UNIVERSIDADE FEDERAL DO MARANHÃO - UFMA
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIAS – CCET
COORDENAÇÃO DO CURSO DE ENGENHARIA DA COMPUTAÇÃO

STEFHANE PEREIRA COSTA

ANÁLISE COMPARATIVA DE INTELIGÊNCIAS ARTIFICIAIS GENERATIVAS NA INTERPRETAÇÃO DE PROJETOS PEDAGÓGICOS DOS CURSOS DE ENGENHARIAS VINCULADOS AO BICT: um estudo quantitativo e qualitativo

SÃO LUIS – MA
2026

STEFHANE PEREIRA COSTA

ANÁLISE COMPARATIVA DE INTELIGÊNCIAS ARTIFICIAIS GENERATIVAS NA INTERPRETAÇÃO DE PROJETOS PEDAGÓGICOS DOS CURSOS DE ENGENHARIAS VINCULADOS AO BICT: um estudo quantitativo e qualitativo

Trabalho de Conclusão de Curso II (TCC II) apresentado para obtenção do título de Bacharel em Engenharia da Computação, pela Universidade Federal do Maranhão - UFMA, Cidade Universitária: Campus Dom Delgado, São Luís, MA.

Orientador (a): Prof (a). Dr (a). Alex Oliveira Barradas Filho

São Luís – Ma, 20 de Julho de 2026.

BANCA EXAMINADORA

Documento assinado digitalmente
ALEX OLIVEIRA BARRADAS FILHO
Data: 24/07/2026 18:29:19-0300
Verifique em https://validar.iti.gov.br

_____

Prof. (Alex Oliveira Barradas Filho)
Afiliações

Documento assinado digitalmente
CLAUDIO MANOEL PEREIRA AROUCHA
Data: 21/07/2026 10:02:35-0300
Verifique em https://validar.iti.gov.br

_____

Prof. (Claudio Manoel Pereira Aroucha)
Afiliações

Documento assinado digitalmente
DAVI VIANA DOS SANTOS
Data: 21/07/2026 09:40:11-0300
Verifique em https://validar.iti.gov.br

_____

Prof. (Davi Viana dos Santos)
Afiliações

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Pereira Costa, Stefhane.
ANÁLISE COMPARATIVA DE INTELIGÊNCIAS ARTIFICIAIS GENERATIVAS NA INTERPRETAÇÃO DE PROJETOS PEDAGÓGICOS DOS CURSOS DE ENGENHARIAS VINCULADOS AO BICT : um estudo quantitativo e qualitativo / Stefhane Pereira Costa. - 2026.
54 p.

Orientador(a): Alex Oliveira Barradas Filho.
Curso de Engenharia da Computação, Universidade Federal do Maranhão, São Luis, 2026.

1. Ia Generativa. 2. Projetos Pedagógicos de Curso. 3. Gemini 2.5 Flash. 4. Llama-3.3 70b Versatile. 5. Mistral-small-latest. I. Oliveira Barradas Filho, Alex. II. Título.

Resumo

O crescimento das inteligências artificiais generativas tem ampliado as possibilidades de análise e interpretação de documentos complexos no contexto da educação superior. Neste trabalho, foi realizada uma comparação entre os modelos Gemini 2.5 Flash, LlaMA-3.3 70B versatile e mistral-small-latest quanto à capacidade de interpretar Projetos Pedagógicos de Curso (PPCs) de cinco cursos de Engenharia vinculados ao Bacharelado Interdisciplinar em Ciência e Tecnologia (BICT) da Universidade Federal do Maranhão. A pesquisa adotou uma abordagem qualitativa, utilizando uma arquitetura baseada em Retrieval-Augmented Generation (RAG), um protótipo web desenvolvido para envio padronizado de perguntas aos modelos e uma rubrica analítica composta pelos critérios de clareza, aderência ao documento, coerência, profundidade e fidelidade das informações. Ao todo, foram avaliadas 150 respostas geradas pelos modelos. Os resultados indicaram que o Gemini 2.5 Flash apresentou o desempenho mais consistente, seguido pelo mistral-small-latest, enquanto o LlaMA-3.3 70b versatile obteve menor desempenho, principalmente nos critérios de profundidade e aderência ao conteúdo dos PPCs. Também foi observado que todos os modelos produziram respostas claras e coerentes, mas ainda apresentaram limitações na interpretação de informações implícitas e na elaboração de análises mais aprofundadas. A análise realizada evidencia a potência que as inteligências artificiais generativas possuem para apoiar a interpretação de documentos acadêmicos, desde que seus resultados sejam utilizados com supervisão humana, contribuindo para aplicações mais seguras e eficientes no ambiente educacional

Palavras-chave: IA Generativa; Projetos Pedagógicos de Curso; Gemini 2.5 Flash; LlaMA-3.3 70B versatile; mistral-small-latest

Abstract

The growth of generative artificial intelligence has expanded the possibilities for analyzing and interpreting complex documents within the context of higher education. This study compared the Gemini 2.5 Flash, LlaMA-3.3 70B Versatile, and Mistral-Small-Latest models regarding their ability to interpret Course Pedagogical Projects (PPCs) from five engineering programs linked to the Interdisciplinary Bachelor's Degree in Science and Technology (BICT) at the Federal University of Maranhão. The research adopted a qualitative approach, utilizing an architecture based on Retrieval-Augmented Generation (RAG), a web prototype developed for the standardized submission of questions to the models, and an analytical rubric comprising criteria for clarity, adherence to the document, coherence, depth, and information fidelity. A total of 150 model-generated responses were evaluated. The results indicated that Gemini 2.5 Flash demonstrated the most consistent performance, followed by Mistral-Small-Latest, while LlaMA-3.3 70B Versatile performed less effectively, particularly regarding depth and adherence to the PPC content. It was also observed that, while all models produced clear and coherent responses, they still exhibited limitations in interpreting implicit information and generating in-depth analyses. The analysis highlights the potential of generative AI to support the interpretation of academic documents—provided the results are used with human oversight—thereby contributing to safer and more efficient applications within the educational environment.

Keywords: Generative AI; Course Pedagogical Projects; Gemini 2.5 Flash; LlaMA-3.3 70B Versatile; Mistral-Small-Latest

SUMÁRIO

1 INTRODUÇÃO

O crescente uso da inteligência artificial (IA) em diversos setores da sociedade tem impulsionado mudanças significativas na forma como se produz, acessa e interpreta o conhecimento. De acordo com Rich e Knight (1994 apud Oliveira *et al*., 2023), a IA pode ser compreendida como "a ciência por trás da criação de máquinas inteligentes" ou ainda como "o estudo de como fazer com que os computadores executem tarefas que, atualmente, os seres humanos realizam com mais habilidade".

Dentre os avanços mais recentes no campo da inteligência artificial, destacam-se as IA generativas, capazes de produzir textos, imagens, códigos e outros conteúdos a partir de instruções em linguagem natural, os chamados *prompts.* De acordo com Sampaio, Sabbatini e Limongi (2024), essas ferramentas permitem a automação de diversas atividades, como a elaboração de textos, resumos, síntese de informações, organização textual, tradução, preenchimento de formulários e redação de conteúdos padronizados, tornando-se recursos acessíveis até mesmo para usuários sem formação técnica especializada.

Nesse contexto, destaca-se a possibilidade de utilização da IA generativa na interação com documentos institucionais, como os Projetos Pedagógicos de Curso (PPC). O PPC “[...]representa o planejamento e a organização do curso, sendo insumo formal e estruturante da oferta de serviço de ensino.” (INEP, 2018). Contudo, por sua linguagem técnica e detalhada, pode apresentar dificuldades de compreensão para discentes e até mesmo para docentes e gestores.

Apesar dos avanços dos modelos de linguagem e da utilização crescente da arquitetura Retrieval-Augmented Generation (RAG) para reduzir alucinações (Lewis *et al*., 2020), ainda são poucos os estudos que avaliam o desempenho desses modelos na interpretação de documentos acadêmicos. Chang *et al*. (2024) apontam que a maior parte das pesquisas se concentra em tarefas gerais de Processamento de Linguagem Natural, enquanto aplicações em documentos especializados ainda são menos exploradas. Além disso, Silva e Paula (2024) destacam que o uso da Inteligência Artificial em pesquisas qualitativas ainda é um campo em desenvolvimento, reforçando a necessidade de estudos que investiguem sua aplicação em diferentes contextos. Diante desse cenário, esta pesquisa busca comparar o desempenho de diferentes modelos de linguagem na interpretação de Projetos Pedagógicos de Curso utilizando uma arquitetura baseada em Retrieval-

Augmented Generation (RAG), contribuindo para ampliar o conhecimento sobre a aplicação dessas tecnologias em documentos acadêmicos. Especificamente, o estudo analisa o comportamento desses modelos diante de documentos educacionais estruturados, como os PPCs, que, nos cursos de Engenharia, especialmente aqueles vinculados ao Bacharelado Interdisciplinar em Ciência e Tecnologia (BICT), desempenham papel fundamental na comunicação das diretrizes curriculares, das abordagens interdisciplinares e dos objetivos formativos. Ao avaliar o desempenho desses modelos diante desse tipo de material, busca-se compreender seus potenciais e limitações, contribuindo para o debate sobre o uso da IA generativa como ferramenta de apoio à interpretação e à compreensão de conteúdos pedagógicos complexos no contexto educacional.

1.1 Justificativa

Os Projetos Pedagógicos de Curso (PPCs) são documentos essenciais para a organização curricular dos cursos de ensino superior, pois estruturam objetivos formativos, diretrizes e matriz curricular. Embora fundamentais para a organização curricular das instituições de ensino superior, muitas vezes apresentam uma linguagem mais técnica, extensa e de difícil compreensão para alunos, professores e gestores.

Nesse contexto, o avanço da Inteligência artificial, especialmente os modelos generativos de linguagem natural, tem impulsionado novas formas de interação com documentos mais complexos e estruturados. Ferramentas como LLaMa, Gemini e Mistral, vêm se mostrando capazes de interpretar linguagem natural, sintetizar informações e responder perguntas subjetivas de forma coerente. Porém, ainda são pouco exploradas as aplicações dessas ferramentas na área educacional com foco em interpretação de documentos institucionais.

