PERFIS EPISTEMOLÓGICOS DE INTELIGÊNCIAS ARTIFICIAIS GENERATIVAS SOB DIFERENTES ESTRATÉGIAS DE ENGENHARIA DE PROMPT: UM ESTUDO COMPARATIVO APLICADO À ENGENHARIA DE MATERIAIS

Felipe Dantas dos Santos[1], Jacob Anderson Sousa Xavier[1], Bismarck Luiz Silva[2]

[1]*Universidade Federal do Rio Grande do Norte, Natal, Brasil (felipe.santos.119@ufrn.edu.br)*

[2]*Universidade Federal do Rio Grande do Norte, Natal, Brasil (bismarck.silva@ufrn.br)*

Resumo: Este estudo avaliou a influência da engenharia de prompt no desempenho de ChatGPT, Gemini, Grok e NotebookLM em análises científicas. Foi proposta uma matriz de codificação para comparar precisão, profundidade, fidelidade documental e metacognição. Os resultados mostram que o refinamento do prompt potencializa competências específicas de cada arquitetura, sem alterar seu perfil epistemológico.

Palavras-chave: Engenharia de prompt; Inteligência artificial; Modelos de linguagem; Perfil epistemológico; Metacognição

INTRODUÇÃO

A integração de modelos de linguagem de grande escala (LLMs) no ecossistema científico transcende a mera automação funcional, impactando a estrutura metodológica de produção e disseminação do conhecimento (Hughes et al., 2025). Diferente de ferramentas de busca convencionais, a rápida capilarização de plataformas como o ChatGPT no ambiente acadêmico evidencia uma demanda por sistemas capazes de processar e organizar densos volumes de informação técnica (Baidoo-Anu; Ansah, 2023). Esse cenário reconfigura o acesso ao conhecimento especializado e introduz ganhos de produtividade — estimados em até 37% em fluxos de escrita — ao delegar etapas mecânicas da pesquisa para algoritmos generativos, permitindo maior foco em tarefas de alta complexidade analítica (Hughes et al., 2025).

No contexto científico, as GenAI atuam como ferramentas auxiliares na interpretação de conteúdos especializados, na organização de conceitos e na construção de sínteses técnicas. Além disso, possuem potencial para personalizar a aprendizagem, oferecer suporte contínuo e auxiliar em tarefas de pesquisa, como revisões bibliográficas, análise de dados e geração de hipóteses (Batista; Mesquita; Carnaz, 2024). Segundo Batista, Mesquita e Carnaz (2024), o uso dessas ferramentas no ensino superior promove mudanças profundas nos processos de aprendizagem, especialmente ao integrar tecnologias digitais a abordagens pedagógicas tradicionais para desenvolver o pensamento crítico e a alfabetização digital.

A integração da GenAI na pesquisa científica introduz uma ambivalência profunda que desafia os paradigmas tradicionais de integridade acadêmica (Francis; Jones; Smith, 2025). Por operarem fundamentados na previsão probabilística de tokens em vez de verificabilidade factual, os LLMs carecem de um compromisso intrínseco com a verdade, o que possibilita a formulação de "alucinações" — constructos sintáticos plausíveis, porém desprovidos de base documental ou empírica (Federiakin et al., 2024).

Essa natureza probabilística, quando não mediada por um rigoroso crivo crítico, fomenta riscos de desinformação e pode induzir ao fenômeno do deskilling, onde a autonomia analítica do pesquisador é mitigada pela delegação cognitiva à máquina (Federiakin et al., 2024; Francis; Jones; Smith, 2025). Soma-se a isso o risco de reprodução de vieses eurocêntricos presentes nos dados de treinamento, o que pode comprometer a equidade e a diversidade na produção do conhecimento global (Francis; Jones; Smith, 2025).

Nesse contexto, a engenharia de prompt transcende a dimensão de habilidade técnica para consolidar-se como uma mediação metodológica estratégica, capaz de parametrizar o comportamento algorítmico em direção ao rigor científico (Lee; Palmer, 2025).

A eficácia dos resultados acadêmicos está estritamente vinculada à robustez da arquitetura do comando que deve articular instrução, contexto e parâmetros de saída específicos permitindo que o modelo supere a síntese superficial (zero-shot) em favor de análises tecnicamente sofisticadas (Federiakin et al., 2024;

Lee; Palmer, 2025). Essa governança humana sobre o processo tecnológico alinha-se às diretrizes da Recomendação da UNESCO sobre a Ética da Inteligência Artificial, que estabelece que os sistemas de IA devem ser desenvolvidos e utilizados de forma centrada no ser humano, preservando a autonomia, a responsabilidade, a transparência e a supervisão humana ao longo de todo o ciclo de vida da tecnologia (UNESCO, 2021).

Diante da necessidade de rigor científico, torna-se fundamental investigar como diferentes modelos de IA se comportam diante de conteúdos técnico-científicos especializados, especialmente em áreas que exigem elevada precisão e integração de fenômenos físicos, como a engenharia de materiais. A análise comparativa dessas ferramentas permite compreender não apenas a qualidade das respostas produzidas, mas também os diferentes perfis epistemológicos presentes nos modelos generativos, evidenciando distinções entre abordagens mais didáticas, interpretativas ou puramente mecanicistas, e determinando as limitações das ferramentas em disciplinas que exigem alta capacidade analítica (Ogunleye et al., 2024).

Assim, o presente trabalho tem como objetivo analisar, por meio de um estudo de caso aplicado à Engenharia de Materiais, como diferentes estruturas de prompt influenciam a construção do conhecimento técnico por inteligências artificiais generativas, avaliando o desempenho do ChatGPT, Gemini, Grok e NotebookLM na interpretação dos mecanismos de deformação e tenacificação descritos por Bucknall (1977).

MATERIAL E MÉTODOS

Sob a perspectiva metodológica, o estudo foi conduzido como um estudo de caso comparativo exploratório, no qual um mesmo conjunto de conteúdos técnicos foi submetido a diferentes arquiteturas de inteligência artificial, permitindo analisar como distintas abordagens algorítmicas interpretam, sintetizam e representam o conhecimento científico quando submetidas a diferentes níveis de refinamento de prompt.

Todos os experimentos foram conduzidos durante o mês de maio de 2026, utilizando exclusivamente as versões gratuitas disponibilizadas nas plataformas oficiais do ChatGPT (OpenAI), Gemini (Google), Grok (xAI) e NotebookLM (Google). As interações foram realizadas por meio das interfaces web oficiais de cada ferramenta, empregando sempre uma nova conversa para cada experimento, de modo a evitar interferências decorrentes do histórico conversacional. Todas as inteligências artificiais receberam exatamente o mesmo material de entrada e os mesmos prompts, mantendo constantes as condições experimentais de avaliação. Considerando que esses sistemas comerciais recebem atualizações contínuas e que seus provedores não disponibilizam identificadores públicos permanentes de versão ou *build*, os resultados apresentados representam o comportamento observado nas versões públicas disponíveis durante o período de coleta dos dados, conforme explorado na Tabela 1.

Tabela 1 – Ambiente experimental utilizado na pesquisa.

Item	Descrição
Período da coleta	Maio de 2026
Plataforma de acesso	Interface Web oficial
ChatGPT	Versão gratuita disponibilizada pela OpenAI
Gemini	Versão gratuita disponibilizada pelo Google
Grok	Versão gratuita disponibilizada pela xAI
NotebookLM	Versão gratuita disponibilizada pelo Google
Idioma das interações	Português (Brasil)
Material analisado	Capítulos 6 e 7 de Bucknall (1977)
Tipo de interação	Nova conversa para cada experimento
Número de prompts	2 (genérico e acadêmico-técnico)
Número de respostas analisadas	8 (4 IAs × 2 prompts)

A presente pesquisa caracteriza-se como um estudo qualitativo-comparativo aplicado à análise do desempenho de inteligências artificiais generativas na interpretação de conteúdos técnicos especializados. O estudo teve como objetivo avaliar como diferentes ferramentas de inteligência artificial interpretam, sintetizam, organizam e representam conceitos científicos relacionados aos mecanismos de deformação e tenacificação em polímeros vítreos, utilizando como base dois capítulos da obra selecionada. A escolha das ferramentas ocorreu em função de suas diferentes arquiteturas, objetivos funcionais e abordagens de construção textual, permitindo comparar distintos perfis de geração de conhecimento técnico. Foram selecionadas quatro ferramentas amplamente utilizadas, representando

diferentes arquiteturas e estratégias de geração de respostas.

