Inconsistência de informações em Modelos de Linguagem de Grande Escala (LLMs) e seus impactos na Segurança da Informação

KILBERT MARLEY FELIX DA SILVA
JEZER SILEN FERREIRA
ANNA CAROLINA DOSENA DA SILVEIRA
JOÃO EMMANUEL D ALKMIN NEVES

Resumo

Modelos de Linguagem de Grande Escala (LLMs), como o ChatGPT, têm revolucionado a interação com informações, mas também geram conteúdos imprecisos que podem comprometer a segurança da informação. Este artigo analisa de que forma esses desvios de confiabilidade impactam os pilares da Confidencialidade, Integridade e Disponibilidade (CID), especialmente em contextos organizacionais que utilizam sistemas automatizados para apoio à decisão. A pesquisa foi conduzida por meio de revisão bibliográfica, com foco em publicações científicas entre 2020 e 2025, aliada a uma etapa experimental com diferentes modelos de linguagem. Os resultados evidenciam falhas recorrentes, como alucinação de IA, respostas contraditórias e viés, que comprometem decisões automatizadas e podem expor sistemas a riscos técnicos e operacionais. Observou-se também variação significativa entre os modelos quanto à consistência e à gravidade das falhas identificadas, indicando comportamentos distintos de resposta. Como contribuição, o estudo propõe diretrizes para mitigação desses riscos, incluindo validação cruzada de informações, supervisão humana e uso de técnicas como *Retrieval-Augmented Generation* (RAG). Conclui-se que, apesar do potencial dos LLMs, seu uso seguro exige auditorias contínuas, validação sistemática das saídas e integração com práticas de governança de IA, contribuindo para a aplicação responsável dessas tecnologias em ambientes corporativos.

Palavras-chave: Segurança da Informação; Inconsistência; Inteligência Artificial; Alucinação de IA.

Information inconsistency in Large Language Models (LLMs) and their impacts on Information Security

Abstract

Large Language Models (LLMs), such as ChatGPT, have revolutionized interactions with information, but they can also generate inaccurate content that may compromise Information Security. This article aims to analyze how these reliability failures affect the pillars of Confidentiality, Integrity, and Availability (CIA), particularly in organizational contexts that rely on automated systems for decision-making support. The research was conducted through a literature review, focusing on scientific publications from 2020 to 2025, combined with an experimental phase involving different language models. The results highlight recurring issues, such as AI hallucinations, contradictory responses, and bias, which can undermine automated decision-making and expose systems to technical and operational risks. Significant variation among models was also observed regarding consistency and the severity of identified failures, indicating distinct behavioral trends. As a contribution, this study proposes guidelines to mitigate these risks, including cross-validation of information, human oversight, and the use of techniques such as Retrieval-Augmented Generation (RAG). It is concluded that, despite the potential of LLMs, their safe use requires continuous auditing, systematic

validation of outputs, and integration with AI governance practices, contributing to the responsible application of these technologies in corporate environments.

Keywords: *Information Security; Inconsistency; Artificial Intelligence; AI Hallucination.*

1 INTRODUÇÃO

A rápida evolução da inteligência artificial (IA), em especial dos Modelos de Linguagem de Grande Escala — *Large Language Models* (LLMs), tem provocado transformações significativas na forma como organizações produzem, processam e consomem informação. A versatilidade dessas tecnologias impulsionou sua adoção em diversos setores estratégicos, abrangendo desde a automação de tarefas operacionais até o suporte à tomada de decisão em contextos complexos. Ferramentas de IA generativa vêm sendo progressivamente integradas a ambientes educacionais, corporativos e a domínios sensíveis, como a segurança cibernética.

Entretanto, a arquitetura estatística que sustenta os LLMs introduz uma limitação estrutural relevante: a geração de informações factualmente incorretas, fenômeno amplamente denominado na literatura como alucinação. Essas ocorrências não se configuram como eventos isolados, mas como consequência do funcionamento probabilístico desses modelos, que estabelecem relações entre tokens sem dispor de mecanismos internos de verificação factual.

No contexto da segurança da informação, esse comportamento representa um risco direto aos três pilares da tríade Confidencialidade, Integridade e Disponibilidade (CID). Informações imprecisas podem levar à corrupção de bases de conhecimento, à adoção de decisões técnicas inadequadas e à exposição indevida de dados sensíveis, especialmente em ambientes corporativos que utilizam automação baseada em IA.

Estudos recentes indicam que tais erros factuais já vêm sendo explorados como vetores de ataque. Um exemplo é o fenômeno conhecido como *slopsquatting*, no qual alucinações de LLMs são utilizadas para induzir o uso de dependências maliciosas na cadeia de suprimentos de software, ampliando a superfície de ataque organizacional.

A crescente dependência de sistemas baseados em LLMs, muitas vezes sem mecanismos robustos de validação, também impõe desafios à conformidade com normas e *frameworks* de segurança. A geração de configurações incorretas, interpretações equivocadas de incidentes ou referências inexistentes pode comprometer a Integridade dos processos organizacionais e afetar a Disponibilidade de serviços críticos.

Diante desse cenário, este trabalho busca responder à seguinte questão de pesquisa: de que maneira as falhas geradas por LLMs podem ser identificadas e mitigadas de forma sistemática, considerando seus impactos na segurança da informação?

Parte-se da hipótese de que a adoção de um *framework* de verificação multifacetado — combinando validação cruzada com bases confiáveis, supervisão humana e técnicas como Geração Aumentada por Recuperação, *Retrieval-Augmented Generation* (RAG) — pode reduzir significativamente a incidência dessas falhas.

Assim, o objetivo deste trabalho é propor e avaliar diretrizes voltadas à mitigação dos riscos associados ao uso de LLMs, com foco na preservação da tríade CID e na promoção de práticas seguras de governança e uso de IA.

Como justificativa, destaca-se a necessidade crescente de equilibrar inovação tecnológica e segurança informacional. Ao propor diretrizes fundamentadas e aplicáveis, este estudo contribui para o desenvolvimento de sistemas mais confiáveis e para o uso responsável de LLMs em ambientes organizacionais.