Dessa maneira, torna-se relevante investigar como diferentes Inteligências Artificiais generativas interpretam conteúdos pedagógicos formais, como os PPCs, verificando se essas ferramentas são capazes de apresentar respostas coerentes, claras e compatíveis com as informações presentes nos documentos. Espera-se que os resultados contribuam para ampliar a compreensão sobre as potencialidades e limitações desses modelos na interpretação de documentos acadêmicos, oferecendo subsídios para futuras pesquisas e aplicações no contexto da educação superior.

1.2 Objetivos

1.2.1 Geral

Comparar o desempenho dos modelos Gemini-2.5-Flash, LlaMA-3.3 70B versatile e mistral-small-latest na resposta a perguntas subjetivas baseadas nos Projetos Pedagógicos de Curso (PPCs) de cinco cursos de Engenharia, avaliando as respostas segundo os critérios de Clareza, Aderência, Coerência, Profundidade e Fidelidade às informações dos documentos.

1.2.2 Específicos

- Formular um conjunto de perguntas subjetivas relacionadas ao conteúdo dos PPCs;
- Desenvolver uma interface web para integrar os três modelos de linguagem e permitir o envio padronizado das perguntas e do contexto recuperado dos PPCs.;
- Obter respostas dos três modelos de IA generativa com base no contexto recuperado dos PPCs;
- Avaliar as respostas utilizando os critérios de Clareza, Aderência, Coerência, Profundidade e Fidelidade, comparando o desempenho dos modelos;
- Identificar padrões, vantagens e limitações do uso de modelos de IA generativa na interpretação de documentos pedagógicos.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Inteligência artificial e sua evolução

Segundo Müller (2023), o termo “Inteligência Artificial” foi utilizado pela primeira vez em 1956, durante um evento realizado pelo Dartmouth College, nos Estados Unidos. Nesse evento, dez cientistas foram convocados para participar do “Projeto de Pesquisa de Verão de Dartmouth sobre Inteligência Artificial”, com o objetivo de promover avanços significativos na área. Atualmente, a IA apresenta-se como uma das áreas de maior crescimento. A evolução da cultura digital ao longo do tempo possibilitou o seu surgimento e a sua consolidação.

A Inteligência Artificial (IA) é caracterizada pela “[...]capacidade dos dispositivos eletrônicos de operarem de forma análoga ao pensamento humano, o que inclui a habilidade de compreender variáveis, tomar decisões e solucionar problemas.” (Silva *et al.*, 2023). Strykker e Kavlakoglu (2024), descrevem a IA como sendo “[...]uma tecnologia que permite que computadores e máquinas simulem o aprendizado, a compreensão, a resolução de problemas, a tomada de decisões, a criatividade e a autonomia dos seres humanos.”

Segundo Pereira, Bonin e Soares (2024), a IA passou a ocupar um papel de destaque na sociedade contemporânea, especialmente com a popularização recente dos grandes modelos de linguagem (LLMs), como o ChatGPT, da OpenAI; o Gemini, da Google; e o Claude, da Anthropic.

Entre as diversas áreas em que a IA pode produzir impactos significativos a longo prazo, a educação destaca-se como um dos campos mais discutidos. De acordo com Clarke e Whittlestone (2022 apud Pereira, Bonin e Soares, 2024), esse setor apresenta amplo potencial para a aplicação dessas tecnologias, tanto no apoio aos processos de ensino e aprendizagem quanto na gestão e análise de informações educacionais.

2.2 Inteligências Artificiais Generativas e Grandes Modelos de Linguagem (LLMs)

“De forma simples, IA generativa é uma divisão da inteligência artificial que se concentra na criação de novos dados a partir de padrões aprendidos a partir de grandes bases de dados.” (Carvalho, 2024). Diversas plataformas baseadas em IA

generativa se destacaram, cada uma com características e aplicações específicas. Entre elas, destacam-se o Gemini, LLaMa e Mistral, ferramentas amplamente utilizadas por sua versatilidade e potencial de aplicação em múltiplas áreas, incluindo a educação e a engenharia.

2.2.1 Gemini

"O Gemini é a resposta do Google à ampla adoção das IA generativas, enfrentando a crescente utilização dessas tecnologias." (Carvalho, 2024). Essa ferramenta adota um modelo que combina a compreensão da linguagem natural com a geração de conteúdo, possuindo uma grande habilidade de fazer a integração de diferentes formatos de dados, como imagens, textos e outros em uma única estrutura.

O Google (2024 apud Carvalho, 2024) destaca que um de seus principais diferenciais está relacionado com uma integração fluida com os diversos serviços que o Google oferece, possibilitando uma adaptação às várias necessidades dos usuários.

2.2.2 LlaMa

Outra ferramenta relevante no cenário das inteligências artificiais generativas é a LLaMa, criado pela Meta (Facebook). "Sua principal finalidade não é ser um chatbot público**,** mas sim um conjunto de modelos de código aberto disponível para todos que solicitarem acesso." (Sánchez, 2024).

Segundo Touvron *et al*. (2023 apud Silva 2025) os modelos desenvolvidos pela Meta priorizam a eficiência e a capacidade de escalonamento, refletindo os mais recentes avanços da empresa em modelos de linguagem de larga escala e apresentando desempenho satisfatório em uma ampla gama de tarefas.

2.2.3 Mistral AI

Outra solução que tem se destacado no campo das IA generativas é a Mistral. Essa empresa surgiu "com a intenção de criar modelos de inteligência artificial que oferecem alto desempenho com baixo custo de recursos, proporcionando eficiência em cada um de seus desenvolvimentos." (Sánchez, 2024). Fundada em 2023 por ex-pesquisadores da Meta Platforms e da Google DeepMind, a Mistral AI concentra seus esforços no desenvolvimento de modelos de linguagem de código aberto. Seu

primeiro modelo, o Mistral-7B, possui aproximadamente 7,3 bilhões de parâmetros e foi otimizado para aplicações de geração e compreensão de texto. Além disso, o modelo apresentou desempenho superior ao LLaMA 2 13B nos benchmarks avaliados, demonstrando que modelos menores podem alcançar resultados competitivos mesmo demandando menos recursos computacionais (Jiang et al., 2023).

2.3 Uso da IA na educação e Projeto Pedagógico de Curso

A crescente integração da Inteligência Artificial (IA) no campo educacional tem provocado grandes mudanças no cenário atual. A sua aplicação no âmbito educacional representa um campo de pesquisa interdisciplinar, que une conhecimentos da ciência da computação e das ciências da aprendizagem. CIEB (2019) cita que um dos principais objetivos dessa abordagem é promover a criação de ambientes de aprendizagem altamente adaptativos, ou seja, a metodologia de ensino e o ritmo são ajustados de acordo com a necessidade individual de cada aluno.

Cardoso *et al*. (2023) relatam que a IA tem potencial para transformar, de forma significativa, a educação. Entre suas principais contribuições destacam-se a personalização da aprendizagem, por meio da adaptação de conteúdos, do fornecimento de feedback imediato e da elaboração de planos de estudo conforme as necessidades de cada aluno, tornando o processo de ensino mais adaptativo e individualizado; a automação de tarefas repetitivas; a ampliação do acesso à educação, permitindo que os conteúdos estejam disponíveis a qualquer hora e em diferentes locais; e a análise de grandes volumes de dados educacionais, possibilitando a identificação de padrões e tendências que contribuem para a melhoria da qualidade do ensino.

Além dos benefícios amplos que a IA traz para a educação, um campo importante de aplicação envolve a utilização dessas tecnologias para a análise e interpretação de documentos acadêmicos institucionais, como os Projetos Pedagógicos de Curso (PPCs). Conforme a Coordenadoria de Projetos e Acompanhamento Curricular (COPAC, 2022), "o Projeto Pedagógico de um Curso de Graduação é o documento que expressa a sua identidade", apresentada à comunidade acadêmica e à sociedade as características, a estrutura e a organização do curso, de acordo com suas escolhas, trajetórias e objetivos formativos. Em razão

dessa importância, torna-se necessário compreender de forma mais detalhada sua finalidade e seus principais elementos constituintes.

O Projeto Pedagógico de Curso (PPC) é o principal documento orientador da organização acadêmica e pedagógica de um curso de graduação. Segundo Medeiros (2024), trata-se de um "instrumento de gestão de natureza acadêmica que, com base nas Diretrizes Curriculares Nacionais – DCNs e demais normativas [...] orienta o currículo para o perfil do egresso/profissional desejado". Complementando essa perspectiva, a Universidade Federal de Mato Grosso do Sul – UFMS (2021) define o PPC como um "instrumento político, filosófico e teórico-metodológico que norteia as práticas pedagógicas no curso, considerando sua trajetória histórica, inserção regional, vocação, missão, visão e objetivos".

Quanto à sua estrutura, o Projeto Pedagógico de Curso reúne os elementos que orientam a organização acadêmica e pedagógica da formação. De acordo com as orientações da UFMS (2021), o documento contempla a concepção do curso, seus objetivos, o perfil do egresso, as metodologias de ensino, os processos de avaliação da aprendizagem e a organização curricular, incluindo a matriz curricular e o ementário. Além disso, o PPC incorpora as políticas institucionais e assegura o atendimento aos requisitos legais e normativos, abrangendo temas como as relações étnico-raciais, os direitos humanos e a educação ambiental. Assim, a interpretação automatizada dos PPCs requer métodos computacionais capazes de compreender a linguagem presente nesses documentos, destacando o papel do Processamento de Linguagem Natural (PLN).

2.4 Processamento de Linguagem Natural (PLN) e Interpretação de Textos Acadêmicos

O Processamento de Linguagem Natural (PLN) é um ramo da inteligência artificial dedicado ao desenvolvimento de métodos computacionais que possibilitam a interpretação e a geração da linguagem humana de forma automatizada. Com o avanço das tecnologias educacionais, “[...]o PLN tem se destacado como uma ferramenta necessária no cenário educacional, promovendo interações mais inteligentes e personalizadas entre educadores e alunos.” (Santana e Magalhães, 2024).

A utilização de técnicas de Processamento de Linguagem Natural (PLN) na análise de documentos acadêmicos apresenta diversos desafios relevantes, dentre elas “A ambiguidade linguística representa outro obstáculo relevante. Muitas palavras possuem múltiplos significados, cuja interpretação correta depende do contexto em que estão inseridas.” (Silva, 2025) e a variedade de temas nos materiais pedagógicos que dificultam sua interpretação por ferramentas computacionais. Com o avanço do PLN, surgiram arquiteturas como o Retrieval-Augmented Generation (RAG), desenvolvidas para melhorar a recuperação e a interpretação de informações em documentos.