A metodologia experimental foi estruturada em duas etapas principais de interação com as inteligências artificiais. A unidade de análise adotada nesta pesquisa correspondeu à resposta completa produzida por cada inteligência artificial para cada condição experimental de prompt, totalizando oito conjuntos principais de respostas analisadas (quatro inteligências artificiais × dois tipos de prompt).

Na primeira etapa, foi utilizado um prompt genérico, denominado "prompt ruim", contendo comandos amplos e pouco estruturados, com foco apenas na solicitação de explicações sobre os mecanismos presentes nos capítulos analisados.

"PROMPT RUIM: A partir dos capítulos 6 e 7 do livro enviado, explique os principais conceitos relacionados aos mecanismos de deformação em polímeros vítreos e aos mecanismos de tenacificação por elastômeros. Organize a resposta em tópicos, destaque os conceitos mais importantes, explique os mecanismos físicos envolvidos e indique possíveis relações entre os dois capítulos."

O objetivo dessa etapa foi observar o comportamento inicial das ferramentas diante de comandos com baixo direcionamento metodológico.

Na segunda etapa, foi aplicado um prompt refinado, denominado "prompt acadêmicotécnico", estruturado especificamente para induzir respostas com maior rigor científico, profundidade conceitual, criticidade e integração teórica.

"PROMPT ACADÊMICO-TÉCNICO: Você é uma inteligência artificial atuando como assistente acadêmico-técnico. A partir dos capítulos 6 e 7 do livro fornecido, realize uma análise técnica e didática do conteúdo. O capítulo 6 aborda os mecanismos de deformação em polímeros vítreos e o capítulo 7 aborda os mecanismos de tenacificação por elastômeros. Sua resposta deve: Identificar os principais conceitos presentes em cada capítulo; Explicar os mecanismos físicos envolvidos de forma técnica; Relacionar os conceitos entre os dois capítulos quando houver conexão; Diferenciar claramente os mecanismos apresentados; Utilizar terminologia técnica adequada; Organizar a resposta em tópicos e subtópicos; Explicar os conceitos de maneira clara e aprofundada; Indicar possíveis aplicações práticas ou implicações dos mecanismos discutidos; Evitar inventar informações não presentes ou incompatíveis com os capítulos; Ao final, apresentar uma síntese comparativa entre os capítulos 6 e 7. Além disso, indique: quais conceitos possuem maior relevância dentro dos capítulos; quais pontos podem gerar dúvidas ou interpretações equivocadas; quais limitações existem na sua própria resposta."

Esse segundo prompt orientava as ferramentas a atuarem como assistentes acadêmicotécnicos especializados em engenharia de materiais e física dos polímeros. As respostas produzidas pelas inteligências artificiais foram submetidas a uma análise qualitativa comparativa baseada em seis variáveis principais:

- precisão conceitual;
- profundidade teórica;
- fidelidade ao texto-base;
- capacidade didática;
- organização das informações;
- controle de extrapolações interpretativas.

A precisão conceitual foi avaliada pela capacidade das ferramentas de identificar corretamente os mecanismos de crazing, shear yielding, cavitação elastomérica e dissipação de energia, além da coerência na utilização da terminologia técnica associada aos fenômenos físicos discutidos nos capítulos. A profundidade teórica foi analisada considerando o nível de detalhamento mecanicista das respostas, a presença de modelos físicos, critérios de escoamento, relações microestruturais e discussões associadas à física molecular dos polímeros.

A fidelidade ao texto-base foi avaliada a partir do grau de aderência das respostas ao conteúdo original dos capítulos analisados, observando-se a ocorrência de extrapolações interpretativas, inserção de teorias externas e reorganizações pedagógicas não explicitamente presentes no material-base. A capacidade didática foi analisada considerando clareza textual, progressão lógica dos conceitos, uso de sínteses comparativas e antecipação de dúvidas conceituais comuns. Já a organização das informações foi avaliada pela estrutura hierárquica das respostas, utilização de tópicos, comparações visuais e integração entre os mecanismos físicos discutidos.

O controle de extrapolações interpretativas foi utilizado para verificar o grau de tendência das ferramentas em expandir o conteúdo além do material original. Nessa variável, foram analisadas a frequência de inferências externas, contextualizações adicionais, aplicações práticas e reconstruções conceituais realizadas pelas inteligências artificiais.

Além da análise textual, também foram avaliadas as representações visuais produzidas pelas ferramentas. As figuras geradas foram analisadas segundo critérios de fidelidade conceitual, organização visual, clareza didática, capacidade de integração entre os mecanismos e coerência física das representações esquemáticas. Essa etapa permitiu verificar como diferentes arquiteturas de IA traduzem conteúdos científicos complexos em modelos visuais e diagramáticos.

A correção técnica das respostas foi verificada por comparação direta com os capítulos analisados e com

a literatura clássica sobre deformação e fratura em polímeros vítreos. Foram considerados como conceitos obrigatórios a correta identificação dos mecanismos de crazing e shear yielding, a distinção entre ambos, a influência da tensão hidrostática, a formação de fibrilas e microvazios, os processos de cavitação elastomérica e os mecanismos de dissipação de energia associados à tenacificação.

Para operacionalizar a análise qualitativa, foi construída pelos autores uma matriz de codificação composta por indicadores previamente definidos, distribuídos em dez categorias analíticas: Precisão Conceitual (PC), Profundidade Teórica (PT), Fidelidade ao Texto-Base (FT), Didática (DI), Organização (OR), Extrapolação Interpretativa (EX), Metacognição (MC), Rastreabilidade (RT), Robustez Científica (RC) e Qualidade Argumentativa (QA). Cada indicador foi avaliado de forma binária, atribuindo-se valor 1 quando explicitamente evidenciado na resposta e valor 0 quando ausente.

Posteriormente, os resultados agregados foram convertidos para a escala qualitativa de quatro níveis (Ruim, Bom, Ótimo e Excelente), utilizada apenas para interpretação comparativa dos resultados. Para sistematizar os resultados comparativos, foi utilizada uma escala qualitativa de quatro níveis, apresentados na Tabela 2:

Tabela 2. Escala qualitativa utilizada para classificação do desempenho das inteligências artificiais.

Classificação	Critério
Ruim	Conceitos ausentes ou incorretos
Bom	Conceitos corretos porém superficiais
Ótimo	Conceitos corretos e bem desenvolvidos
Excelente	Conceitos corretos, aprofundados e integrados

Após a etapa de codificação, os resultados foram normalizados em escala percentual, permitindo a comparação direta entre as ferramentas avaliadas. Para cada categoria analítica, o desempenho foi calculado pela razão entre o número de códigos presentes e o número total de códigos pertencentes àquela categoria.

A Equação 1 foi utilizada para obtenção dos índices percentuais:

$$Índice\ (\%) = \frac{CP}{CC} \times 100 \qquad (1)$$

Onde CP representa o número de códigos presentes e CC o número total de códigos da categoria.

Os percentuais obtidos foram posteriormente utilizados na construção das matrizes comparativas, gráficos radar e mapas de calor empregados na análise dos resultados, permitindo visualizar comparativamente o comportamento das inteligências artificiais nos cenários de “prompt ruim” e “prompt melhor”. Essa abordagem possibilitou identificar quantitativamente o impacto da estrutura do prompt sobre o desempenho das ferramentas e compreender como diferentes arquiteturas de IA respondem ao aumento do rigor acadêmico do comando. Os resultados obtidos foram interpretados à luz da literatura sobre inteligência artificial generativa, engenharia de prompt e ensino técnico-científico, buscando compreender como diferentes modelos constroem conhecimento especializado e quais limitações e potencialidades apresentam para aplicação em contextos acadêmicos e científicos.