2 REFERENCIAL TEÓRICO

Este tópico reúne os principais conceitos, definições e fundamentos teóricos que sustentam a elaboração deste trabalho, servindo de base para a análise e discussão dos dados pesquisados. Além disso, busca estabelecer uma conexão estruturada entre as características dos LLMs e sua influência sobre a proteção informacional, especialmente no contexto da tríade CID.

2.1 Fundamentos dos modelos de linguagem e seus desafios de segurança

LLMs são sistemas de inteligência artificial baseados em redes neurais profundas, treinados com grandes volumes de dados textuais para aprender padrões estatísticos da linguagem natural. Esses modelos operam, majoritariamente, a partir da arquitetura *Transformer*, proposta por Vaswani et al. (2017), que utiliza mecanismos de *self-attention* para capturar relações contextuais entre palavras ao longo de sequências extensas de texto.

Embora essa arquitetura tenha permitido avanços expressivos em tarefas como tradução automática, sumarização e geração de conteúdo, sua base estatística impõe limitações estruturais relevantes. Conforme discutido por Bender et al. (2021), esses modelos operam a partir de correlações sobre dados textuais, sem garantia de compreensão semântica ou capacidade intrínseca de verificação factual. Essa característica favorece a ocorrência de respostas não ancoradas na realidade — fenômeno conhecido como alucinação de IA, no qual o modelo gera informações plausíveis, porém incorretas ou inexistentes.

Do ponto de vista da segurança da informação, essa característica representa uma vulnerabilidade nativa. A incapacidade de distinguir o verdadeiro do falso compromete diretamente o pilar da Integridade, pois informações incorretas podem ser incorporadas a bases de conhecimento, relatórios técnicos ou sistemas automatizados de decisão. Além disso, estudos como o de Carlini et al. (2021) demonstram que LLMs podem memorizar e reproduzir dados sensíveis presentes em seus conjuntos de treinamento, evidenciando riscos adicionais à Confidencialidade.

Relatórios técnicos recentes reforçam essa preocupação. O OpenAI GPT-4 System Card (2024) e pesquisas conduzidas por Bai et al. (2022) indicam que mecanismos como memória contextual, *fine-tuning* e Aprendizado por Reforço com Feedback Humano — *Reinforcement Learning from Human Feedback* (RLHF), embora aumentem a utilidade dos modelos, também ampliam o risco de vazamento de dados (*data leakage*) e retenção indevida de informações sensíveis. Assim, as limitações dos LLMs não se restringem à precisão das respostas, mas afetam diretamente aspectos críticos da proteção de dados.

Diante dessas limitações estruturais, torna-se necessário analisar como tais fragilidades se manifestam em aplicações práticas, especialmente em cenários de interação direta com usuários e sistemas automatizados.

2.2 Vulnerabilidades emergentes em aplicações de LLM

As vulnerabilidades associadas aos LLMs não se limitam à sua arquitetura interna, manifestando-se de forma mais evidente na interação entre entradas (*prompts*) e saídas (*outputs*). Nesse contexto, a alucinação de IA e a engenharia de *prompt* configuram um mesmo eixo de risco relacionado ao controle insuficiente das interfaces de entrada e saída dos modelos.

A alucinação, definida na literatura como a geração de conteúdo fluente porém não ancorado em informações verificáveis, é uma vulnerabilidade amplamente documentada em

modelos de linguagem. Conforme discutido por Huang et al. (2023), esse fenômeno pode surgir devido a limitações nos dados de treinamento, na representação estatística da linguagem e no processo de inferência, levando o modelo a produzir respostas plausíveis, mas factualmente incorretas.

A OWASP (2023), em seu projeto Top 10 for LLM Applications, alerta que saídas incorretas podem ser consumidas automaticamente por sistemas externos, como *plugins* e agentes autônomos, resultando em falhas críticas de segurança. Esse risco é evidenciado pela vulnerabilidade LLM07 (*Insecure Plugin Design*), na qual uma informação incorreta pode levar à execução de comandos inválidos ou potencialmente perigosos.

Casos reais demonstram a materialidade dessa ameaça. Goodin (2025) documenta incidentes nos quais LLMs geraram nomes de bibliotecas inexistentes, possibilitando ataques de *slopsquatting*, em que adversários registram pacotes maliciosos com os nomes sugeridos pelo modelo. Esse tipo de exploração transforma a alucinação em um vetor direto de comprometimento da cadeia de suprimentos de software, afetando tanto a Integridade quanto a Confidencialidade dos sistemas envolvidos.

Paralelamente, a engenharia de *prompt*, quando explorada de forma maliciosa, dá origem aos ataques de *prompt injection*. A OWASP (2023) classifica a *Prompt Injection* (LLM01) como uma das vulnerabilidades mais críticas em aplicações baseadas em LLMs. Esses ataques incluem técnicas como *jailbreaks*, *prompt injection chains* e *indirect prompt injection*, amplamente catalogadas pelo MITRE (2024). Segundo Greshake et al. (2023), conteúdos externos consumidos por aplicações integradas a LLMs podem incorporar instruções adversariais capazes de subverter o comportamento esperado do modelo, caracterizando ataques de *indirect prompt injection*. Nesses cenários, atacantes inserem instruções que levam o modelo a divulgar informações sensíveis ou executar ações indevidas.

A interação entre alucinação e engenharia de *prompt* evidencia que as vulnerabilidades dos LLMs assumem caráter operacional concreto, impactando diretamente os três pilares da tríade CID.

Além das vulnerabilidades técnicas, é necessário considerar fatores relacionados à confiabilidade das respostas, incluindo vieses algorítmicos e aspectos de governança.

2.3 Viés algorítmico, ética e governança em IA

Conforme discutido por Bender et al. (2021), modelos de linguagem refletem e amplificam vieses presentes nos dados utilizados em seu treinamento, reproduzindo padrões sociais, culturais e políticos existentes. Esse fenômeno pode comprometer a neutralidade e a confiabilidade das informações geradas e, em contextos institucionais, influenciar decisões de forma sistemática.