2.5 RAG

O conceito de RAG foi introduzido no trabalho "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", publicado no NeurIPS 2020. Nesse trabalho, foi proposto um modelo que combina a recuperação de informações com um modelo de IA generativa, permitindo consultar uma base de conhecimento durante a geração das respostas. Com isso, as respostas tornam-se mais baseadas em informações reais e há uma redução na geração de informações incorretas ou inventadas (Lewis *et al*., 2020).

No contexto brasileiro, Silva (2025) mostrou que soluções de RAG simplificadas são eficazes na recuperação de informações em documentos acadêmicos. Utilizando o modelo DeepSeek com documentos da UFMG, o autor obteve respostas corretas para todas as perguntas, inclusive as que exigiam maior contexto.

Além de reduzir a ocorrência de informações incorretas, o RAG permite que as respostas sejam fundamentadas em documentos específicos da aplicação. Zhang *et al*. (2026) destacam que essa arquitetura produz respostas mais precisas e contextualizadas quando comparada ao uso isolado de modelos de linguagem.

2.6 Avaliação de Modelos de Linguagem (LLMs)

A avaliação de Modelos de Linguagem de Grande Escala (LLMs) constitui uma etapa fundamental para verificar sua adequação a contextos específicos de uso, uma vez que seu desempenho pode variar significativamente em função da tarefa, do domínio do conhecimento e dos critérios adotados para mensuração da qualidade das respostas (Chang *et al*., 2024).

É importante considerar que modelos de linguagem podem apresentar limitações mesmo quando produzem respostas bem escritas. Wang *et al*. (2025) observam que, devido ao caráter probabilístico dos LLMs, respostas diferentes podem ser produzidas para uma mesma entrada, reforçando a necessidade de avaliações qualitativas conduzidas por pesquisadores.

Para compreender melhor essas características, o desempenho das LLMs pode ser avaliado por diferentes abordagens. Segundo Rodrigues (2025), as principais abordagens de avaliação são a avaliação automática por benchmarks, a avaliação humana e a avaliação conhecida como LLM-as-a-Judge.

A avaliação automática por benchmarks utiliza métricas objetivas, como acurácia, F1-score e BLEU, além de conjuntos padronizados de tarefas, como o MMLU e o ENEM em língua portuguesa (Rodrigues, 2025). Contudo, essas métricas apresentam limitações quando aplicadas a tarefas de geração de texto aberto, pois tendem a privilegiar a similaridade de superfície em detrimento da qualidade semântica e da profundidade analítica das respostas (Rodrigues, 2025).

Em contrapartida, a avaliação humana surge como uma alternativa mais adequada para tarefas subjetivas e de natureza qualitativa, pois permite avaliar aspectos que dificilmente são capturados por métricas automáticas, como clareza, coerência, profundidade, fidelidade e aderência ao documento de referência. Kloss, Cordovez-Fernández e Bustamante (2026) conduziram um estudo comparativo entre a avaliação realizada por LLMs e por especialistas humanos em ensaios argumentativos universitários, utilizando a rubrica analítica RUBRIAR. Os resultados revelaram que a avaliação dos modelos foi similar à do especialista humano em dimensões como propósito comunicativo, mas apresentou baixa precisão nas relações de coesão e coerência, indicando que os modelos ainda têm dificuldades em capturar aspectos mais sutis da qualidade textual. Esse achado reforça a importância de rubricas analíticas bem estruturadas como instrumentos de avaliação, especialmente quando o objetivo é avaliar respostas geradas por IA em contextos acadêmicos.

A abordagem LLM-as-a-Judge, por sua vez, consiste em utilizar um modelo de linguagem mais poderoso como avaliador das respostas geradas por outros modelos. Conforme apontado por Assumpção e Brito (2025) em estudo sobre a avaliação automatizada de redações do ENEM, essa abordagem apresenta potencial para padronizar avaliações em larga escala, mas exige cuidado na definição dos prompts e das rubricas utilizadas, uma vez que a qualidade da avaliação é diretamente

dependente da clareza dos critérios fornecidos ao modelo avaliador. Os autores observaram variações significativas de desempenho entre os modelos testados, com diferenças de latência e custo computacional que também devem ser consideradas em aplicações práticas.

Outra limitação observada nos LLMs é o fenômeno das alucinações, definido como a geração de informações incorretas, imprecisas ou não fundamentadas no contexto fornecido. Segundo Zhang *et al*. (2024 apud Rodrigues, 2025), as alucinações representam um dos maiores desafios para a confiabilidade dos modelos em tarefas que exigem fidelidade às informações de um documento de referência. No contexto de aplicações baseadas em RAG, estudos como o de Bao *et al*. (2025) demonstraram que, mesmo quando os modelos recebem contexto relevante, ainda é frequente a introdução de informações não presentes nos documentos. Esse aspecto reforça a importância de avaliar a fidelidade das respostas ao documento original.

A variação de desempenho entre modelos em diferentes domínios do conhecimento também constitui um aspecto relevante na avaliação comparativa de LLMs. Silva, L. F. M. da (2025) observaram que modelos generativos apresentam comportamentos distintos ao processar documentos acadêmicos, com diferenças na capacidade de sintetizar informações, manter a coerência temática e reproduzir com precisão o conteúdo dos textos originais. Segundo os autores, essas diferenças tendem a ser mais evidentes em domínios técnicos e especializados, nos quais a terminologia específica e a estrutura dos documentos influenciam diretamente a qualidade das respostas.

Considerando essas evidências e as diferentes abordagens de avaliação discutidas nesta seção, esta pesquisa adota uma avaliação humana qualitativa baseada em respostas de referência elaboradas a partir dos PPCs e em uma rubrica analítica composta por cinco critérios. Essa estratégia permite analisar não apenas a correção das informações, mas também aspectos como clareza, coerência, profundidade, aderência ao PPC e fidelidade ao conteúdo do documento, sendo adequada ao objetivo de comparar qualitativamente diferentes modelos de linguagem na interpretação de documentos institucionais.

3 METODOLOGIA

Este estudo trata-se uma pesquisa qualitativa e comparativa, com foco na análise de conteúdo e interpretação textual. A escolha dessa abordagem está relacionada à necessidade de avaliar não apenas a correção das respostas geradas pelos modelos de linguagem, mas também sua capacidade de interpretar informações presentes nos Projetos Pedagógicos de Curso (PPCs). Conforme destacam Silva e Paula (2024), pesquisas que utilizam Inteligência Artificial exigem supervisão humana para assegurar a confiabilidade e a validade das análises produzidas.

Este projeto seguiu as seguintes etapas: Seleção dos documentos, formulação das perguntas, escolha de modelos de IA, criação de um protótipo de interface web interativa, envio dos prompts, registro das respostas, critérios de avaliação e análise dos dados. Inicialmente, foi realizada a seleção de documentos, compreendendo os Projetos Pedagógicos de Curso (PPCs) dos cinco cursos de engenharias vinculados ao Bacharelado Interdisciplinar em Ciência e Tecnologia (BICT), disponibilizados pela Universidade Federal do Maranhão (UFMA). A partir desses documentos, foram formuladas dez perguntas abordando aspectos como objetivos do curso, perfil profissional dos estudantes, metodologia e avaliação, estrutura curricular e impactos sociais e valores.

Após a seleção dos Projetos Pedagógicos de Curso (PPCs), foram elaboradas dez perguntas abertas, aplicadas de forma padronizada aos cinco cursos analisados. As questões foram desenvolvidas com o objetivo de avaliar a capacidade dos modelos de linguagem em interpretar diferentes dimensões dos PPCs, contemplando aspectos como objetivos do curso, perfil do egresso, metodologias de ensino, organização curricular e impacto social. As perguntas completas utilizadas na pesquisa são apresentadas no Apêndice A.

Para cada pergunta, foi elaborada uma resposta de referência baseada nas informações contidas nos PPCs. Essas respostas serviram como padrão para a etapa de avaliação, sendo utilizadas em conjunto com o documento original para comparar as respostas geradas pelos modelos de linguagem.

Para garantir um padrão e controle sobre o processo de coleta das respostas, foi desenvolvido um protótipo de interface web funcional. Nesse protótipo foram carregados os PPCs dos cursos e as perguntas formuladas. A ferramenta integra às APIs dos modelos de linguagens selecionados e permite o upload dos arquivos dos

PPCs, seleção das perguntas diretamente na plataforma, o envio simultâneo dos prompts para todos os modelos e o registro automático das respostas em arquivo estruturado no formato CSV, com exportação em Excel para a análise comparativa dos dados.

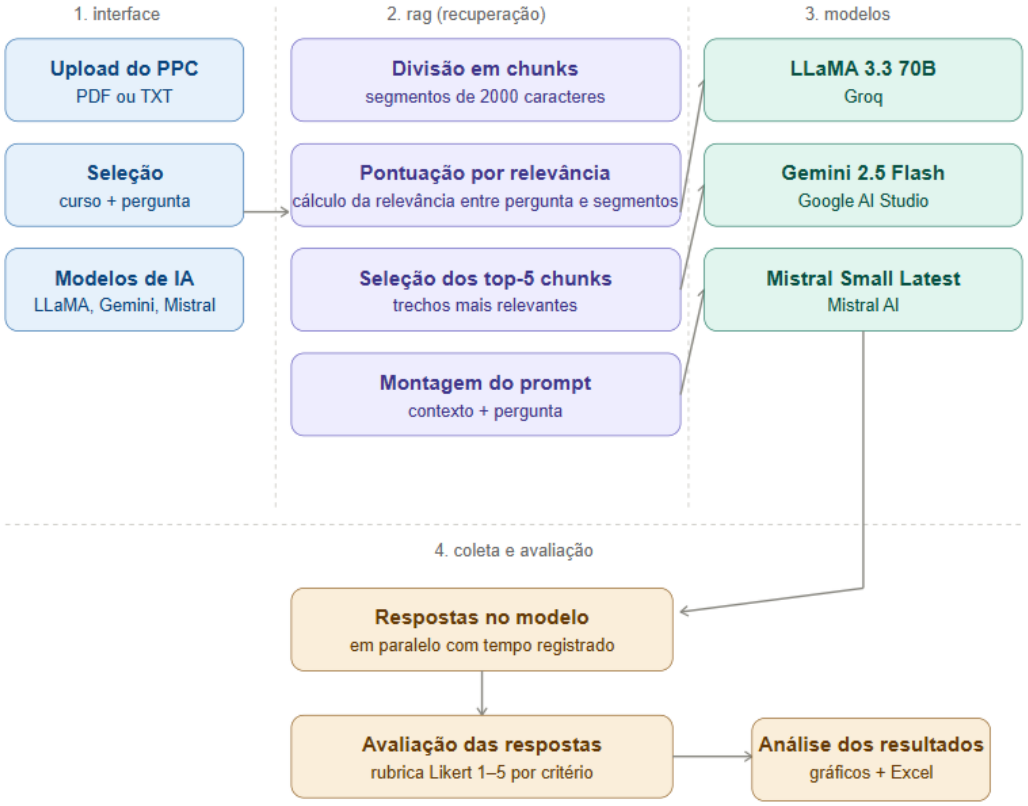
Os descritores de cada nível foram elaborados com base na revisão de literatura realizada, adaptados ao contexto específico desta pesquisa, que tem como objeto as respostas geradas por Inteligência Artificial a partir dos Projetos Pedagógicos de Curso.