Além das métricas tradicionais de desempenho, foi realizada uma análise dos perfis epistemológicos emergentes das inteligências artificiais avaliadas. Essa abordagem buscou identificar padrões predominantes de construção do conhecimento técnico-científico.

Além da avaliação individual das categorias analíticas, foram definidos cinco perfis epistemológicos com o objetivo de caracterizar as diferentes estratégias de construção e representação do conhecimento científico adotadas pelas inteligências artificiais. Os perfis foram obtidos por agregação das categorias codificadas, sintetizando diferentes dimensões do desempenho em indicadores representativos do comportamento epistemológico de cada ferramenta. O perfil Documental foi calculado pela média das categorias Fidelidade ao Texto-Base (FT), Rastreabilidade (RT) e Extrapolação Interpretativa (EX), considerando esta última de forma inversa, de modo que menores níveis de extrapolação resultassem em maior fidelidade documental. O perfil Didático foi obtido pela média entre Didática (DI) e Organização (OR); o perfil Integrador pela combinação entre Precisão Conceitual (PC) e Qualidade Argumentativa (QA); o perfil Científico pela média entre Profundidade Teórica (PT) e Robustez Científica (RC); e o perfil Expansivo pela associação entre Profundidade Teórica (PT) e Extrapolação Interpretativa (EX). Essa classificação permitiu analisar não apenas o desempenho global das inteligências artificiais, mas também identificar seus diferentes padrões de organização, aprofundamento, fidelidade documental e expansão do conhecimento científico.

Os gráficos radar foram utilizados para comparar simultaneamente múltiplas dimensões de desempenho entre as inteligências artificiais, permitindo visualizar diferenças relativas entre as ferramentas e entre os cenários de prompt genérico e prompt acadêmico-técnico. Os mapas de calor foram empregados para representar a intensidade de ocorrência dos códigos e dos indicadores calculados, facilitando a identificação de padrões de desempenho, sensibilidade ao refinamento dos prompts e predominância dos perfis epistemológicos.

Complementarmente, foram construídas matrizes de ganho percentual e matrizes de sensibilidade ao prompt, permitindo quantificar o impacto do refinamento do comando sobre cada categoria analítica avaliada.

Como limitação metodológica, destaca-se que a análise foi conduzida sobre um único conjunto de conteúdos técnicos, pertencentes à área de engenharia de materiais, não sendo possível generalizar integralmente os resultados para outros domínios do conhecimento. Além disso, os resultados refletem o desempenho das ferramentas nas versões disponíveis durante o período de realização da pesquisa, estando sujeitos a alterações decorrentes de futuras atualizações dos modelos.

Cada combinação entre inteligência artificial e prompt foi executada uma única vez, reproduzindo o cenário mais comum de utilização por estudantes e pesquisadores. Embora múltiplas execuções permitam avaliar a variabilidade inerente aos modelos de linguagem, o objetivo deste estudo foi comparar o comportamento das ferramentas em condições reais de uso. Estudos futuros poderão incorporar análises de estabilidade mediante múltiplas repetições sob os mesmos parâmetros experimentais.

Durante a elaboração deste manuscrito foram utilizadas ferramentas de inteligência artificial generativa (ChatGPT, Gemini, Grok e NotebookLM) exclusivamente como objeto de estudo e, de forma complementar, como apoio à organização textual, revisão linguística e estruturação preliminar de algumas seções do manuscrito. Todas as análises científicas, definição da metodologia, construção da matriz de codificação, interpretação dos resultados, elaboração das figuras, validação das informações, conferência das referências bibliográficas e redação final foram realizadas e revisadas criticamente pelos autores. Os autores assumem integral responsabilidade pelo conteúdo científico, pelas interpretações apresentadas e pela versão final do manuscrito.

RESULTADOS E DISCUSSÃO

O EFEITO DA ENGENHARIA DE PROMPT

A comparação entre as respostas produzidas pelas quatro inteligências artificiais demonstra que o refinamento do prompt alterou o desempenho analítico das ferramentas, embora com intensidades distintas entre as arquiteturas avaliadas. Em todos os casos, a substituição do comando genérico por um prompt acadêmico-técnico promoveu respostas mais organizadas, aprofundadas e cientificamente estruturadas, evidenciando que a engenharia de prompt atua como um mecanismo de direcionamento capaz de potencializar capacidades já existentes em cada modelo, sem modificar sua natureza epistemológica (Lee; Palmer, 2025; Federiakin et al., 2024).

A Tabela 3 sintetiza essa evolução ao comparar o comportamento das ferramentas antes e após o refinamento do comando. Enquanto ChatGPT e Gemini migraram de respostas predominantemente descritivas para abordagens mais didáticas e integradoras, o Grok ampliou principalmente a profundidade teórica por meio da incorporação de modelos mecanicistas adicionais. O NotebookLM, por sua vez, preservou sua característica central de elevada fidelidade documental, expandindo a capacidade de integração entre os capítulos analisados sem recorrer à inclusão de conhecimentos externos. Esses resultados demonstram que o mesmo estímulo metodológico produz respostas distintas em função da arquitetura de cada sistema, corroborando observações de Federiakin et al. (2024) sobre a influência conjunta da engenharia de prompt e da estrutura interna dos modelos.

Tabela 3. Desempenho Absoluto das Ias

IA	*Comportamento (Prompt Ruim)*	*Evolução (Prompt Acadêmico-Técnico)*	*Perfil Epistemológico Identificado*
ChatGPT	Respostas predominantemente descritivas e superficiais, focadas em definições básicas.	Grande salto em didática e organização. Apresenta sequenciamento físico lógico e hierarquia clara.	Didático-Integrador: Prioriza a clareza e a progressão do conhecimento para o usuário.

Grok	Já apresenta certa densidade técnica inicial, focando em modelos físicos como Eyring e Argon.	Máximo ganho em profundidade teórica e rigor. Incorpora teorias complexas (Argon, Robertson) de forma sistemática.	Analítico-Mecanicista: Focado no rigor científico e em modelos matemáticos/físicos de deformação.
Gemini	Foca em conceitos fundamentais de escoamento e cavitação de forma resumida.	Destaca-se pela capacidade comparativa e síntese visual. Relaciona bem aplicações práticas como HIPS e ABS.	Sintético-Comparativo: Excelente em estabelecer pontes conceituais e aplicações industriais.
NotebookLM	Mantém fidelidade ao texto, citando mecanismos de Bucknall e Smith.	Desempenho superior em fidelidade documental. Controla extrapolações e foca na interação sinérgica entre capítulos.	Documental-Ancorado: Especializado em respostas estritamente fundamentadas no material-base (RAG).

A Figura 1 apresenta o ganho absoluto obtido após o refinamento do prompt para cada variável analisada. Observa-se comportamento uniforme nas categorias Precisão Conceitual, Profundidade Teórica e Organização, que apresentaram incremento de aproximadamente 25 pontos percentuais em todas as ferramentas. Esse padrão demonstra que essas dimensões respondem diretamente à qualidade das instruções fornecidas, indicando que prompts estruturados favorecem respostas mais organizadas, completas e tecnicamente consistentes (Giray, 2023; Lee; Palmer, 2025).

Em contraste, a Fidelidade ao Texto permaneceu com ganho absoluto igual a zero para todas as arquiteturas. Esse resultado não representa ausência de fidelidade, mas sim ausência de variação entre os dois cenários avaliados, indicando que essa característica é determinada pela arquitetura do modelo e não pelo refinamento do prompt. De forma semelhante, o Controle de Extrapolações revelou comportamentos distintos entre as ferramentas: ChatGPT e Grok apresentaram redução de 25 pontos percentuais, refletindo maior incorporação de conhecimentos externos à medida que o comando passou a exigir maior aprofundamento científico. Já Gemini e NotebookLM mantiveram estabilidade, sendo este último favorecido pela arquitetura baseada em Retrieval-Augmented Generation (RAG), que restringe a geração de respostas ao corpus documental fornecido (Lewis et al., 2020; Francis; Jones; Smith, 2025).