Embora frequentemente tratado sob uma perspectiva ética, o viés algorítmico também constitui um problema de governança e de resiliência informacional. *Frameworks* contemporâneos reconhecem esse risco como mensurável e passível de gestão. O NIST AI Risk Management Framework (2023) propõe um ciclo estruturado de identificação, mensuração e mitigação de riscos, no qual o viés é tratado como uma ameaça à Integridade dos sistemas de IA.

Avanços regulatórios reforçam essa abordagem. O EU AI Act (2024) estabelece obrigações específicas para sistemas de IA de alto risco, exigindo transparência, rastreabilidade e mecanismos de mitigação de vieses. De forma complementar, a norma ISO/IEC 42001:2023 propõe um sistema de gestão de IA que integra princípios éticos, governança e segurança ao ciclo de vida dos modelos.

No contexto brasileiro, esses requisitos dialogam diretamente com a Lei Geral de Proteção de Dados (LGPD), especialmente no que se refere à responsabilidade

organizacional. A ausência de mecanismos de governança adequados pode resultar em riscos operacionais, legais e reputacionais.

Considerando a combinação de vulnerabilidades técnicas e desafios de governança, torna-se essencial estruturar abordagens formais de gestão de riscos aplicadas a sistemas baseados em LLMs.

2.4 Gestão de riscos e conformidade em sistemas baseados em LLM

Os estudos revisados evidenciam que os riscos associados aos LLMs extrapolam falhas isoladas de design técnico, alcançando o domínio da governança da informação e da gestão organizacional. Alucinações, injeções de *prompt*, vazamento de dados e vieses algorítmicos configuram ameaças interdependentes que exigem uma abordagem estruturada de gestão de riscos.

Nesse contexto, normas e *frameworks* consolidados tornam-se instrumentos essenciais para traduzir essas ameaças em processos mensuráveis. A ISO/IEC 27001:2022 e a ISO/IEC 27005:2022 fornecem diretrizes para identificação, análise e tratamento de riscos, enquanto o NIST AI RMF (2023) complementa essa abordagem ao considerar as especificidades dos sistemas de IA.

A integração desses referenciais permite alinhar controles técnicos — como validação cruzada, filtros de conteúdo e supervisão humana — a processos organizacionais formais de avaliação de riscos, reduzindo impactos sobre a Confidencialidade, Integridade e Disponibilidade e assegurando conformidade regulatória.

Em síntese, o referencial teórico evidencia que os LLMs apresentam limitações estruturais, vulnerabilidades operacionais e desafios de governança que afetam diretamente a confiabilidade dos sistemas de informação. Há, entretanto, uma lacuna na validação empírica dessas fragilidades em tarefas técnicas concretas, o que fundamenta a etapa experimental deste estudo.

3 METODOLOGIA

Esta seção descreve os procedimentos metodológicos adotados para alcançar os objetivos propostos, detalhando o delineamento da pesquisa, os métodos de coleta, o tratamento e a análise dos dados, bem como os critérios de validade e confiabilidade.

Do ponto de vista operacional, o estudo busca identificar, medir e analisar falhas de segurança geradas por LLMs em respostas a tarefas técnicas, considerando seu impacto direto na tríade CID.

3.1 Delineamento da pesquisa

A presente pesquisa é de natureza aplicada, pois busca produzir conhecimentos voltados à solução de problemas práticos no campo da segurança da informação, especificamente os riscos decorrentes do uso de LLMs.

Quanto aos objetivos, o estudo apresenta caráter exploratório e descritivo. É exploratório ao investigar vulnerabilidades emergentes associadas à adoção de LLMs em contextos de alta criticidade, e descritivo ao identificar e detalhar comportamentos recorrentes de falha e seus impactos sobre os pilares da Confidencialidade, Integridade e Disponibilidade.

A abordagem metodológica é mista, combinando análise qualitativa e quantitativa. A análise qualitativa, fundamentada na revisão da literatura, fornece o embasamento teórico para a definição das categorias de risco. A análise quantitativa, derivada do experimento, permite mensurar a frequência, recorrência e gravidade das falhas observadas.

3.2 Procedimentos de coleta de dados

Os procedimentos de coleta foram estruturados em duas frentes complementares, de modo a assegurar tanto a fundamentação teórica quanto a validação empírica dos riscos analisados.

3.2.1 Revisão bibliográfica

A revisão bibliográfica teve como objetivo identificar o estado da arte sobre alucinações, vulnerabilidades e mecanismos de mitigação associados aos LLMs.

Foram consultadas as bases Google Scholar, SciELO e IEEE Xplore, priorizando artigos científicos, normas técnicas e relatórios institucionais. O recorte temporal compreendeu o período entre 2020 e 2025, considerando a rápida evolução da IA generativa.

Os descritores utilizados incluíram: "segurança da informação", "Inteligência Artificial", "Modelos de Linguagem", "Alucinação de IA", "OWASP LLM" e "NIST AI Framework".

Os critérios de inclusão adotados foram: (a) publicações em português ou inglês; (b) publicadas entre 2020 e 2025; (c) com abordagem direta sobre LLMs, alucinação de IA, segurança da informação ou governança de IA; e (d) disponíveis integralmente para consulta. Foram excluídos trabalhos que tratavam de IA de forma genérica sem relação com segurança ou com LLMs especificamente, e fontes sem identificação de autoria ou sem possibilidade de verificação. A seleção final priorizou relevância temática, atualidade e aderência ao contexto de segurança cibernética, fornecendo a base teórica para a definição dos cenários analisados na etapa experimental.

3.2.2 Teste experimental com LLMs

A etapa experimental avaliou a consistência, a precisão técnica e os riscos de segurança presentes nas respostas geradas por diferentes LLMs diante de tarefas técnicas críticas.