3.1 Desenvolvimento do protótipo web para coleta e avaliação de respostas

3.1.1 Desenvolvimento da interface

Para viabilizar a coleta sistemática e a avaliação das respostas geradas pelos modelos de linguagem utilizados nesta pesquisa, foi desenvolvida uma interface web interativa utilizando o framework Streamlit, na linguagem de programação Python. O Streamlit foi escolhido por permitir o desenvolvimento rápido de aplicações para visualização de dados e prototipagem de sistemas, características que facilitaram a implementação e os testes realizados durante a pesquisa. A Figura 1 apresenta a arquitetura geral do sistema, destacando os principais componentes e o fluxo de processamento das informações.

Figura 1 – arquitetura geral da aplicação desenvolvida



Fonte: acervo da autora, 2026.

A interface foi estruturada seguindo o princípio da separação de responsabilidades, organizando o código em módulos independentes distribuídos em cinco diretórios principais:

apis/: responsável pela integração com os modelos de linguagem. Cada arquivo corresponde a um modelo distinto — LLaMA-3.3-70b-versatile, Gemini 2.5 Flash e mistral-small-latest;

core/: contém a lógica central da aplicação, incluindo o módulo de leitura e extração de texto de arquivos PDF (pdf_reader.py), o construtor de prompts (prompt_builder.py) e o orquestrador de requisições paralelas (orchestrator.py).

ui/: agrupa as telas da interface, separadas por funcionalidade: coleta de respostas (config.py) e visualização e avaliação dos dados (dados.py).

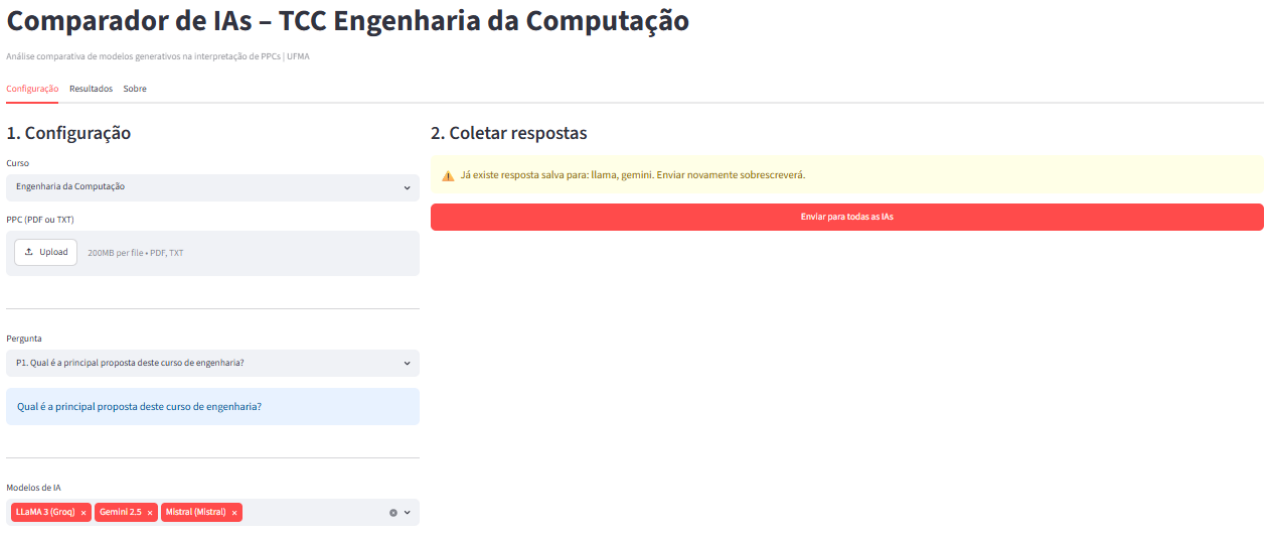
utils/: armazena constantes globais como as perguntas, os cursos e os modelos utilizados (constants.py), os descritores da rubrica de avaliação (rubrica.py), as funções de leitura e escrita do arquivo CSV (exporter.py) e o gerar_excel.py que gera planilha Excel comparando respostas dos modelos.

data/: diretório destinado ao armazenamento dos arquivos PDF dos PPCs, os arquivos em .docx das respostas de referência e do arquivo respostas.csv, gerado automaticamente durante a coleta.

3.1.2 Telas da aplicação

A Figura 2 apresenta a tela inicial da aplicação utilizada durante os experimentos. Nela, o usuário pode fazer o upload do Projeto Pedagógico de Curso (PPC), selecionar qual curso o PPC pertence, escolher uma das dez perguntas elaboradas e escolher o modelo de linguagem. Durante o processo de consulta, o sistema identifica os trechos do PPC mais relacionados à pergunta formulada e constrói um único prompt contendo as instruções ao modelo, os trechos recuperados e a pergunta do usuário. Esse mesmo prompt é enviado aos modelos de linguagem selecionados de forma paralela, garantindo que todos recebam exatamente o mesmo contexto durante a geração das respostas. As respostas produzidas são então armazenadas para posterior avaliação.

Figura 2 – Tela inicial da aplicação



Fonte: Acervo da autora, 2026.

A Figura 3 apresenta um exemplo de resposta produzida pelo sistema após o processamento da consulta. Além de exibir a resposta gerada pelo modelo, a aplicação registra automaticamente os dados da consulta, como o curso, a pergunta,

o modelo utilizado e a resposta produzida. Esses registros foram utilizados posteriormente na etapa de avaliação das respostas, conforme a rubrica analítica apresentada na Seção 3.2.

Figura 3 - Exemplo de resposta gerada pela aplicação

Fonte: Acervo da autora, 2026.

3.1.3 Integração com os modelos de linguagem

Foram integrados três modelos de linguagem de grande escala (LLMs) à interface:

- LLama (Meta AI), acessado por meio da plataforma Groq, que oferece infraestrutura de inferência de alta velocidade com acesso gratuito via API, utilizando o modelo LLama-3.3 70B Versatile.
- Gemini (Google DeepMind), acessado por meio da biblioteca oficial google-generativeai, utilizando o modelo Gemini 2.5 Flash.
- Mistral, acessado por meio da API oficial da Mistral AI, utilizando o modelo mistral-small-latest.

As credenciais de acesso às APIs foram armazenadas em variáveis de ambiente por meio de um arquivo .env, carregado pela biblioteca python-dotenv, garantindo que as chaves não fossem expostas diretamente no código-fonte.

3.1.4 Mecanismo de envio paralelo

Para otimizar o tempo de coleta, o módulo orchestrator.py implementa o envio de forma simultânea do prompt para todos os modelos selecionados, utilizando a classe ThreadPoolExecutor da biblioteca concurrent.futures. A utilização desse mecanismo permite que as requisições sejam disparadas em paralelo, reduzindo o tempo total de espera em comparação ao envio sequencial. Para cada resposta obtida, o tempo de processamento individual é registrado em segundos.

3.1.5 Construção do prompt

O prompt enviado aos modelos foi padronizado por meio do módulo *prompt_builder.py*. Inicialmente, o sistema utilizava apenas os primeiros 20.000 caracteres do Projeto Pedagógico de Curso (PPC) para compor o contexto enviado aos modelos, como forma de respeitar as limitações da janela de contexto. Entretanto, durante o desenvolvimento da aplicação, observou-se que essa estratégia frequentemente desconsiderava informações relevantes localizadas nas partes finais dos documentos, especialmente em PPCs mais extensos.

Diante dessa limitação, optou-se pela adoção de uma estratégia de recuperação de informação de busca por relevância, baseada na arquitetura Retrieval-Augmented Generation (RAG)**.** Conforme proposto por Lewis *et al.* (2020), essa arquitetura combina a recuperação de informações com a geração de texto, permitindo que as respostas sejam fundamentadas em documentos específicos. Além disso, Zhang *et al*. (2026) destacam que essa abordagem contribui para a geração de respostas mais contextualizadas e alinhadas às fontes utilizadas.

Na implementação desenvolvida nesta pesquisa, o texto completo do PPC é dividido em segmentos de tamanho fixo. Em seguida, cada segmento recebe uma pontuação de relevância com base na correspondência entre os termos da pergunta e o conteúdo do segmento. Os segmentos mais relevantes são selecionados e utilizados na construção do prompt enviado ao modelo, permitindo que qualquer parte do documento possa ser considerada, independentemente de sua posição no arquivo.

Essa estratégia foi adotada em todos os experimentos realizados nesta pesquisa, buscando fornecer aos modelos um contexto mais relevante para cada pergunta e favorecer a geração de respostas mais aderentes ao conteúdo dos PPCs.

3.1.6 Fluxo de coleta e avaliação

O fluxo de utilização da interface foi dividido em duas etapas:

Coleta: O usuário realiza o upload do PPC em formato PDF, seleciona o curso, a pergunta e os modelos desejados, e realiza o envio. As respostas são exibidas na tela com o respectivo tempo de resposta de cada modelo e, em seguida, salvas no arquivo respostas.csv sem notas atribuídas.