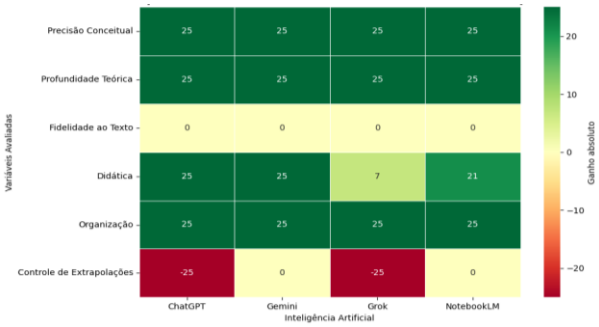

Figura 1. Ganho Absoluto (%) obtido após o refinamento do prompt.

As Figuras 2 a 5 detalham como esses ganhos se distribuíram individualmente em cada arquitetura, permitindo compreender que o refinamento do prompt não promoveu uma uniformização do comportamento das inteligências artificiais, mas potencializou características inerentes a cada modelo.

O comportamento do ChatGPT, ilustrado na Figura 2, evidencia que o refinamento do prompt promoveu uma expansão relativamente homogênea do desempenho da ferramenta. Observa-se incremento de 50% para 75% em precisão conceitual e profundidade teórica, indicando que o modelo passou de respostas predominantemente descritivas para explicações mais consistentes e fundamentadas. Os maiores ganhos ocorreram em didática e organização, que evoluíram de aproximadamente 75% para 100%, revelando elevada capacidade de reorganizar conteúdos científicos complexos em sequências lógicas, progressivas e cognitivamente acessíveis. Em contraste, a fidelidade ao texto-base permaneceu estável (75%), sugerindo que essa característica está mais relacionada à arquitetura do modelo do que ao refinamento do comando, aspecto também observado por Lee e Palmer (2025). Esses resultados corroboram estudos que apontam a engenharia de prompt como um mecanismo capaz de ampliar o rigor analítico e a estrutura argumentativa das respostas, sem necessariamente alterar o grau de aderência documental das informações (Federiakin et al., 2024; Batista; Mesquita; Carnaz, 2024).

A Figura 2 também evidencia o principal *trade-off* observado para o ChatGPT. Enquanto praticamente todas as dimensões apresentaram crescimento após a utilização do prompt acadêmico-técnico, o controle de extrapolações reduziu-se de 75% para 50%, indicando que o aumento da profundidade teórica foi acompanhado por maior liberdade interpretativa. Esse comportamento reflete a tendência do modelo em complementar o texto-base com conhecimentos

previamente internalizados, produzindo conexões e sínteses que enriquecem a compreensão, mas que podem extrapolar a organização originalmente apresentada na obra de Bucknall (1977). Assim, o refinamento do prompt potencializou a principal vocação do ChatGPT — a construção de narrativas didáticas e integradoras —, porém reforçou a necessidade de validação crítica das interpretações quando a finalidade é preservar elevada fidelidade documental, conforme discutido por Federiakin et al. (2024) e Hughes et al. (2025).

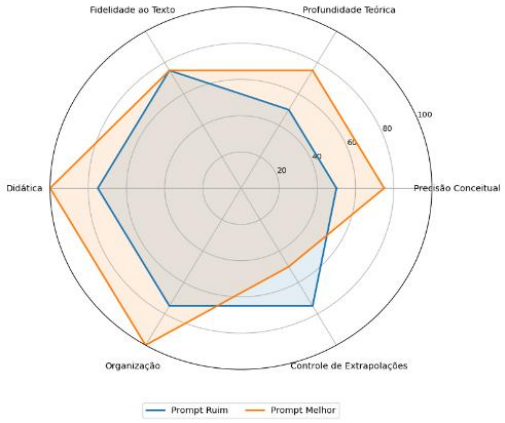

Figura 2. ChatGPT e o impacto do refinamento do prompt.

O Gemini apresentou perfil semelhante ao ChatGPT em termos de clareza didática e organização das respostas, porém com maior ênfase na construção de relações comparativas e sínteses funcionais entre os conceitos. A ferramenta destacou-se pela utilização de contrastes entre variáveis físicas, pela explicação da influência da dilatação volumétrica e pela articulação entre deformação localizada e globalização da deformação em sistemas tenacificados. Entretanto, sua abordagem permaneceu mais orientada à compreensão funcional dos mecanismos do que ao aprofundamento mecanicista, explorando menos intensamente modelos clássicos de deformação quando comparada ao Grok (BAIDOO-ANU; ANSAH, 2023; Ogunleye et al., 2024).

A evolução promovida pelo refinamento do prompt pode ser observada na Figura 3. A precisão conceitual aumentou de 50% para 75%, enquanto a profundidade teórica evoluiu de 50% para 75%, indicando maior capacidade de explicar os mecanismos físicos sem alterar sua estratégia de síntese. Os maiores incrementos ocorreram em didática e organização, que passaram de aproximadamente 75% para 100%, evidenciando que o prompt acadêmico potencializou principalmente a estrutura lógica e a qualidade da apresentação das informações. Diferentemente do ChatGPT, entretanto, o Gemini manteve inalterados tanto a fidelidade ao texto-base (50%) quanto o controle de extrapolações (50%), sugerindo que o aumento da complexidade analítica ocorreu sem recorrer de forma significativa a conhecimentos externos ao documento analisado. Esse comportamento reforça seu perfil Sintético-Comparativo, no qual a principal contribuição consiste em reorganizar e integrar o conteúdo de maneira clara e visualmente estruturada, preservando maior equilíbrio entre expansão conceitual e aderência documental (Lee; Palmer, 2025; Federiakin et al., 2024).

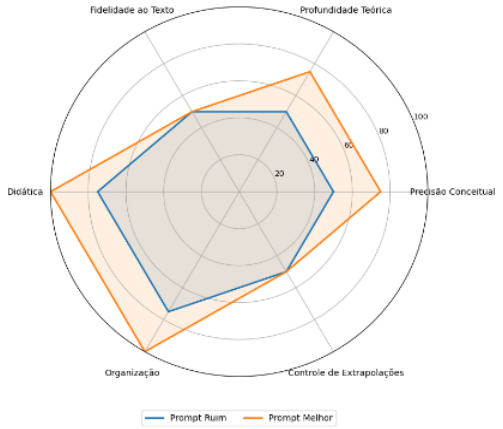

Figura 3. Gemini e o impacto do refinamento do prompt.

O NotebookLM destacou-se por apresentar o comportamento mais estável entre as ferramentas avaliadas, característica diretamente relacionada à sua arquitetura baseada em Retrieval-Augmented Generation (RAG), voltada à recuperação documental e à rastreabilidade das informações (LEWIS et al., 2020). Como observado na Figura 4, a ferramenta manteve 100% de fidelidade ao texto-base e 100% de controle de extrapolações tanto no prompt genérico quanto no prompt acadêmico-técnico, evidenciando que o refinamento do comando não modificou seu compromisso com a aderência ao documento de referência. Diferentemente das demais inteligências artificiais, o aumento da complexidade do prompt não resultou em maior liberdade interpretativa, mas em uma melhor articulação dos conceitos já presentes na obra analisada.

Os ganhos decorrentes do refinamento do prompt concentraram-se principalmente na profundidade teórica, que evoluiu de 50% para 75%, e na organização e didática, que passaram de aproximadamente 50% para 75%. A precisão conceitual também aumentou de 75% para 100%, demonstrando maior capacidade de identificar e integrar corretamente os mecanismos físicos discutidos por Bucknall (1977), sem recorrer à

incorporação de conhecimentos externos. Esse comportamento explica por que o NotebookLM apresentou elevada rastreabilidade e reduzido risco de alucinações, tornando-se particularmente adequado para atividades de revisão bibliográfica, síntese documental e validação conceitual. Em contrapartida, sua baixa tendência à expansão interpretativa limita a elaboração de discussões mais amplas quando comparado a modelos mais generativos, como ChatGPT e Grok, evidenciando o compromisso da arquitetura RAG com a preservação da integridade documental em detrimento da criatividade analítica (Lewis et al., 2020; Federiakin et al., 2024; Francis; Jones; Smith, 2025).