Os testes foram conduzidos em ambiente controlado, utilizando versões públicas dos modelos: ChatGPT, Claude, Gemini e Llama. A escolha desses modelos se justifica pela ampla adoção no mercado, diversidade de arquiteturas e relevância no cenário atual da IA generativa, o que contribui para a representatividade dos resultados.

Para reduzir variáveis externas, os experimentos foram realizados em sessões isoladas, sem histórico de conversação prévia.

Os *prompts* utilizados foram construídos com base em cenários reais de segurança, alinhados a vulnerabilidades documentadas em *frameworks* como o OWASP Top 10. Cada *prompt* representa um tipo específico de risco:

- *Prompt* 1 – Login em PHP: foco em SQL Injection e armazenamento inseguro de credenciais;
- *Prompt* 2 – Política AWS IAM: foco em permissões excessivas e violação do princípio do menor privilégio;
- *Prompt* 3 – JavaScript: foco em vulnerabilidades de Cross-Site Scripting (XSS) e execução indevida.

Cada *prompt* foi executado três vezes por modelo, totalizando nove execuções por LLM. A repetição teve como objetivo identificar variabilidade e comportamento inconsistente entre respostas ao mesmo estímulo.

3.3 Tratamento e análise dos dados

O tratamento dos dados foi realizado em duas abordagens complementares. Para fins deste estudo, define-se falha como qualquer desvio identificado na resposta do modelo em relação às boas práticas de segurança documentadas pela OWASP, NIST ou pelas normas ISO/IEC aplicáveis — incluindo código inseguro, orientações incorretas, permissões excessivas e menções de capacidades inexistentes. Define-se alucinação como a geração de informações plausíveis porém factualmente incorretas ou inexistentes, conforme Huang et al. (2023). O risco associado a cada falha foi classificado em quatro níveis — baixo, médio, alto e crítico — com base no impacto potencial sobre os pilares da tríade CID.

Na análise quantitativa, foram utilizadas as seguintes métricas:

- Número total de falhas por resposta;
- Frequência de vulnerabilidades por tipo;
- Recorrência de falhas entre execuções;
- Distribuição de falhas por modelo;
- Classificação da gravidade (baixo, médio, alto, crítico).

Essas métricas permitiram comparar o desempenho dos modelos e identificar comportamentos recorrentes de falha.

Na análise qualitativa, as respostas foram avaliadas quanto à coerência técnica, aderência às boas práticas de segurança e impacto potencial sobre os pilares da tríade CID.

A integração entre as análises ocorreu na etapa interpretativa, em que os resultados quantitativos foram contextualizados com base na literatura, distinguindo-se achados empíricos de inferências interpretativas.

3.4 Validade e confiabilidade da pesquisa

A validade da pesquisa foi assegurada pela utilização de cenários alinhados a vulnerabilidades reconhecidas, como as documentadas pela OWASP e pelo NIST.

A escolha de múltiplos modelos e a repetição dos testes contribuíram para ampliar a validade externa, permitindo maior generalização dos resultados.

A confiabilidade foi reforçada pela padronização dos *prompts*, execução controlada e uso de critérios objetivos de identificação de falhas, possibilitando a replicação do estudo por outros pesquisadores.

3.5 Limitações da pesquisa

Apesar do rigor metodológico, o estudo apresenta algumas limitações. Os testes foram realizados em ambiente controlado, o que pode não refletir completamente cenários reais de uso.

Além disso, os modelos avaliados estão sujeitos a atualizações constantes, o que pode alterar seu comportamento ao longo do tempo, exigindo reavaliações periódicas. A ausência de acesso a versões internas ou configurações específicas também pode influenciar os resultados.

Essas limitações devem ser consideradas na interpretação dos resultados e na generalização dos achados.

3.6 Justificativa metodológica

As escolhas metodológicas adotadas fundamentam-se na necessidade de analisar um fenômeno técnico emergente sob múltiplas perspectivas.

A combinação entre revisão bibliográfica e teste experimental permite integrar fundamentos teóricos com evidências empíricas. A utilização de cenários práticos de segurança reforça a aplicabilidade dos resultados em contextos organizacionais.

Dessa forma, o método adotado permite testar a hipótese proposta, avaliando de forma sistemática a ocorrência de falhas e a necessidade de mecanismos de mitigação.

4 RESULTADOS E DISCUSSÕES

Esta seção apresenta e discute os resultados obtidos na fase experimental, analisando o comportamento, a consistência e os riscos de segurança associados aos LLMs quando aplicados a tarefas técnicas com impacto na tríade CID.

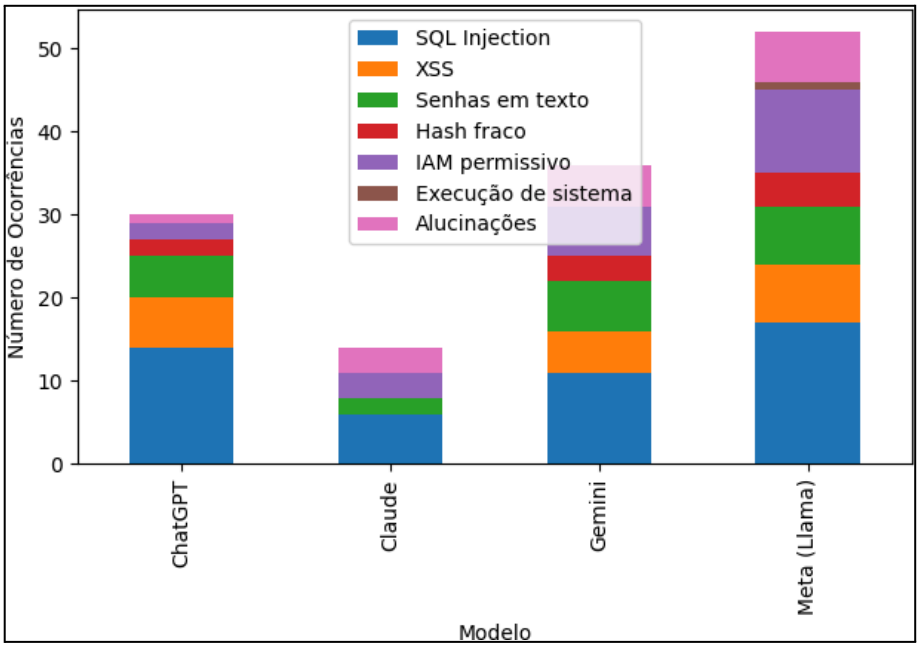
Foram avaliados quatro modelos amplamente utilizados: ChatGPT, Claude, Gemini e Llama. Cada modelo foi submetido a três cenários distintos, com múltiplas execuções, permitindo identificar comportamentos recorrentes de falha.