Avaliação: Em Avaliar respostas na aba Resultados, o pesquisador acessa cada resposta individualmente, aplica as notas referentes aos cinco critérios da rubrica de avaliação como Clareza, Aderência, Coerência, Profundidade e Fidelidade em escala Likert de 1 a 5, e salva a avaliação diretamente na linha correspondente do CSV. É possível editar avaliações previamente realizadas, filtrar por curso, pergunta e o status, se a resposta está avaliada ou pendente.

3.1.7 Armazenamento dos dados

Todas as respostas geradas e respectivas avaliações foram armazenadas em um arquivo Excel, contendo informações como data e hora da coleta, curso analisado, identificador e texto da pergunta, modelo utilizado, resposta gerada, tempo de resposta em segundos, pontuações atribuídas a cada critério da rubrica e a média geral das avaliações.

3.1.8 Visualização dos resultados

A interface disponibiliza diferentes técnicas de visualização dos dados coletados, implementadas com a biblioteca Plotly, com o objetivo de facilitar a análise comparativa entre os modelos de linguagem.

Entre as visualizações utilizadas, destaca-se o gráfico de barras agrupadas, que permite comparar a média dos critérios da rubrica de avaliação entre os modelos. Foi utilizado também o gráfico radar, que representa simultaneamente o desempenho dos modelos nos cinco critérios da rubrica, possibilitando uma análise visual do perfil geral de cada modelo, tanto de forma geral quanto por curso, permitindo observar variações no desempenho dos modelos de acordo com o PPC analisado. Foi utilizado

também um mapa de calor (heatmap), construído a partir do cruzamento entre cursos e modelos, exibindo a nota média obtida em cada combinação em uma escala de cores.

Complementando essas visualizações, os resultados também são apresentados em tabelas, permitindo a consulta detalhada das respostas geradas pelos modelos e das notas atribuídas em cada critério de avaliação.

3.2 Instrumento de Avaliação: Rubrica Analítica

Para a avaliação das respostas geradas pelas Inteligências Artificiais generativas a partir dos Projetos Pedagógicos de Curso (PPCs), foi construída uma rubrica analítica composta por cinco critérios: Clareza, Aderência ao PPC, Coerência, Profundidade e Fidelidade das Informações. Cada critério é avaliado em uma escala de 1 a 5 pontos, totalizando uma pontuação máxima de 25 pontos por resposta.

As respostas de referência foram elaboradas manualmente, a partir da leitura integral de cada PPC. Essas respostas serviram como base para a aplicação da rubrica analítica, permitindo comparar as informações apresentadas pelos modelos de linguagem com o conteúdo dos documentos originais.

A utilização de uma escala de cinco pontos e de critérios independentes é respaldada por Shukla *et al*. (2025), que empregaram uma escala Likert para avaliar a qualidade das respostas produzidas por diferentes LLMs, e por Yavuz, Çelik e Çelik (2024), que utilizaram uma rubrica analítica multidimensional para avaliar a confiabilidade e a validade de modelos de linguagem em atividades acadêmicas. A adoção dessa abordagem contribui para reduzir a subjetividade da avaliação e permite analisar diferentes aspectos das respostas produzidas pelos modelos.

Os critérios definidos para esta pesquisa foram escolhidos por contemplarem diferentes dimensões da qualidade das respostas. Enquanto Clareza e Coerência avaliam a organização e a compreensão do texto, Aderência e Fidelidade verificam o alinhamento das respostas ao conteúdo dos PPCs. Já o critério de Profundidade busca avaliar a capacidade do modelo de interpretar e desenvolver as informações presentes nos documentos, indo além da simples recuperação de trechos. Conforme discutido por Montes *et al*. (2025), esse tipo de avaliação é importante para identificar se o modelo consegue captar interpretações implícitas do texto, e não apenas reproduzir informações superficiais. Dessa forma, a combinação desses critérios

permite uma avaliação mais abrangente e consistente do desempenho dos modelos analisados.

Os critérios, seus respectivos descritores e a escala de pontuação utilizada nesta pesquisa são apresentados no Quadro 1**.**

Quadro 1 – Rubrica analítica para avaliação de respostas

Critério	1 muito baixo	2 baixo	3 regular	4 bom	5 excelente
Clareza	Texto confuso, desorganizado e de difícil compreensão.	Texto com estrutura precária e vocabulário limitado	Texto compreensível, mas com trechos ambíguos	Resposta clara e bem estruturada	Texto extremamente claro e acessível
Aderência ao PPC	fora do documento.	Utiliza poucas informações relacionadas ao PPC e apresenta desvios relevantes.	Apresenta algumas informações do PPC, mas deixa pontos importantes de fora ou inclui informações pouco relacionadas.	Baseia-se principalmente no PPC e responde adequadamente à pergunta.	respondendo exatamente ao que foi solicitado.
Coerência	Apresenta contradições, ideias sem relação ou raciocínio incorreto.	Possui inconsistências e organização lógica fraca.	As ideias fazem sentido parcialmente, mas há problemas na sequência ou conexão.	Apresenta ideias organizadas e logicamente relacionadas.	Apresenta raciocínio lógico, consistente e com boa relação entre as informações apresentadas.
Profundidade	Resposta superficial, vaga ou sem explicar o	Apresenta poucas informações e não	Responde ao básico, mas poderia explorar	Apresenta informações relevantes e	Explora os principais pontos do PPC, trazendo

	conteúdo solicitado.	desenvolve o tema.	melhor o conteúdo do PPC.	algum detalhamento.	detalhes relevantes sem fugir do tema.
Fidelidade das informações do PPC	Contém informações inventadas, incorretas ou contradiz o PPC.	Possui muitos erros ou informações não presentes no PPC.	Parte das informações está correta, mas apresenta generalizações ou pequenos erros.	Informações predominantemente corretas, com pequenas omissões ou imprecisões.	Todas as informações apresentadas estão de acordo com o PPC, sem adicionar dados não encontrados.

Fonte: Elaborado pela autora, 2026, com base em Shukla *et al*. (2025), Yavuz, Çelik e Çelik (2024), Lima e Serrano (2024), Mugaanyi *et al*. (2024) e Miranda, Rosini e Macedo (2025).

3.3 Justificativa dos critérios

A avaliação das respostas geradas pelos modelos foi realizada por meio de uma rubrica analítica composta por cinco critérios: Clareza, Aderência e fidelidade ao PPC, Coerência e Profundidade. Segundo Yavuz, Çelik e Çelik (2024), a utilização de critérios independentes contribui para uma avaliação mais confiável, uma vez que o desempenho dos modelos pode variar conforme a dimensão analisada. De forma complementar, Shukla *et al*. (2025) destacam que o uso de rubricas analíticas favorece comparações consistentes entre diferentes modelos de linguagem.

O critério de Clareza refere-se à organização e a compreensibilidade da resposta gerada pela IA. Shukla *et al*. (2025) tratam a clareza como um domínio independente, denominado *clarity and accessibility*. Em avaliações de modelos de linguagem em contextos acadêmicos, os autores observaram desempenhos elevados de modelos como o Claude nesse aspecto, demonstrando a importância da clareza para a qualidade das respostas.

Aderência ao PPC diz respeito ao nível de proximidade entre a resposta gerada e o conteúdo presente no documento institucional. Yavuz, Çelik e Çelik (2024) identificam que modelos de linguagem apresentam variações de desempenho no domínio de conteúdo, o qual se relaciona diretamente à capacidade de responder de acordo com o tema proposto. Em contexto brasileiro, Miranda, Rosini e Macedo (2025) evidenciam que é possível produzir conteúdo alinhados às diretrizes curriculares,

sendo esse alinhamento um aspecto importante na avaliação da qualidade das respostas.

Coerência refere-se à forma como as ideias são organizadas e articuladas ao longo da resposta, garantindo uma sequência lógica e facilitando a compreensão do texto. Yavuz, Çelik e Çelik (2024) destacam a organização e a coerência como aspectos relevantes em rubricas utilizadas para avaliar textos gerados por modelos de linguagem. Já Lima e Serrano (2024) observam que esses modelos ainda podem apresentar falhas na lógica ou inconsistências em determinadas situações de geração de texto.

A Profundidade está relacionada ao nível de detalhamento e a capacidade da resposta de explorar adequadamente os aspectos solicitados. Em aplicações na área clínica, Cassar, Galea e Ferry (2026) observaram diferenças de desempenho entre critérios relacionados à completude e à clareza, indicando que modelos podem produzir respostas claras, mas pouco aprofundadas. Lima e Serrano (2024) também destacam que modelos generativos podem sintetizar informações sem necessariamente aprofundar a análise.

A Fidelidade das Informações do PPC, corresponde a capacidade do modelo de reproduzir corretamente o conteúdo do documento, sem acréscimos, omissões relevantes ou distorções. Mugaanyi *et al*. (2024) analisam a confiabilidade de modelos de linguagem em tarefas de citação e referência, identificando variações de desempenho e ocorrência de erros factuais. Zhou *et al*. (2023) e Kasneci *et al*. (2023), citados por Lima e Serrano, 2024) reforçam que as alucinações, geração de informações incorretas ou inexistentes, representam uma limitação relevante desses modelos em contextos acadêmicos. Neste trabalho, esse critério recebeu atenção especial por estar diretamente relacionado ao objetivo da pesquisa, que consiste em avaliar a capacidade dos modelos de interpretar os PPCs sem acrescentar informações inexistentes ou distorcer o conteúdo dos documentos.

Além da definição dos critérios de avaliação, optou-se pela utilização de perguntas abertas e interpretativas. Essa escolha deve-se ao fato de que os Projetos Pedagógicos de Curso apresentam informações que vão além de dados objetivos. Dessa forma, perguntas subjetivas permitem avaliar não apenas a recuperação de informações, mas também a capacidade dos modelos de interpretar o conteúdo dos documentos.

Essa abordagem favorece especialmente a análise dos critérios de Profundidade e Fidelidade, considerados essenciais nesta pesquisa. Conforme observado por Montes *et al*. (2025), modelos de linguagem podem apresentar dificuldades na interpretação de significados implícitos presentes em textos complexos, tornando necessária uma avaliação que considere aspectos além da simples recuperação de informações.