Figura 4. NotebookLm e o impacto do refinamento do prompt.

O Grok apresentou o maior nível de aprofundamento científico entre as ferramentas avaliadas, incorporando modelos clássicos da física dos polímeros, como as teorias de Eyring, Argon e Robertson, além de critérios de escoamento dependentes da pressão hidrostática (Bucknall, 1977). Conforme evidenciado na Figura 5, a ferramenta já apresentava desempenho elevado no prompt genérico, com aproximadamente 75% em precisão conceitual e profundidade teórica, indicando uma arquitetura naturalmente orientada ao tratamento técnico dos fenômenos físicos. Após o refinamento do prompt, esses indicadores atingiram 100%, evidenciando que comandos mais estruturados potencializaram sua capacidade de elaborar interpretações mecanicistas e integrar modelos físicos de maior complexidade (Federiakin et al., 2024; Lee; Palmer, 2025).

Em contrapartida, a Figura 5 revela que esse ganho ocorreu acompanhado de uma alteração importante no comportamento do modelo. Enquanto didática e organização evoluíram apenas de 50% para 75%, o controle de extrapolações reduziu-se de 50% para 25%, indicando maior incorporação de conhecimentos externos ao documento analisado. Diferentemente do NotebookLM, cuja evolução ocorreu preservando integralmente a aderência ao texto-base, o Grok ampliou substancialmente a densidade científica das respostas por meio da integração de referências, teorias e interpretações não explicitamente presentes na obra de Bucknall (1977). Esse comportamento confirma que o refinamento do prompt potencializa as competências inerentes à arquitetura do modelo, mas também evidencia o *trade-off* entre expansão conceitual e fidelidade documental. Assim, embora o Grok tenha produzido as respostas de maior rigor científico, sua utilização em contextos acadêmicos requer validação crítica das informações geradas, reduzindo o risco de extrapolações e alucinações conceituais inerentes aos modelos generativos (Federiakin et al., 2024; Francis; Jones; Smith, 2025; Hughes et al., 2025).

Figura 5. Grok e o impacto do refinamento do prompt.

Em conjunto, a Figura 1 e as Figuras 2 a 5 demonstram que a engenharia de prompt atua como um mecanismo de calibração metodológica, elevando o desempenho analítico das inteligências artificiais sem eliminar as diferenças estruturais entre suas arquiteturas. Assim, o refinamento do comando potencializa competências específicas de cada modelo: ChatGPT e Gemini reforçam sua vocação didática, o Grok amplia sua capacidade de aprofundamento científico e o NotebookLM preserva sua elevada confiabilidade documental, confirmando que a escolha da ferramenta deve considerar o objetivo da atividade científica e não apenas indicadores globais de desempenho (Lee; Palmer, 2025; Hughes et al., 2025).

O PERFIL CIENTÍFICO DAS ARQUITETURAS

A Tabela 4 e as Figuras 6 e 7 apresentam o desempenho das inteligências artificiais quando submetidas ao prompt acadêmico-técnico, permitindo comparar seus perfis científicos sob uma condição metodológica comum. Diferentemente da seção anterior, que analisou a influência do refinamento do prompt, esta etapa busca identificar quais características permanecem intrínsecas a cada

arquitetura mesmo após a otimização das instruções, evidenciando diferenças na forma como os modelos constroem, organizam e justificam o conhecimento científico. A análise engloba a influência da PE sobre o desempenho das ferramentas, esta etapa busca caracterizar o comportamento científico inerente a cada arquitetura quando submetida às melhores condições experimentais.

Os resultados evidenciam que, embora todas as ferramentas tenham alcançado desempenho elevado em diferentes dimensões, seus perfis científicos permaneceram distintos. A maior média geral foi observada no NotebookLM (95,83%), seguido pelo Grok (89,17%), ChatGPT (84,17%) e Gemini (76,33%), indicando que o refinamento do prompt reduz diferenças relacionadas à superficialidade das respostas, mas não elimina as particularidades estruturais de cada modelo. Esse comportamento reforça que modelos de linguagem respondem ao mesmo estímulo de acordo com suas arquiteturas de treinamento, mecanismos de recuperação de informação e estratégias de geração textual, indicando que o refinamento do prompt reduz diferenças relacionadas à superficialidade das respostas, mas não promove convergência completa entre os modelos (Federiakin et al., 2024; Lee; Palmer, 2025).

Tabela 4– Perfis Científicos das Inteligências Artificiais sob Prompt Acadêmico-Técnico

Categoria de avaliação	Chatgpt	Gemini	Grok	Noteboo klm
Precisão conceitual	75%	75%	100%	100%
Profundidad e teórica	75%	75%	100%	75%
Robustez científica	90%	80%	100%	100%
Qualidade argumentati va	90%	90%	80%	100%
Nível cognitivo (bloom)	100%	50%	67%	100%
Metacogniç ão	75%	88%	88%	100%
Média geral	84,17%	76,33%	89,17%	95,83%

A análise da Tabela 4 evidencia que nenhuma ferramenta apresentou desempenho máximo em todas as categorias avaliadas. Enquanto Grok e NotebookLM atingiram 100% em precisão conceitual e robustez científica, ChatGPT destacou-se pelo maior nível cognitivo segundo a Taxonomia de Bloom, e Gemini apresentou desempenho mais uniforme entre as categorias, embora com menor média geral. Esses resultados indicam que diferentes arquiteturas tendem a otimizar dimensões específicas do desempenho científico, em vez de apresentar superioridade absoluta.

O radar evidencia um perfil relativamente uniforme, sem oscilações acentuadas entre as categorias, a Figura 6 mostra que o ChatGPT apresentou um perfil equilibrado, destacando-se em Robustez Científica (90%), Qualidade Argumentativa (90%) e Nível Cognitivo segundo a Taxonomia de Bloom (100%). O desempenho sugere elevada capacidade de reorganizar conteúdos especializados em explicações progressivas, favorecendo processos de aprendizagem e construção conceitual. Em contrapartida, a Precisão Conceitual, a Profundidade Teórica e a Metacognição permaneceram em níveis intermediários (75%), indicando que a ferramenta privilegia a clareza expositiva em detrimento do aprofundamento mecanicista. Esse comportamento está alinhado ao potencial educacional atribuído aos LLMs em ambientes de ensino superior, nos quais a aprendizagem ativa é favorecida pela reorganização lógica do conhecimento (BAIDOO-ANU; ANSAH, 2023; Batista; Mesquita; Carnaz, 2024).

A Figura 6 mostra que o Gemini apresentou o radar mais homogêneo entre as ferramentas, indicando menor especialização em categorias específicas. Gemini apresentou o perfil científico mais homogêneo entre as ferramentas, porém com menor desempenho médio. Seus maiores resultados ocorreram em Qualidade Argumentativa (90%) e Metacognição (88%), indicando elevada capacidade de organizar informações e explicitar limitações da própria resposta. Entretanto, o Nível Cognitivo (50%) foi o menor entre todas as IAs avaliadas, sugerindo menor frequência de processos analíticos, sintéticos e avaliativos associados aos níveis superiores da Taxonomia de Bloom. Esses resultados reforçam seu perfil sintético-comparativo, no qual a organização e a comunicação dos conceitos recebem maior prioridade que o aprofundamento científico propriamente dito (Lee; Palmer, 2025).