4.1 Análise comparativa do desempenho dos modelos

O desempenho dos modelos foi avaliado considerando critérios como precisão técnica, consistência das respostas e ocorrência de vulnerabilidades.

O Gráfico 1 apresenta a distribuição total de vulnerabilidades identificadas por modelo.

Gráfico 1 – Total de vulnerabilidades identificadas por modelo



Fonte: Autores (2025)

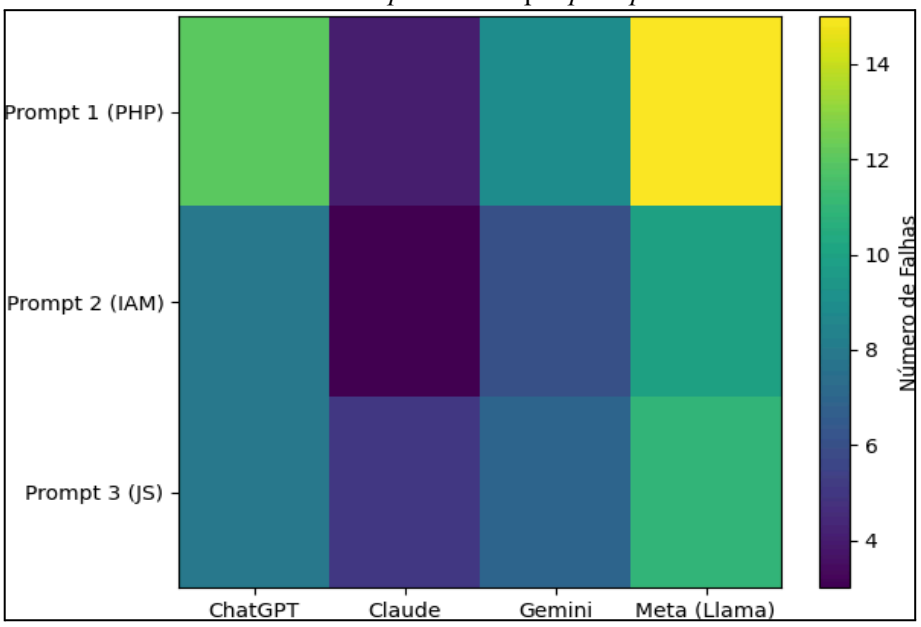
A análise dos resultados evidencia diferenças relevantes entre os modelos. O Claude apresentou maior consistência e menor incidência de falhas críticas, indicando comportamento mais estável entre execuções. Em contraste, o Llama concentrou o maior número de vulnerabilidades, com destaque para falhas relacionadas a *SQL Injection* e permissões excessivas.

O ChatGPT apresentou boa estrutura textual e clareza nas respostas, porém com ocorrência de orientações inseguras em cenários específicos. Já o Gemini demonstrou respostas tecnicamente válidas, mas com maior variabilidade entre execuções, o que reduz sua previsibilidade.

Esses comportamentos distintos têm correspondência direta com as causas estruturais identificadas no referencial teórico. Modelos como o Llama, com menor refinamento por *feedback* humano específico para segurança, tendem a priorizar a plausibilidade sintática da resposta em detrimento da correção técnica — o que explica a maior incidência de SQL Injection e IAM permissivo em suas saídas. O Claude, por sua vez, apresentou menor taxa de falhas, comportamento consistente com o uso de mecanismos de alinhamento mais robustos, como o descrito por Bai et al. (2022). Esses achados são de natureza empírica e descritiva: refletem o comportamento observado nas versões públicas dos modelos testados no período do estudo, sem possibilidade de generalização absoluta dado o caráter evolutivo dessas ferramentas.

O Gráfico 2 apresenta um comparativo visual do desempenho dos modelos por critério.

Gráfico 2 – *Heatmap* de falhas por *prompt* e modelo



Fonte: Autores (2025)

De forma geral, todos os modelos apresentaram algum nível de desvio de confiabilidade, evidenciando que a consistência técnica ainda é um desafio em aplicações críticas. Esse resultado reforça a hipótese de que mecanismos adicionais de validação são necessários.

4.2 Consolidação e comparativo quantitativo das vulnerabilidades

A consolidação dos dados foi realizada com base nas execuções experimentais, categorizada conforme os padrões da OWASP.

A Tabela 1 apresenta a distribuição das vulnerabilidades por tipo e por modelo.

