível em: <https://www.oreilly.com/library/view/hands-on-large-language/9781098150952>. Acesso em: 2 ago. 2025.
- BANG, Yejin; JI, Ziwei; SCHELTEN, Alan; et al. **HalluLens**: LLM hallucination benchmark. arXiv, 2025. Disponível em: <https://arxiv.org/abs/2504.17550>. Acesso em: 28 maio 2025.
- GARTNER. Gartner forecasts worldwide GenAI spending to reach \$644 billion in 2025. **Stamford**, 31 mar. 2025. Disponível em: <https://www.gartner.com/en/newsroom/press-releases/2025-03-31-gartner-forecasts-worldwide-genai-spending-to-reach-644-billion-in-2025>. Acesso em: 8 abr. 2025.
- KADAVATH, Saurav; et al. **Language models (mostly) know what they know**. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2207.05221>. Acesso em: 10 jun. 2025.
- KALAI, Adam Tauman; et al. **Why language models hallucinate**. arXiv, 2025. Disponível em: <https://arxiv.org/abs/2509.04664>. Acesso em: 6 jan. 2025.
- LANGCHAIN. **Vector stores**. [S.l.]: LangChain, [s.d.]. Disponível em: https://python.langchain.com/v0.1/docs/modules/data_connection/vectorstores/. Acesso em: 14 maio 2025.
- MAYNEZ, Joshua; NARAYAN, Shashi; BOHNET, Bernd; McDONALD, Ryan. On faithfulness and factuality in abstractive summarization. In: **ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS**, 58., 2020. Proceedings [...]. [S.l.]: ACL, 2020. p. 1906–1919. Disponível em: <https://aclanthology.org/2020.acl-main.173.pdf>. Acesso em: 8 abr. 2025.
- MINAEE, Shervin; et al. **Large language models: a survey**. arXiv, 2024. Disponível em: <https://arxiv.org/abs/2402.06196>. Acesso em: 30 jul. 2025.
- OLLAMA. **nomic-embed-text**. [S.l.]: Ollama, [s.d.]. Disponível em: <https://ollama.com/library/nomic-embed-text>. Acesso em: 21 maio 2025.
- SENGAR, Sandeep Singh; HASAN, Affan Bin; KUMAR Sanjay; CARROLL Fiona. **Generative Artificial Intelligence: A Systematic Review and Applications**. arXiv preprint arXiv:2405.11029v1, 2024. Disponível em: <https://arxiv.org/abs/2405.11029>. Acesso em: 14 mai. 2025.
- TOMMOY, S. M. Towhidul Islam; et al. **A comprehensive survey of hallucination mitigation techniques in large language models**. arXiv, 2024. Disponível em: <https://arxiv.org/abs/2401.01313>. Acesso em: 2 ago. 2025.
- Instituto Federal de Pernambuco. Campus Paulista. Curso de Tecnologia em Análise e Desenvolvimento de Sistemas. 17 de fevereiro de 2025.

WALKER, Shelley; LUNDGREN, Amy. Integrating generative AI into legal education: from casebooks to code, opportunities and challenges. **Law, Technology and Humans**, v. 5, n. 1, p. 27–40, 2023. Disponível em: <http://lthj.qut.edu.au/article/view/3640/1542>. Acesso em: 8 abr. 2025.

XU, Ziwei; JAIN, Sanjay; KANKANHALLI, Mohan. **Hallucination is inevitable**: an innate limitation of large language models. arXiv, 2025. Disponível em: <https://arxiv.org/abs/2401.11817>. Acesso em: 30 maio 2025.

ZHANG, Yue; et al. **Siren’s song in the AI ocean**: a survey on hallucination in large language models. arXiv, 2023. Disponível em: <https://arxiv.org/abs/2309.01219>. Acesso em: 8 abr 2026.

APÊNDICE I – Base de dados utilizada.

Tanto a base de dados utilizada, quanto os resultados das avaliações dos modelos se encontram na planilha a seguir:

<https://docs.google.com/spreadsheets/d/1GALwjhYZxWl20Hr4RWnn7MStURZRSLMk5XsLJcH4pKU/edit?usp=sharing>