4 RESULTADOS E DISCUSSÕES

4.1 Visão Geral dos Dados Coletados

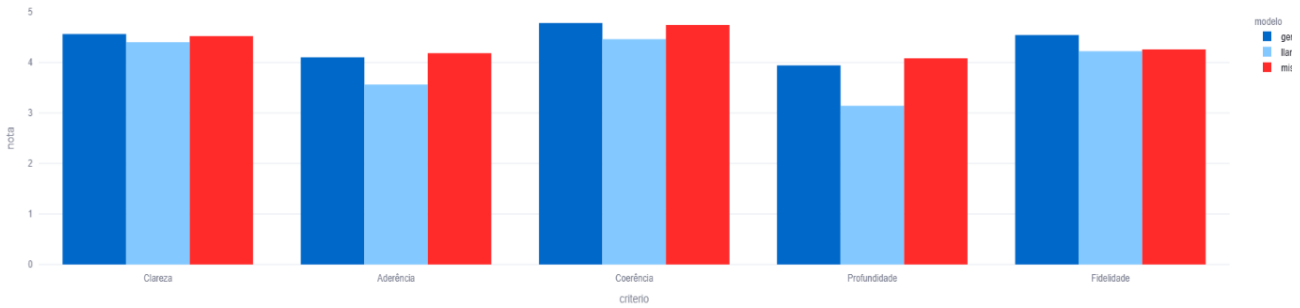
Para a realização desta pesquisa, foram coletadas e avaliadas 150 respostas geradas por três modelos de linguagem de grande escala (LLMs): LlaMA-3.3 70b versatile (Groq), Gemini 2.5 Flash (Google) e mistral-small-latest (Mistral AI). As respostas foram obtidas a partir de 10 perguntas aplicadas sobre os Projetos Pedagógicos de Curso (PPCs) de cinco cursos de engenharia da Universidade Federal do Maranhão (UFMA) vinculados ao BICT: Engenharia da Computação, Engenharia Aeroespacial, Engenharia Civil, Engenharia Mecânica e Engenharia Ambiental, totalizando 30 respostas por curso (10 perguntas × 3 modelos).

Cada resposta foi avaliada manualmente com base em uma rubrica analítica composta por cinco critérios: Clareza, Aderência e Fidelidade ao PPC, Coerência e Profundidade, em escala Likert de 1 a 5, conforme apresentado no Quadro 1. A avaliação foi realizada com consulta simultânea ao PPC correspondente e à resposta de referência elaborada de forma manual.

4.2 Desempenho por Critério

A Figura 4 apresenta o gráfico de barras agrupadas com as médias obtidas por cada modelo em cada critério da rubrica, considerando a totalidade das respostas coletadas.

Figura 4 — Gráfico de barras agrupadas por critério e modelo



Fonte: Acervo da autora, 2026.

Os resultados obtidos demonstram que os três modelos de linguagem avaliados foram capazes de interpretar os Projetos Pedagógicos de Curso (PPCs), apresentando desempenho satisfatório em todos os critérios da rubrica analítica. Entre os modelos avaliados, o Gemini-2.5-Flash apresentou o desempenho mais consistente, seguido pelo mistral-small-latest, enquanto o LLaMA-3.3 70B versatile obteve médias inferiores principalmente nos critérios de Aderência e Profundidade.

Os critérios de Clareza e Coerência apresentaram as maiores médias entre os modelos, indicando que as respostas geradas foram, em geral, bem estruturadas, organizadas e de fácil compreensão. Esse resultado demonstra que os modelos atuais possuem elevada capacidade de produzir textos linguisticamente consistentes.

Este achado confirma o que é observado na literatura sobre a "polidez" linguística dos modelos generativos. Montes *et al*. (2025) destacam que avaliadores humanos frequentemente preferem resultados de IAs (61% das vezes em seu estudo) devido à sua "legibilidade e polimento superficial", que fazem os códigos e textos parecerem mais precisos em comparações diretas. Da mesma forma, Liu, Ye e Yan (2026) apontam que os LLMs priorizam a precisão lexical e a fluência, o que explica as altas notas de clareza encontradas nesta análise. Assim como a clareza, a coerência obteve pontuações elevadas, refletindo a habilidade dos modelos em organizar ideias de forma lógica e conectada, garantindo que o texto faça sentido do início ao fim.

O critério Profundidade apresentou as menores médias entre os modelos avaliados. Os resultados indicam que, embora os modelos localizem informações, possuem limitações quando a tarefa exige maior capacidade de contextualização e detalhamento de informações implícitas ou estruturais. Este resultado converge com as limitações apontadas por Montes *et al*. (2025), que observaram que as IAs tendem a fragmentar dados desnecessariamente e perder interpretações. O LLaMA-3.3 70B apresentou maior variabilidade de desempenho, com reduções acentuadas neste critério. O Gemini e o mistral-small-latest mostraram-se mais consistentes, mantendo um desempenho mais uniforme nesse critério.

A variação na aderência pode ser explicada pela teoria da "Engenharia de Contexto" proposta por Gentil e Oliveira (2025). Eles argumentam que a aderência e a consistência das respostas em cenários complexos dependem da curadoria das informações que condicionam a resposta. Nessa perspectiva, os resultados desta pesquisa indicam que o Gemini-2.5-Flash e o mistral-small-latest apresentaram maior

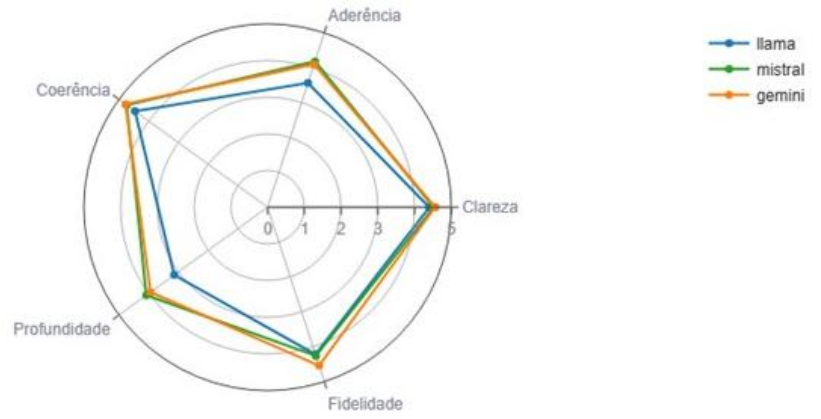
alinhamento entre as respostas geradas e as diretrizes curriculares presentes nos PPCs, superando o desempenho do LLaMA nesse critério e evidenciando menor ocorrência de interpretações incorretas ou inconsistentes.

A superioridade de alguns modelos em relação a fidelidade reflete a eficácia da redução de "alucinações", um desafio central discutido por Anjos et al. (2024). De acordo com Silva e Paula (2024), o uso de IA em análises textuais exige que o pesquisador assegure a validade e confiabilidade dos dados para evitar distorções que invalidariam a pesquisa. os resultados desta pesquisa indicam que a escolha do modelo de linguagem influencia diretamente a fidelidade das respostas geradas em relação ao conteúdo dos Projetos Pedagógicos de Curso.

4.3 Perfil Geral dos Modelos

Os modelos Gemini-2.5-Flash e mistral-small-latest apresentaram desempenho bastante semelhante na maioria dos critérios avaliados, como apresentado na figura 5, indicando comportamento consistente na interpretação dos Projetos Pedagógicos de Curso quando submetidos às mesmas condições experimentais. Esse resultado está alinhado ao estudo de Zhang *et al*. (2026), que demonstraram que uma arquitetura baseada em Retrieval-Augmented Generation (RAG) favorece a produção de respostas mais corretas, completas e claras. Complementarmente. Além disso, Gentil e Oliveira (2025) destacam que a qualidade das respostas também está relacionada à organização e à seleção das informações fornecidas como contexto ao modelo.

Figura 5 — Gráfico radar geral dos modelos



Valores por modelo

	llama	mistral	gemini
Clareza	4.4	4.52	4.56
Aderência	3.56	4.18	4.1
Coerência	4.46	4.74	4.78
Profundidade	3.14	4.08	3.94
Fidelidade	4.22	4.26	4.54

Fonte: Acervo da autora, 2026.

Por outro lado, o LLaMA-3.3 70B versatile apresentou maior variabilidade de desempenho, com reduções mais acentuadas nos critérios de Profundidade e Aderência ao PPC. Essa concentração da instabilidade em critérios específicos é consistente com a literatura de Yavuz, Çelik e Çelik (2024), que destacam que a qualidade da IA deve ser analisada em múltiplas dimensões, uma vez que o desempenho pode variar drasticamente conforme o aspecto avaliado. De forma semelhante, Silva (2025) observou diferenças de desempenho entre modelos de linguagem na interpretação de documentos acadêmicos, indicando que a estabilidade das respostas pode variar mesmo quando os modelos são avaliados sob um mesmo protocolo experimental.

Um aspecto observado nos resultados foi que todos os modelos apresentaram médias elevadas nos critérios de Clareza e Coerência, enquanto obtiveram desempenho inferior em Profundidade. Esse comportamento indica que, embora os modelos sejam capazes de produzir respostas bem estruturadas, organizadas e de fácil compreensão, ainda apresentam limitações quando a tarefa exige maior

contextualização e interpretação das informações presentes nos Projetos Pedagógicos de Curso.

Esse resultado está em consonância com Montes *et al*. (2025), que identificaram dificuldades dos modelos de linguagem em interpretar significados implícitos e produzir respostas mais aprofundadas em tarefas complexas. Da mesma forma, Liu, Ye e Yan (2026) destacam que os LLMs apresentam elevado desempenho na produção de textos fluentes e linguisticamente consistentes, característica que explica as altas pontuações obtidas nos critérios de Clareza e Coerência.

Dessa forma, os resultados desta pesquisa indicam que, embora os modelos avaliados apresentem boa capacidade de organização textual e comunicação das informações, ainda encontram limitações na elaboração de respostas mais aprofundadas e na interpretação de aspectos implícitos presentes nos PPCs.