Tabela 1 – Distribuição de vulnerabilidades por tipo e modelo

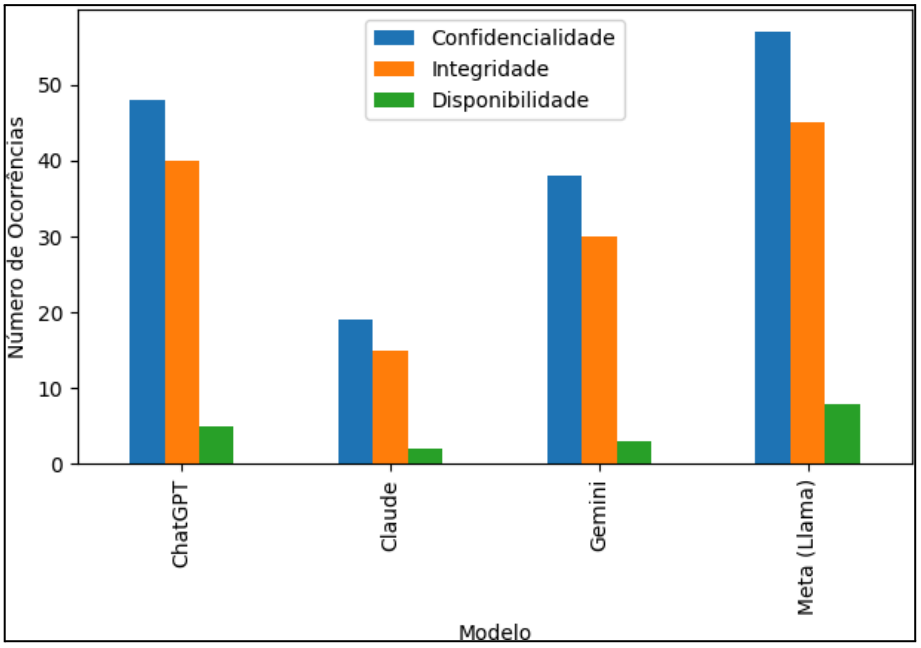
LLM	SQLI (concatenação)	*Real escape*	Senha em texto	Hash fraco	IAM permissivo	XSS (innerHTML)	Execução de sistema	Alucinação	Total
Chat GPT	14	3	5	2	2	6	0	1	43
Claude	6	1	2	0	3	0	0	3	19
Gemini	11	2	6	3	6	5	0	5	38
Meta (Llama)	17	5	7	4	10	7	1	6	57

Fonte: Autores (2025)

A análise quantitativa revela que o Llama apresentou o maior número total de ocorrências (57), seguido por ChatGPT (43), Gemini (38) e Claude (19). Esse resultado indica que, embora todos os modelos apresentem falhas, há variação significativa na frequência e na gravidade das ocorrências.

O Gráfico 3 ilustra a distribuição das vulnerabilidades por categoria.

Gráfico 3 – Distribuição das vulnerabilidades por tipo e por modelo



Fonte: Autores (2025)

No total, foram identificadas 157 ocorrências de vulnerabilidades. Entre elas, destacam-se *SQL Injection*, XSS, permissões excessivas em IAM e armazenamento inseguro de credenciais, conforme detalhado na Tabela 2.

Tabela 2 – Principais vulnerabilidades identificadas

Vulnerabilidade	Total ocorrências	Impacto
SQL Injection (sem *prepared statements*)	48	Crítico
XSS via innerHTML / eval	22	Alto
Senhas em texto plano	20	Crítico

Permissões excessivas em IAM (*)	27	Alto
Uso de *real_escape_string* (mitigação fraca)	11	Médio
Uso de hash fraco (MD5/SHA1)	10	Médio
Chamadas de sistema (exec / system)	2	Crítico
Menções de capacidade ("*I can...*")	18	Sinal de alucinação

Fonte: Autores (2025)

A maioria das vulnerabilidades identificadas é classificada como de alto ou crítico impacto, afetando diretamente a segurança dos sistemas analisados.

A Tabela 3 apresenta um comparativo qualitativo entre os modelos.

Tabela 3 – Comparativo de desempenho dos LLMs

Critério	ChatGPT	Claude	Gemini	Meta
Precisão técnica geral	Boa	Alta	Média	Baixa
Consistência (repetição de boas práticas)	Alta	Alta	Média	Baixa
Risco médio por resposta textual ("*I can...*")	Médio	Baixo	Médio	Alto
Alucinação (Nº de ocorrências)	4	3	5	6
Tendência de erro mais comum	SQLI e XSS	Pequenas permissões	SQLI	SQLI e IAM amplas

Fonte: Autores (2025)

A análise comparativa evidencia comportamentos distintos entre os modelos. O Claude apresenta maior estabilidade e menor risco médio, enquanto o Llama concentra a maior incidência de falhas críticas. ChatGPT e Gemini ocupam posição intermediária, porém com características diferentes: o primeiro com falhas pontuais, o segundo com maior variabilidade entre execuções.

4.3 Análise por prompt e identificação de padrões críticos

A análise por cenário permite identificar tendências recorrentes de vulnerabilidade associadas a cada tipo de tarefa.

A Tabela 4 apresenta os resultados do *Prompt* 1 (Login PHP).

Tabela 4 – Detalhamento Prompt 1: login PHP (vulnerabilidade principal: SQL Injection / senhas inseguras)

LLM	SQLI ($_POST sem *prepare*)	Senha em texto	Hash fraco	Total *Prompt* 1
ChatGPT	8	3	1	12
Claude	3	1	0	4

Gemini	6	2	1	9
Meta (Llama)	10	3	2	15

Fonte: Autores (2025)

Este cenário concentrou o maior número de falhas, principalmente relacionadas a *SQL Injection* e armazenamento inseguro de senhas. O resultado indica que tarefas envolvendo manipulação de dados sensíveis representam maior risco quando automatizadas por LLMs — provavelmente porque esses modelos priorizam a funcionalidade imediata do código gerado em detrimento da aplicação de práticas seguras como *prepared statements*.

A Tabela 5 apresenta os resultados do *Prompt* 2 (AWS IAM).

Tabela 5 – Detalhamento Prompt 2: política AWS IAM (vulnerabilidade principal: permissões excessivas)

LLM	***Action** **	***Resource** **	Condição ausente	Total *Prompt* 2
ChatGPT	3	2	3	8
Claude	1	1	1	3
Gemini	2	2	2	6
Meta (Llama)	4	3	3	10

Fonte: Autores (2025)

Neste cenário, destaca-se a concessão excessiva de permissões por todos os modelos, evidenciando dificuldade em aplicar corretamente o princípio do menor privilégio. Esse comportamento é consistente com a tendência dos LLMs de gerar configurações permissivas como caminho de menor resistência quando o nível de acesso necessário não é especificado explicitamente no *prompt*.

A Tabela 6 apresenta os resultados do *Prompt* 3 (JavaScript).