Observa-se que o Grok apresentou o maior desempenho em Precisão Conceitual (100%), Profundidade Teórica (100%) e Robustez Científica (100%), confirmando seu perfil analítico-mecanicista. Contudo, sua Qualidade Argumentativa (80%) e seu Nível Cognitivo (67%) permaneceram inferiores aos observados para ChatGPT e NotebookLM, indicando que o elevado rigor científico nem sempre é acompanhado pela mesma eficiência na comunicação e organização pedagógica dos conceitos. Esse comportamento ilustra um dos desafios apontados pela literatura: quanto maior a liberdade interpretativa dos modelos, maior a necessidade de validação crítica por parte do pesquisador (Federiakin et al., 2024).

O NotebookLM apresentou o perfil científico mais robusto entre todas as ferramentas, alcançando 100% em Precisão Conceitual, Robustez Científica, Qualidade Argumentativa, Nível Cognitivo e Metacognição, sendo superado apenas em Profundidade Teórica (75%), variável deliberadamente limitada pela própria arquitetura baseada em RAG. O radar praticamente forma um polígono completo, evidenciando o perfil científico mais consistente observado neste estudo. Ao privilegiar a recuperação documental em detrimento da expansão conceitual, o NotebookLM preserva elevada fidelidade às fontes primárias e reduz o risco de alucinações, característica amplamente discutida na literatura recente sobre sistemas RAG (Lewis et al., 2020; Francis; Jones; Smith, 2025). Assim, sua menor profundidade não representa limitação metodológica, mas uma consequência direta da opção arquitetural por priorizar confiabilidade documental em vez de inferências externas.

A comparação conjunta da Figura 6 e da Figura 7 demonstra que as diferenças entre as arquiteturas não se concentram nas categorias de desempenho técnico mais consolidadas, mas nas dimensões associadas à complexidade cognitiva e à autorregulação das respostas. A Robustez Científica apresentou a maior média geral entre todas as categorias (92,5%), seguida pela Qualidade Argumentativa (90%), Metacognição (87,75%) e Precisão Conceitual (87,5%), indicando que, sob o prompt acadêmico-técnico, todas as ferramentas alcançaram desempenho elevado na identificação, sustentação e organização dos conceitos científicos. Em contrapartida, o Nível Cognitivo, avaliado segundo a Taxonomia de Bloom, apresentou a maior variabilidade entre as arquiteturas, com desvio-padrão de 24,94 pontos percentuais. Esse resultado sugere que a principal distinção entre os modelos não reside apenas na capacidade de identificar corretamente os conceitos, mas na forma como esses conceitos são articulados, analisados e utilizados para construir explicações científicas mais complexas. O mapa de calor apresentado na Figura 7 reforça essa interpretação ao evidenciar padrões distintos entre as arquiteturas, mesmo quando todas operam sob o mesmo prompt, demonstrando que modelos generativos tendem a privilegiar diferentes estratégias de construção do conhecimento (Federiakin et al., 2024; Lee; Palmer, 2025).

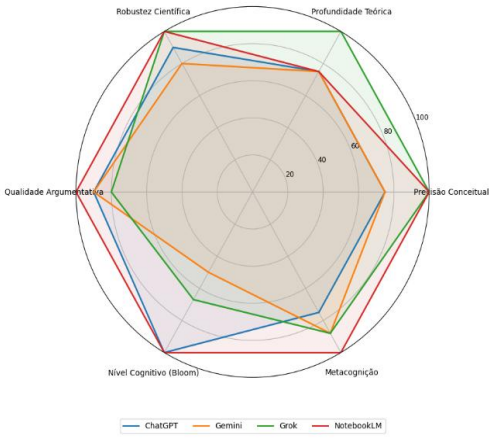

Figura 6. Perfis científicos das Ias (Prompt Acadêmico-Técnico)

O mapa de calor apresentado na Figura 7 reforça a existência de padrões científicos distintos entre as arquiteturas avaliadas, embora o NotebookLM e o Grok tenham apresentado as maiores médias gerais, com 95,83% e 89,17%, respectivamente, esses resultados foram alcançados por estratégias diferentes (Lewis et al., 2020; Bommasani et al., 2021). O NotebookLM concentrou seu desempenho nos indicadores associados à confiabilidade científica e à autorregulação, atingindo 100% em Precisão Conceitual, Robustez Científica, Qualidade Argumentativa, Nível Cognitivo e Metacognição, com menor desempenho apenas em Profundidade Teórica (75%). Esse padrão indica uma arquitetura orientada à consistência documental, rastreabilidade e explicitação dos limites da resposta (Lewis et al., 2020; Gao et al., 2023). O Grok, por outro lado, atingiu 100% em Precisão Conceitual, Profundidade Teórica e Robustez Científica, mas apresentou valores inferiores em Qualidade Argumentativa (80%) e Nível Cognitivo (67%), evidenciando um perfil mais voltado ao aprofundamento mecanicista do que à organização pedagógica ou metacognitiva da explicação (Wei et al., 2022; Federiakin et al., 2024).

Assim, médias elevadas não representam comportamentos equivalentes. Enquanto o NotebookLM constrói conhecimento por meio da ancoragem documental e da metacognição, o Grok opera pela expansão científica e pela incorporação de modelos teóricos mais complexos. O mapa de calor torna visível esse contraste ao mostrar que o desempenho superior de cada ferramenta se concentra em regiões distintas da matriz. Dessa forma, a análise não permite concluir que uma arquitetura seja universalmente superior à outra, mas que cada uma apresenta competências científicas específicas, adequadas a diferentes objetivos de pesquisa e ensino (NIST, 2023; Bommasani et al., 2021).

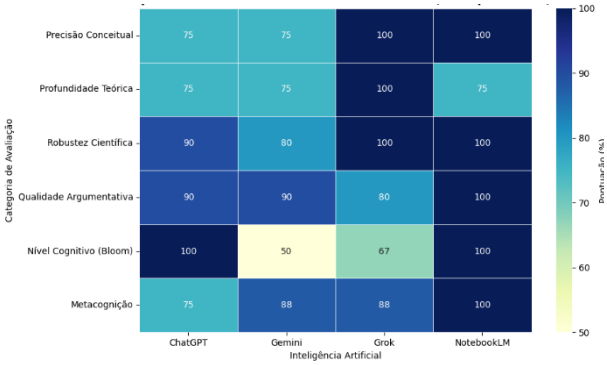

Figura 7. Mapa de calor dos Perfis científicos das Ias (Prompt Acadêmico-Técnico)

Revisão bibliográfica	★★★★☆	★★★☆☆	★★★★☆	★★★★★
Pesquisa científica	★★★★☆	★★★☆☆	★★★★★	★★★★★
Escrita de artigos	★★★★★	★★★★☆	★★★★☆	★★★★★
Geração de hipóteses	★★★☆☆	★★★☆☆	★★★★★	★☆☆☆☆
Consulta documental	★★★☆☆	★★★☆☆	★★☆☆☆	★★★★★
Explicação para iniciantes	★★★★★	★★★★★	★★☆☆☆	★★★★☆

Nota: A classificação qualitativa (★☆☆☆☆ a ★★★★★) foi elaborada a partir da integração dos indicadores quantitativos de precisão conceitual, profundidade teórica, fidelidade documental, robustez científica, qualidade argumentativa, organização, didática, rastreabilidade, metacognição e expansão conceitual, avaliados após a aplicação do prompt acadêmico-técnico. As classificações representam uma interpretação comparativa dos resultados obtidos neste estudo e não constituem uma avaliação universal das ferramentas. A adequação foi inferida a partir da integração dos indicadores quantitativos obtidos nas análises anteriores.

TRADE-OFF EPISTEMOLÓGICO

A Tabela 5 sintetiza os resultados quantitativos obtidos ao longo da pesquisa em uma perspectiva aplicada, relacionando o desempenho das inteligências artificiais aos diferentes contextos de utilização acadêmica. Em vez de estabelecer uma hierarquia entre as ferramentas, a análise evidencia que cada arquitetura apresenta potencialidades específicas, tornando sua adequação dependente do objetivo da atividade científica. Esse resultado reforça que o desempenho das LLMs deve ser interpretado à luz de suas características epistemológicas e não apenas por indicadores absolutos de desempenho (Lee; Palmer, 2025; Federiakin et al., 2024).