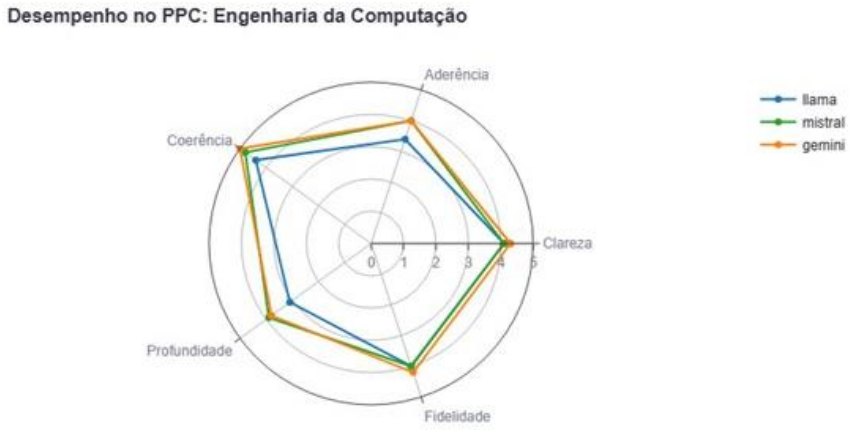
4.4 Perfil dos Modelos por curso

Os gráficos de radar apresentados nesta seção ilustram o desempenho comparativo dos três modelos de linguagem LLaMA-3.3 70B versatile, Gemini-2.5-Flash e mistral-small-latest para cada curso analisado, considerando os cinco critérios da rubrica analítica.

4.4.1 Engenharia da computação

A Figura 6 apresenta o desempenho dos modelos de linguagem na interpretação do Projeto Pedagógico do Curso de Engenharia da Computação, considerando os cinco critérios estabelecidos na rubrica de avaliação.

Figura 6 — Gráfico radar: Engenharia da computação



Valores por modelo

	llama	mistral	gemini
Clareza	4.1	4.1	4.3
Aderência	3.4	4	4
Coerência	4.4	4.8	5
Profundidade	3.1	3.9	3.8
Fidelidade	4	4	4.2

Fonte: Acervo da autora, 2026.

A diferença de desempenho observada no PPC de Engenharia da Computação, principalmente a dificuldade do LLaMA em alcançar altos níveis de Profundidade e Fidelidade, confirma uma limitação apontada por Wang *et al*. (2025). Segundo os autores, os modelos de linguagem costumam analisar o texto com base em características mais superficiais, em vez de compreender seu significado de forma mais profunda. Esse comportamento também foi observado nesta pesquisa, na qual o modelo apresentou interpretações incorretas de informações presentes no PPC, como as metodologias de ensino, indicando ausência de detalhes que, na realidade, estavam descritos no documento.

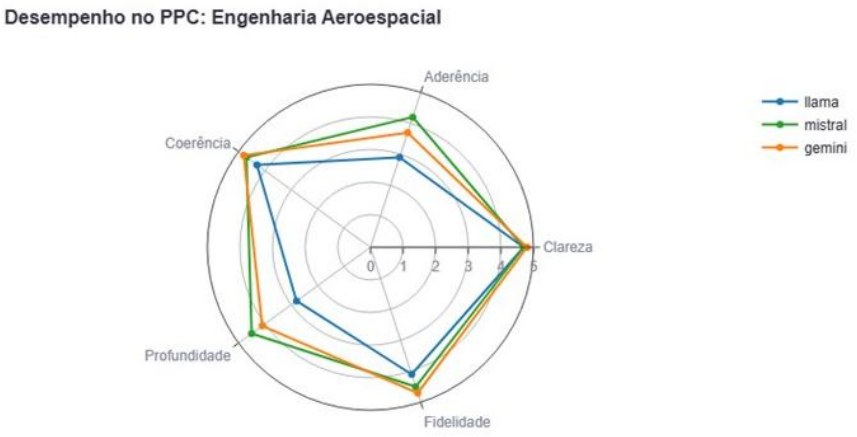
Por outro lado, o bom desempenho de modelos como Gemini-2.5-Flash e mistral-small-latest nos critérios de Clareza e Coerência confirma os achados de Liu, Ye e Yan (2026), que destacam a capacidade desses modelos em relação a precisão das palavras e a fluência da escrita. Entretanto, como destacam Montes *et al.* (2025), essa qualidade na escrita não garante, necessariamente, uma compreensão mais

profunda do conteúdo, uma vez que os modelos ainda apresentam dificuldade relacionadas a interpretações e na identificação de significados implícitos. Diante disso, torna-se importante que as respostas geradas por IA sejam revisadas e validadas pelo pesquisador, garantindo a confiabilidade e a validade dos dados, conforme defendem Silva e Paula (2024).

4.4.2 Engenharia Aeroespacial

Os resultados obtidos para o PPC de Engenharia Aeroespacial são apresentados na Figura 7.

Figura 7 — Gráfico radar: Engenharia Aeroespacial



Valores por modelo

	llama	mistral	gemini
Clareza	4.7	4.7	4.8
Aderência	2.9	4.2	3.7
Coerência	4.3	4.7	4.8
Profundidade	2.8	4.5	4.1
Fidelidade	4.1	4.5	4.7

Fonte: Acervo da autora, 2026.

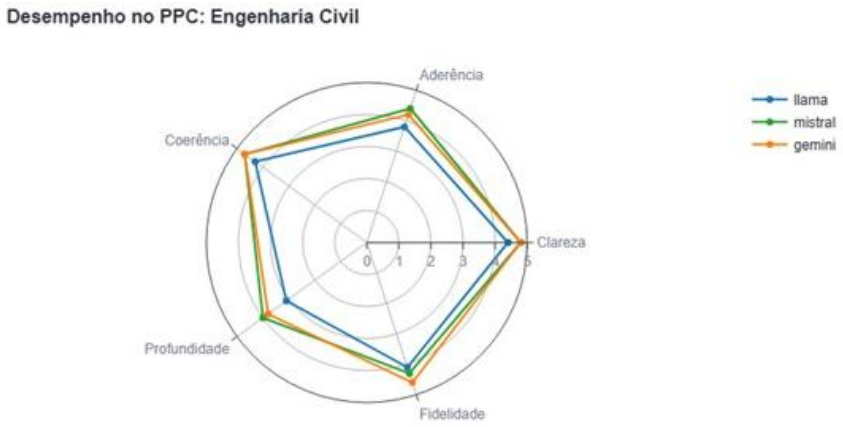
Os resultados obtidos no PPC de Engenharia Aeroespacial destacaram expressiva queda do LLaMA-3.3 70B versatile, ao responder questões relacionadas a temas técnicos, como telemetria e o Centro de Lançamento de Alcântara (CLA). Esse comportamento está de acordo com as observações de Montes *et al*. (2025), que

apontam que modelos de linguagem podem fragmentar informações e apresentar dificuldades na interpretação de significados implícitos em textos complexos. De forma semelhante, Liu, Ye e Yan (2026) destacam que, embora os LLMs produzam respostas linguisticamente fluentes, a qualidade da escrita não garante a correta interpretação de conteúdos específicos e tecnicamente especializados.

4.4.3 Engenharia Civil

O desempenho dos modelos nos critérios avaliados usando o PPC da Engenharia Civil, é ilustrado na Figura 8.

Figura 8 — Gráfico radar: Engenharia Civil



Valores por modelo

	llama	mistral	gemini
Clareza	4.4	4.8	4.8
Aderência	3.8	4.4	4.2
Coerência	4.3	4.7	4.7
Profundidade	3.1	4	3.8
Fidelidade	4.1	4.3	4.6

Fonte: Acervo da autora, 2026.

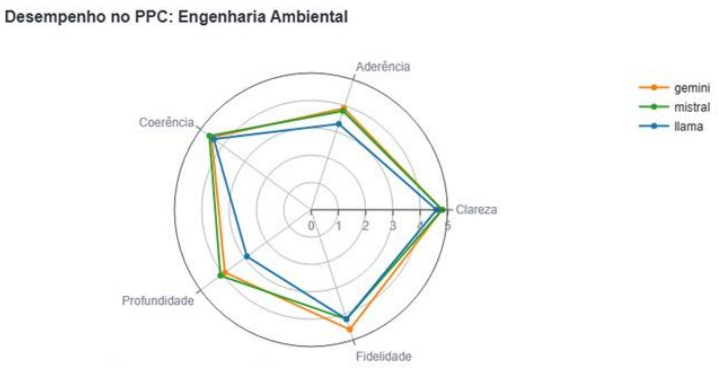
A proximidade dos resultados obtidos pelos modelos Gemini e mistral no PPC da Engenharia Civil indicam desempenho semelhante em relação a interpretação do

documento. Esse resultado está de acordo com Gentil e Oliveira (2025), que destacam a importância de uma boa organização das informações e da qualidade do contexto fornecido para orientar as respostas geradas pela IA. No entanto, o modelo LLaMA apresentou redução em relação ao desempenho nos critérios de Profundidade e Fidelidade, evidenciando que respostas com uma boa estrutura não garante, por si só, uma análise mais completa do conteúdo. Nesse sentido, afirmam Silva e Paula (2024), ressaltam que o uso da IA em pesquisas qualitativas ainda apresenta desafios e não assegura que o modelo consiga interpretar o conteúdo de forma aprofundada. De forma semelhante, Yavuz, Çelik e Çelik (2024), destacam que alguns textos exigem uma compreensão mais detalhada para que a resposta da IA se aproxime da interpretação humana.

4.4.4 Engenharia Ambiental

A análise referente ao PPC de Engenharia Ambiental é apresentada na Figura 9, que reúne o desempenho dos modelos nos cinco critérios da rubrica.

Figura 9 — Gráfico radar: Engenharia Ambiental



Valores por modelo

	gemini	mistral	llama
Clareza	4.8	4.8	4.6
Aderência	3.9	3.8	3.3
Coerência	4.5	4.6	4.4
Profundidade	3.9	4.1	2.9
Fidelidade	4.6	4.2	4.2

Fonte: Acervo da autora, 2026.