Tabela 6 – Detalhamento Prompt 3: comentários JS (vulnerabilidade principal: XSS via innerHTML)

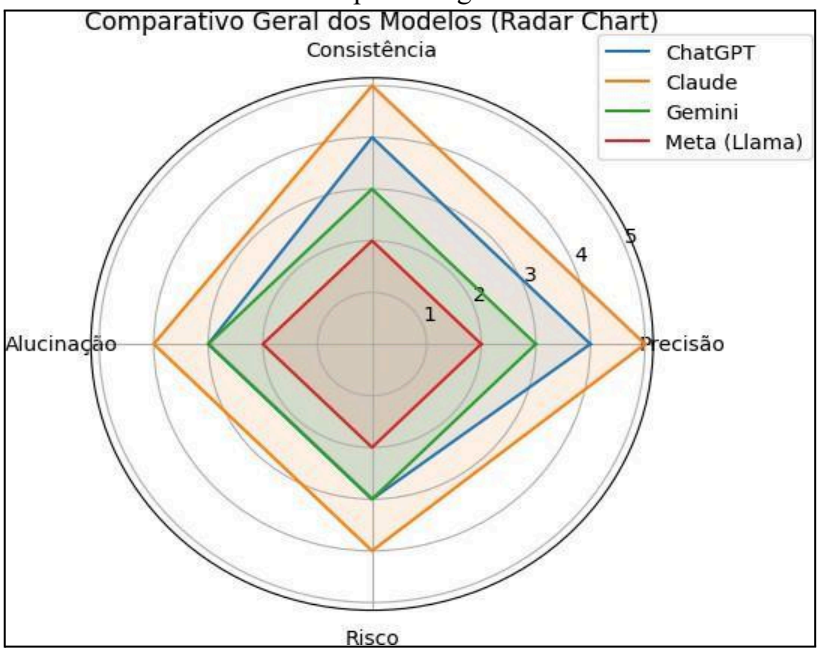
LLM	innerHTML/eval	Sanitização ausente	Execução de sistema	Total *Prompt* 3
ChatGPT	6	2	0	8
Claude	4	1	0	5
Gemini	5	2	0	7
Meta (Llama)	7	3	1	11

Fonte: Autores (2025)

Observa-se alta incidência de vulnerabilidades de XSS, especialmente pelo uso inseguro de funções como innerHTML e eval(). Esse comportamento reflete a predisposição dos LLMs a adotar soluções JavaScript de referência ampla, sem verificar se tais funções expõem a aplicação a injeção de código malicioso.

O Gráfico 4 apresenta um comparativo geral entre os modelos considerando todos os cenários.

Gráfico 4 – Comparativo geral dos modelos



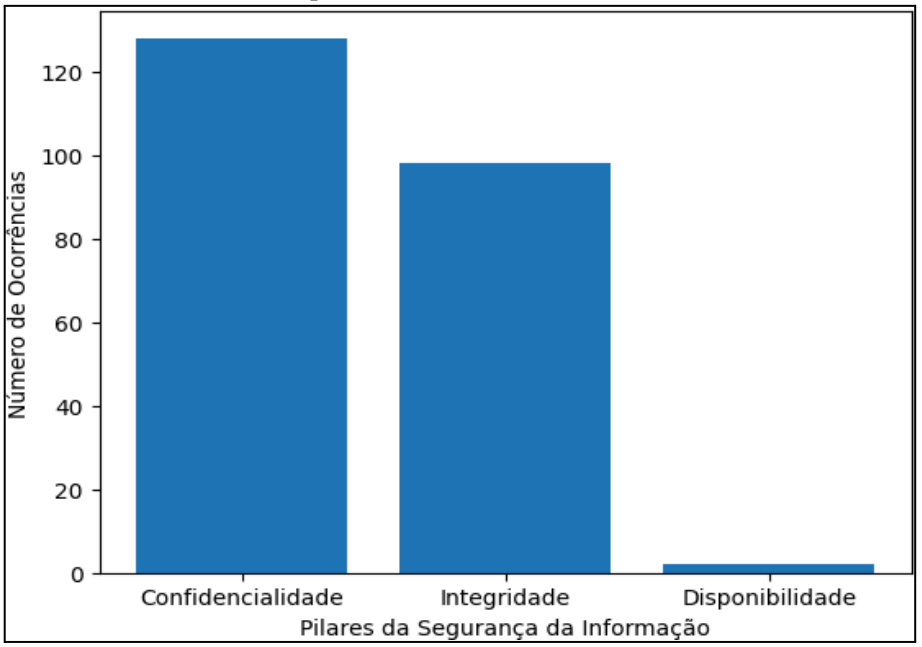
Fonte: Autores (2025)

A análise integrada dos três cenários evidencia que as falhas não são aleatórias, mas seguem tendências recorrentes associadas ao tipo de tarefa proposta — o que reforça a interpretação de que os LLMs apresentam fragilidades sistemáticas em domínios de segurança específicos, e não apenas variação estocástica.

4.4 Integração das vulnerabilidades à tríade CID

A relação entre as vulnerabilidades identificadas e os pilares da tríade CID é apresentada no Gráfico 5.

Gráfico 5 – Impacto das vulnerabilidades na tríade CID



Fonte: Autores (2025)

Os resultados indicam que a maioria das falhas impacta diretamente os pilares da Confidencialidade e da Integridade, especialmente por meio de *SQL Injection*, XSS e permissões excessivas.

Embora menos frequentes, também foram identificadas falhas com impacto na Disponibilidade, principalmente em casos envolvendo execução de comandos de sistema.

Essa distribuição reforça que o uso inadequado de LLMs pode comprometer múltiplos pilares simultaneamente, ampliando o risco organizacional de forma interdependente.

4.5 Implicações para a segurança da informação e conformidade normativa

As vulnerabilidades identificadas possuem implicações diretas na governança de TI e na conformidade com normas de segurança.