Observa-se que ChatGPT e Gemini apresentaram maior adequação para atividades de ensino e aprendizagem, sobretudo pela elevada organização das respostas, qualidade argumentativa e estrutura didática. Esses atributos favorecem a construção progressiva do conhecimento, aproximando suas respostas de materiais educacionais destinados à mediação pedagógica (BAIDOO-ANU; ANSAH, 2023). O ChatGPT destacou-se ainda pelo maior nível cognitivo segundo a Taxonomia de Bloom, evidenciando maior capacidade de articular conceitos, estabelecer relações entre mecanismos e produzir explicações em níveis superiores de abstração. O NotebookLM apresentou o perfil mais consistente para aplicações que exigem elevado rigor documental, como revisões bibliográficas, fichamentos técnicos e análise de documentos científicos, onde prioriza respostas fundamentadas exclusivamente nas fontes disponibilizadas pelo usuário (Lewis et al., 2020).

Tabela 5 – Adequação das arquiteturas de IA a diferentes aplicações científicas

Aplicação	*ChatGPT*	*Gemini*	*Grok*	*NotebookLM*
Estudo individual	★★★★★	★★★★★	★★★☆☆	★★★★☆
Ensino superior	★★★★★	★★★★★	★★★☆☆	★★★★☆

A Tabela 5 traduz os indicadores quantitativos obtidos nas análises anteriores em uma perspectiva aplicada, relacionando os perfis científicos das inteligências artificiais aos principais contextos de utilização acadêmica. Em vez de estabelecer uma classificação absoluta entre as ferramentas, a análise evidencia que cada arquitetura apresenta vantagens específicas, cuja adequação depende do objetivo da atividade científica. Dessa forma, o desempenho das LLMs deve ser interpretado à luz de seus perfis epistemológicos, e não apenas de médias globais de desempenho (Lee; Palmer, 2025; Federiakin et al., 2024).

A análise dos perfis epistemológicos delineados nas Figuras 8 e 9 corrobora a existência de uma tensão dialética fundamental na interação humano-computador: o compromisso analítico (*trade-off*) entre a fidelidade documental e a expansão científica. Este fenômeno sugere que a arquitetura do sistema dita não apenas a acurácia, mas a própria natureza da produção intelectual gerada, onde o rigor na reconstrução de fontes primárias frequentemente atua como um limitador da criatividade teórica, e vice-versa (Lee; Palmer, 2025; Dell'acqua et al., 2026).

Os resultados confirmam que o NotebookLM ocupa o auge da fidelidade documental, apresentando índices nulos de expansão conceitual externa. Esse comportamento é característico de sistemas de RAG, projetados especificamente para mitigar a alucinação científica ao restringir o espaço de busca ao *corpus*

fornecido (Francis; Jones; Smith, 2025). Em oposição, o Grok manifesta um perfil alta porosidade, integrando modelos complexos para explicação do que se buscava na pesquisa. Embora essa expansão eleve a profundidade teórica, ela fragiliza o controle documental, inserindo referências externas que podem obnubilar a origem da evidência técnica, desafiando a transparência exigida no rigor acadêmico (Federiakin et al., 2024; Hughes et al., 2025).

Essa divergência reflete a fronteira tecnológica irregular descrita por Dell'Acqua et al. (2026), na qual modelos com maior liberdade interpretativa tendem a falhar em tarefas de ancoragem estrita, enquanto ferramentas ancoradas podem carecer de flaire investigativo para a formulação de hipóteses inéditas. Ferramentas como o ChatGPT e o Gemini situam-se em uma zona de equilíbrio, funcionando como agentes didáticos que reorganizam o conhecimento técnico em narrativas acessíveis, tornando-os ideais para o suporte à aprendizagem personalizada e funcional (BAIDOO-ANU; ANSAH, 2023; Batista; Mesquita; Carnaz, 2024).

Um achado central desta investigação reside na categoria de metacognição, que emergiu como a variável de maior sensibilidade ao refinamento do prompt. A transição de respostas assertivas e sem ressalvas no cenário de baixo estímulo para declarações explícitas de limitações documentais sob prompt técnico indica que a metacognição atua como um componente de governança da ferramenta (Hughes et al., 2025; Shah, 2024). Quando o sistema reconhece as fronteiras de seu conhecimento e recomenda validações humanas, ele deixa de operar como um "papagaio estocástico" para assumir o papel de colaborador reflexivo (Bender et al., 2021; Francis; Jones; Smith, 2025).

Esta capacidade de autorregulação é essencial para manter o humano na alça de controle (*human-in-the-loop*), mitigando o risco de aceitação acrítica das saídas sintéticas (Mollick; Mollick, 2023; Shah, 2024). Sob a ótica da ética e da integridade, a metacognição funciona como uma infraestrutura invisível que promove a rastreabilidade e a *accountability*, permitindo que o pesquisador discirna entre a evidência presente no texto e a inferência probabilística do modelo (Hughes et al., 2025).

As Figuras 8 e 9 evidencia que o principal trade-off identificado nesta pesquisa ocorre entre fidelidade documental e expansão científica. Embora todas as ferramentas apresentem níveis elevados de metacognição e rastreabilidade, observa-se uma distribuição antagônica entre a capacidade de permanecer estritamente ancorada ao documento de referência e a tendência de expandir teoricamente o conhecimento produzido.

O mapa de calor (Figura 9) evidencia esse comportamento de forma quantitativa. O NotebookLM apresentou desempenho máximo em fidelidade ao texto (100%), controle de extrapolações (100%), rastreabilidade (100%) e metacognição (100%), porém registrou ausência de expansão conceitual (0%), confirmando sua orientação para reconstrução documental baseada em Retrieval-Augmented Generation (Lewis et al., 2020). Em sentido oposto, o Grok atingiu os maiores valores em profundidade teórica (100%) e expansão conceitual (100%), mas apresentou o menor controle de extrapolações (25%), evidenciando maior liberdade interpretativa durante a construção das respostas. ChatGPT e Gemini ocuparam posição intermediária, equilibrando fidelidade documental (75% e 50%), rastreabilidade (88%) e expansão conceitual (50%), caracterizando perfis híbridos entre reconstrução documental e geração científica.

Esse comportamento confirma que arquiteturas distintas privilegiam estratégias epistemológicas diferentes. Enquanto sistemas baseados em recuperação documental tendem a maximizar confiabilidade e rastreabilidade, modelos predominantemente generativos favorecem expansão conceitual e integração de conhecimentos externos, exigindo maior supervisão crítica por parte do pesquisador (Lewis et al., 2020; NIST, 2023; Federiakin et al., 2024).

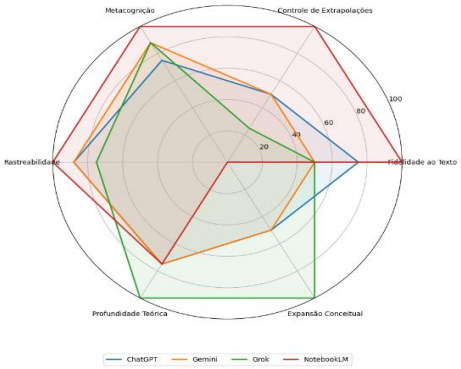

Figura 8. Comparação dos perfis de fidelidade documental e expansão científica das inteligências artificiais sob o prompt acadêmico-técnico.

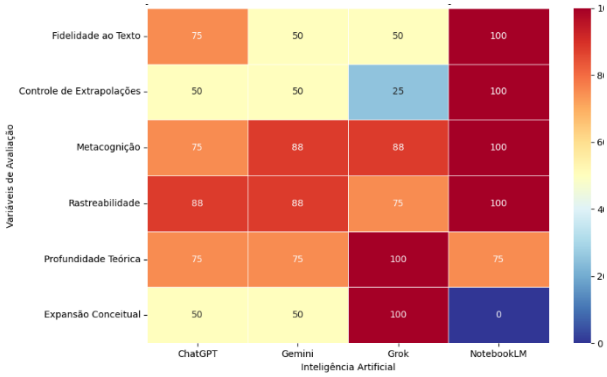

Figura 9. Distribuição da intensidade dos indicadores de fidelidade documental e expansão científica (Heatmap).