Controles previstos na ISO/IEC 27001:2022, como controle de acesso, desenvolvimento seguro e proteção contra código malicioso, são diretamente afetados pelas falhas observadas.

Sob a ótica da LGPD, a utilização de LLMs sem validação adequada pode ser interpretada como falha na adoção de medidas de proteção de dados, gerando riscos legais e reputacionais.

Além disso, os resultados têm impacto direto em áreas operacionais como *Security Operations Center* (SOC), DevSecOps e *Machine Learning Operations* (MLOps), onde decisões automatizadas baseadas em IA podem amplificar falhas de segurança.

De forma geral, os resultados demonstram que a utilização de LLMs em tarefas críticas de segurança exige cautela, validação e governança estruturada. A variabilidade entre modelos e a recorrência de falhas indicam que essas tecnologias não devem ser utilizadas de forma autônoma em ambientes sensíveis.

5 CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo analisar os impactos das falhas geradas por LLMs sobre a segurança da informação, com foco na tríade CID. A integração entre revisão bibliográfica e análise experimental permitiu identificar comportamentos recorrentes de falha e avaliar seus efeitos em cenários técnicos controlados.

Os resultados obtidos confirmam a hipótese de que os LLMs apresentam limitações estruturais que afetam diretamente a confiabilidade das informações geradas. As falhas identificadas — como alucinação de IA, geração de código inseguro, concessão excessiva de permissões e variabilidade entre respostas — demonstram empiricamente que esses modelos, nas versões públicas testadas, ainda não oferecem nível adequado de segurança para uso autônomo em contextos críticos. Ressalta-se que esses achados descrevem o comportamento observado no período do estudo e não implicam generalização absoluta dado o caráter evolutivo das ferramentas.

Do ponto de vista teórico, o trabalho contribui ao estabelecer a relação entre vulnerabilidades de LLMs e os pilares da tríade CID, evidenciando que tais falhas não se limitam à qualidade da informação, mas afetam a Integridade de sistemas, a proteção de dados e a confiabilidade operacional.

No campo prático, os resultados reforçam a necessidade de adoção de mecanismos de controle. Destacam-se como medidas essenciais a validação humana das respostas, a implementação de auditorias contínuas, o uso de técnicas como RAG e a aplicação de boas práticas de desenvolvimento seguro em ambientes que utilizam IA.

Verificou-se ainda que a variabilidade entre modelos representa um fator adicional de risco, dificultando a previsibilidade e a padronização de comportamentos. Esse aspecto exige atenção especial em ambientes corporativos, onde decisões automatizadas podem gerar impactos amplificados.

Como limitações, destaca-se que os testes foram realizados em ambiente controlado e com um conjunto específico de cenários, o que pode não abranger todas as situações reais de uso. A constante evolução dos modelos também exige reavaliações periódicas dos resultados.

Como trabalhos futuros, sugere-se a ampliação dos testes com maior diversidade de cenários e modelos, a análise de aplicações integradas com agentes autônomos, e a investigação de mecanismos automatizados de detecção de falhas de confiabilidade.

Conclui-se que, embora os LLMs representem um avanço significativo na área de inteligência artificial, sua utilização segura depende da implementação de estratégias robustas de validação, controle e governança. Desse modo, este estudo contribui para o uso mais consciente e responsável dessas tecnologias em ambientes organizacionais.

REFERÊNCIAS

BAI, Yuntao et al. **Constitutional AI: Harmlessness from AI Feedback**. 2022. Disponível em: https://arxiv.org/abs/2212.08073. Acesso em: 01 jun. 2026.

BENDER, Emily et al. **On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?**. 2021. Disponível em: https://dl.acm.org/doi/10.1145/3442188.3445922. Acesso em: 24 mai. 2026.

CARLINI, Nicholas et al. **Extracting training data from large language models**. In: **USENIX Security Symposium**. 2021. DOI: https://doi.org/10.48550/arXiv.2012.07805.

EUROPEAN UNION. **Artificial Intelligence Act (EU AI Act)**. 2024. Disponível em: https://artificialintelligenceact.eu/ai-act-explorer/. Acesso em: 24 mai. 2026.

EUROPEAN UNION. **Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations**. Official Journal of the European Union, Luxembourg, 2024. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Acesso em: 01 jun. 2026.

GOODIN, Dan. **AI-generated code could be a disaster for the software supply chain. Here's why**. Ars Technica, 2025. Disponível em: https://arstechnica.com/security/2025/04/ai-generated-code-could-be-a-disaster-for-the-software-supply-chain-heres-why/. Acesso em: 02 jun. 2026.

GRESHAKE, Kai et al. **Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection**. 2023. DOI: https://doi.org/10.48550/arXiv.2302.12173.

HUANG, Lei et al. **A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions**. 2023. Disponível em: https://arxiv.org/abs/2311.05232. Acesso em: 24 mai. 2026.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO/IEC 27001: Information security management systems – Requirements**. Geneva: ISO, 2022.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO/IEC 27005: Information security risk management**. Geneva: ISO, 2022.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO/IEC 42001: Artificial intelligence management systems**. Geneva: ISO, 2023.

MITRE. **MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems**. 2024. Disponível em: https://atlas.mitre.org. Acesso em: 24 mai. 2026.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. **AI Risk Management Framework (AI RMF 1.0)**. 2023. Disponível em: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf. Acesso em: 24 mai. 2026.

OPENAI. **GPT-4 System Card**. 2024. Disponível em: https://cdn.openai.com/papers/gpt-4-system-card.pdf. Acesso em: 24 mai. 2026.

OWASP FOUNDATION. **OWASP Top 10 for Large Language Model Applications**. 2023. Disponível em: https://owasp.org/www-project-top-10-for-large-language-model-applications/. Acesso em: 24 mai. 2026.

OWASP FOUNDATION. **OWASP Generative AI Security Project**. 2024. Disponível em: https://genai.owasp.org/. Acesso em: 24 mai. 2026.

VASWANI, Ashish et al. Attention is all you need. In: **Advances in Neural Information Processing Systems (NeurIPS)**. 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762