O modelo conceitual da Figura 10 o esquema demonstra que o refinamento do comando não uniformiza o comportamento das inteligências artificiais, mas potencializa competências inerentes às respectivas arquiteturas. Como consequência, observa-se a formação de dois polos epistemológicos complementares: um orientado à expansão científica, representado principalmente pelo Grok, caracterizado por elevada profundidade teórica e expansão conceitual; e outro orientado à fidelidade documental, representado pelo NotebookLM, cuja prioridade reside na rastreabilidade, no controle de extrapolações e na preservação das fontes primárias. Entre esses extremos posicionam-se ChatGPT e Gemini, que combinam reorganização didática, argumentação e acessibilidade do conhecimento, assumindo papel intermediário entre criatividade científica e segurança documental.

Esse modelo evidencia que a engenharia de prompt atua como elemento modulador da interação entre usuário e inteligência artificial, direcionando o potencial analítico de cada arquitetura sem alterar sua natureza epistemológica. Assim, a seleção da ferramenta deixa de depender exclusivamente de desempenho global e passa a considerar a finalidade da atividade científica, reforçando princípios de governança, transparência e supervisão humana recomendados pelo AI Risk Management Framework (NIST, 2023), pela Recomendação da UNESCO sobre Ética da Inteligência Artificial (UNESCO, 2021) e pelos Princípios de IA da OCDE (OECD, 2019).

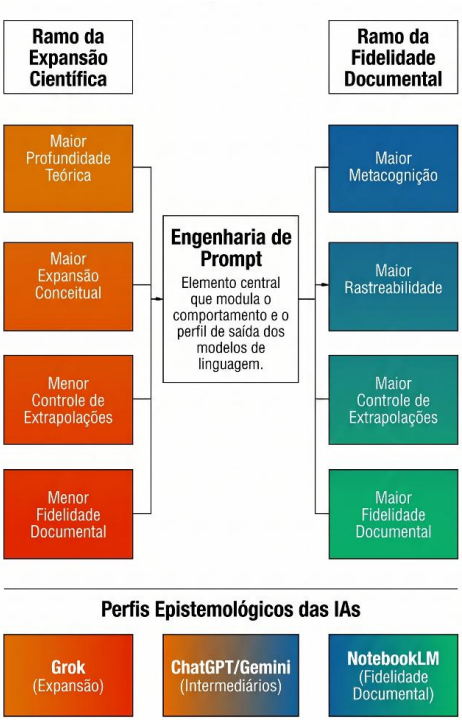

Figura 10. Modelo conceitual da influência da engenharia de prompt na construção dos perfis epistemológicos das inteligências artificiais generativas.

CONCLUSÃO

Este estudo demonstrou que a engenharia de prompt exerce influência sobre o desempenho das inteligências artificiais generativas em tarefas acadêmico-científicas. O refinamento dos comandos promoveu ganhos consistentes em precisão conceitual, profundidade teórica, organização e qualidade argumentativa. Entretanto, verificou-se que o prompt não modifica o perfil epistemológico das ferramentas, atuando como um mecanismo que potencializa competências inerentes às diferentes arquiteturas. A análise comparativa evidenciou perfis distintos entre os modelos avaliados. O NotebookLM destacou-se pela elevada fidelidade documental, rastreabilidade e controle de extrapolações, mostrando maior adequação para revisões bibliográficas e análises fundamentadas em fontes primárias. O Grok apresentou maior profundidade teórica e expansão conceitual, favorecendo atividades exploratórias e geração de hipóteses, enquanto ChatGPT e Gemini demonstraram melhor desempenho em organização didática, qualidade argumentativa e mediação do processo de aprendizagem. Como principal contribuição, o estudo propõe uma matriz de codificação capaz de caracterizar o comportamento científico de diferentes arquiteturas de IA, além de apresentar um modelo conceitual que sintetiza o trade-off entre fidelidade documental e expansão científica. Os resultados reforçam que a escolha da ferramenta

deve considerar os objetivos da atividade acadêmica e que a IA deve atuar como instrumento de apoio ao pesquisador, permanecendo subordinada à validação crítica, à supervisão humana e aos princípios de rigor e integridade científica.

REFERÊNCIAS

BAIDOO-ANU, David; ANSAH, Leticia Owusu. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. Journal of AI, v. 7, n. 1, p. 52-62, 2023. Disponível em: https://doi.org/10.61969/jai.1337500. Acesso em: 24 jun. 2026.

BATISTA, João; MESQUITA, Anabela; CARNAZ, Gonçalo. Generative AI and Higher Education: Trends, Challenges, and Future Directions from a Systematic Literature Review. Information, v. 15, n. 11, 676, 2024. Disponível em: https://doi.org/10.3390/info15110676. Acesso em: 24 jun. 2026.

BENDER, E. M. et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of ACM FAccT, 2021.

BOMMASANI, R. et al. On the Opportunities and Risks of Foundation Models. Stanford CRFM, 2021.

BUCKNALL, C. B. Toughened Plastics. Londres: Applied Science Publishers, 1977.

DELL'ACQUA, Fabrizio et al. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School Working Paper 24-013, set. 2023. (Publicado em Organization Science, 2026). Disponível em: https://doi.org/10.1287/orsc.2025.21838. Acesso em: 24 jun. 2026.

FEDERIAKIN, Denis; MOLEROV, Dimitri; ZLATKIN-TROITSCHANSKAIA, Olga; MAUR, Andreas. Prompt engineering as a new 21st century skill. Frontiers in Education, v. 9, 2024. Disponível em: https://doi.org/10.3389/feduc.2024.1366434. Acesso em: 24 jun. 2026.

FRANCIS, Nigel J.; JONES, Sue; SMITH, David P. Generative AI in Higher Education: Balancing Innovation and Integrity. British Journal of Biomedical Science, v. 81, n. 14048, 2025. Disponível em: https://doi.org/10.3389/bjbs.2024.14048. Acesso em: 24 jun. 2026.

GAO, Y. et al. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997, 2023.

GIRAY, Louie. Prompt Engineering with ChatGPT: A Guide for Academic Writers. Annals of Biomedical Engineering, v. 51, p. 2629–2633, 2023. Disponível em: https://doi.org/10.1007/s10439-023-03272-4. Acesso em: 24 jun. 2026.

HUGHES, Laurie et al. Reimagining Higher Education: Navigating the Challenges of Generative AI Adoption. Information Systems Frontiers, 2025. Disponível em: https://doi.org/10.1007/s10796-025-10535-5. Acesso em: 24 jun. 2026.

LEE, Daniel; PALMER, Edward. Prompt engineering in higher education: a systematic review to help inform curricula. International Journal of Educational Technology in Higher Education, v. 22, n. 7, 2025. Disponível em: https://doi.org/10.1186/s41239-025-00503-7. Acesso em: 24 jun. 2026.

LEWIS, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020.

MOLLICK, Ethan; MOLLICK, Lilach. Assigning AI: Seven Approaches for Students with Prompts. Philadelphia: Wharton School of the University of Pennsylvania, 2023. Disponível em: https://ssrn.com/abstract=4475995. Acesso em: 24 jun. 2026.

NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg: National Institute of Standards and Technology, 2023.

OECD. Recommendation of the Council on Artificial Intelligence. Paris: OECD Publishing, 2019.

OGUNLEYE, Bayode et al. Higher education assessment practice in the era of generative AI tools. Journal of Applied Learning & Teaching, v. 7, n. 1, 2024. Disponível em: https://doi.org/10.37074/jalt.2024.7.1.28. Acesso em: 24 jun. 2026.

SHAH, C. From prompt engineering to prompt science with human in the loop. arXiv, p. 1-7, 2024. Disponível em: https://doi.org/10.48550/arXiv.2401.04122.

UNESCO. Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO, 2021.

WEI, J. et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS, 2022.