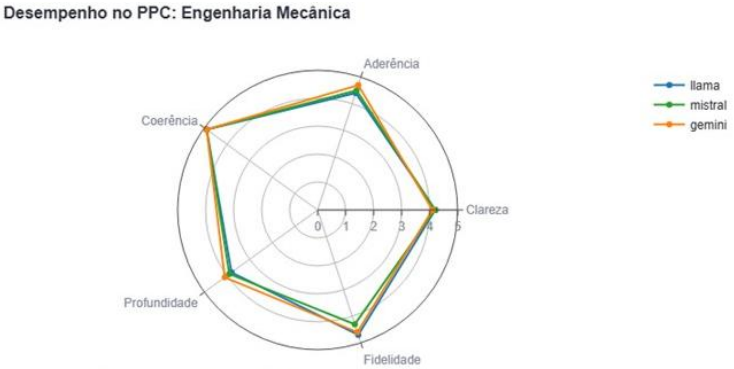
No PPC de Engenharia Ambiental, o LLaMA apresentou as menores notas em Profundidade e Fidelidade, mesmo mantendo um bom desempenho em Clareza e Coerência. Esse resultado está de acordo com Silva e Paula (2024), que apontam que modelos de linguagem podem apresentar dificuldades na análise de textos mais extensos, comprometendo a identificação de informações importantes durante a análise. Por outro lado, Gemini e mistral apresentaram resultados mais consistentes, especialmente nos critérios relacionados à interpretação e à fidelidade das informações. Esse comportamento está de acordo com Anjos *et al*. (2024), que observaram que modelos de linguagem apresentam diferenças na capacidade de produzir respostas fundamentadas nas informações fornecidas. Dessa forma, os resultados mostram que a capacidade de interpretar e utilizar adequadamente o contexto dos PPCs influencia diretamente a qualidade e a fidelidade das respostas geradas pelos modelos.

4.4.5 Engenharia Mecânica

Os resultados obtidos na avaliação utilizando o PPC de Engenharia Mecânica, apresentados na Figura 10, mostram desempenho semelhante entre os modelos avaliados. Esse comportamento está em consonância com a proposta da Engenharia de Contexto (EC) apresentada por Gentil e Oliveira (2025), que destaca a importância da organização e da seleção das informações fornecidas aos modelos de linguagem.

De forma semelhante, Silva e Paula (2024) ressaltam que a qualidade das respostas depende da capacidade dos modelos de utilizar adequadamente as informações disponíveis ao longo do documento. Além disso, a proximidade dos resultados entre os modelos está de acordo com Yavuz, Çelik e Çelik (2024), que destacam que o desempenho dos LLMs pode ser influenciado pelas características do contexto e do texto analisado. Esses resultados também convergem com Liu, Ye e Yan (2026), ao evidenciar que modelos de linguagem são capazes de produzir respostas consistentes quando dispõem de contexto adequado para fundamentar suas respostas.

Figura 10 — Gráfico radar: Engenharia Mecânica



Valores por modelo

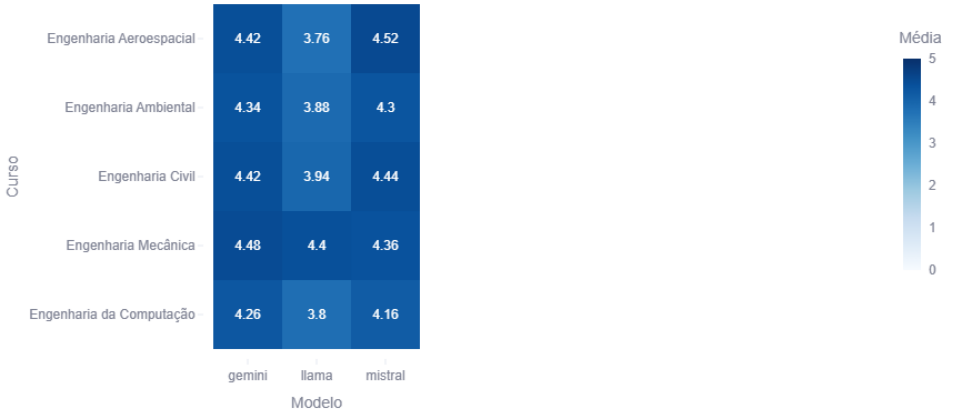
	llama	mistral	gemini
Clareza	4.2	4.2	4.1
Aderência	4.4	4.5	4.7
Coerência	4.9	4.9	4.9
Profundidade	3.8	3.9	4.1
Fidelidade	4.7	4.3	4.6

Fonte: Acervo da autora, 2026.

4.5 Desempenho por Curso

A Figura 11 apresenta o heatmap com as médias obtidas pelos modelos em cada Projeto Pedagógico de Curso (PPC) avaliado. De forma geral, observa-se que o desempenho variou entre os cursos, indicando que as características de cada documento influenciaram a qualidade das respostas geradas pelos modelos.

Figura 11 — Heatmap nota média por curso e modelo



Fonte: Acervo da autora, 2026.

O PPC de Engenharia Mecânica apresentou as maiores médias entre os cursos avaliados, com desempenho semelhante entre os três modelos. Em contrapartida, o LLaMA obteve as menores médias nos PPCs de Engenharia Aeroespacial e Engenharia da Computação, enquanto Gemini e mistral mantiveram resultados mais consistentes. Os cursos de Engenharia Civil e Engenharia Ambiental apresentaram desempenho intermediário, com diferenças menores entre os modelos.

Esses resultados reforçam que o desempenho dos modelos de linguagem variou conforme o Projeto Pedagógico de Curso analisado. Esse comportamento está de acordo com Silva, L. F. M. da (2025), que observaram diferenças no desempenho de modelos generativos ao processar documentos acadêmicos de diferentes áreas do conhecimento. Além disso, Chang *et al*. (2024) destacam que o desempenho dos LLMs pode variar conforme a tarefa, o domínio do conhecimento e os critérios utilizados na avaliação.

4.6 Analise qualitativa das respostas

A análise qualitativa das respostas complementa os resultados quantitativos apresentados nas seções anteriores, permitindo observar comportamentos dos modelos que não podem ser identificados apenas pelas médias obtidas na rubrica analítica. Para isso, foram selecionados três exemplos representativos: uma resposta com interpretação inadequada do conteúdo do PPC, uma resposta considerada satisfatória em todos os critérios avaliados e uma resposta parcialmente correta, que apresentou omissões relevantes.

Quadro 2 – Exemplo de resposta com interpretação divergente do PPC

Curso	Engenharia da computação
Pergunta	As metodologias de ensino propostas no PPC incentivam a autonomia e o pensamento crítico dos alunos?
IA utilizada	Gemini
Resposta de referencia	Sim. O PPC destaca que, por ser um curso de segundo ciclo, o público-alvo é composto por alunos mais

	maduros, preparados para um papel "eminentemente ativo e autônomo na sua formação". [...]
Resposta da IA	Não, os trechos do Projeto Pedagógico de Curso (PPC) fornecidos não detalham as metodologias de ensino específicas que serão propostas para incentivar a autonomia e o pensamento crítico dos alunos. [...]

Fonte: elaborado pela autora, 2026.

O primeiro exemplo demonstra uma situação em que o modelo interpretou de forma inadequada as informações presentes no Projeto Pedagógico de Curso de Engenharia da Computação.

Embora o modelo tenha utilizado trechos do documento como base para sua resposta e não tenha apresentado informações inexistentes, concluiu que o PPC não apresentava evidências suficientes para afirmar que as metodologias de ensino incentivavam a autonomia e o pensamento crítico dos estudantes. Entretanto, a análise do documento mostrou que essas informações estavam presentes no PPC e fundamentaram a resposta de referência.

Dessa forma, a resposta apresentou uma interpretação inadequada do conteúdo recuperado, comprometendo principalmente os critérios de Aderência, Profundidade e Fidelidade, apesar de manter elevada Clareza e Coerência. Esse comportamento está de acordo com Wang *et al*. (2025), que apontam que os modelos de linguagem podem interpretar o texto com base em características mais superficiais, sem compreender plenamente seu significado.

Quadro 3 – Resposta com desempenho máximo

Curso	Engenharia da mecânica
Pergunta	Como é o profissional que o curso quer formar?
IA utilizada	LlaMA-3.3 70B versatille
Resposta da IA	O curso de Engenharia Mecânica da UFMA busca formar profissionais qualificados que tenham uma sólida formação técnica, sem negligenciar as questões éticas e socioambientais, com capacidade de criar soluções tecnológicas e garantir a diversificação da base econômica. Os

	profissionais formados devem ter habilidades pessoais como criatividade, responsabilidade, pró-atividade, trabalho em equipe, ética e responsabilidade social, além de conhecimentos sólidos nas áreas de matemática, física, química e engenharia mecânica. Especificamente, o curso visa formar engenheiros mecânicos que possam atuar em quatro áreas: 1. Projetos Mecânicos: com habilidades para projetar máquinas e equipamentos mecânicos. 2. Materiais e Processos: com conhecimento sobre materiais metálicos e suas propriedades, seleção e caraterização de materiais, métodos de manufatura, etc. 3. Térmica e Fluidos: com habilidades para estudar mecânica dos fluidos, termodinâmica, projeto de sistemas térmicos e equipamentos, etc. 4. Planejamento e Gestão: com habilidades para gerenciar, planejar e organizar sistemas de produção, implantação, manutenção dos processos produtivos de uma organização de forma responsável e crítica. Além disso, o curso busca formar cidadãos conscientes, com atitudes, posturas e valores adequados, que garantam o respeito aos direitos legais e a valorização da identidade.

Fonte: elaborado pela autora, 2026.

O segundo exemplo apresenta uma resposta que recebeu nota máxima em todos os critérios da rubrica analítica.

O modelo respondeu integralmente à pergunta, descrevendo o perfil do egresso, as competências esperadas e as principais áreas de atuação do engenheiro mecânico. Além disso, organizou as informações de forma clara e lógica, explorando os aspectos mais relevantes do PPC sem acrescentar informações externas ou contradizer o documento.

Nesse caso, a resposta apresentou elevado nível de aderência ao conteúdo do PPC, reunindo informações corretas, bem estruturadas e compatíveis com o documento de referência. O exemplo evidencia que os modelos de linguagem podem

produzir respostas completas e fiéis ao conteúdo analisado quando conseguem recuperar e interpretar adequadamente as informações presentes no documento.

O quadro 4 apresenta uma resposta que identificou corretamente diversas estratégias previstas no PPC de Engenharia Ambiental para aproximar os estudantes da prática profissional, como atividades práticas, oficinas interdisciplinares, atividades complementares e participação em eventos técnico-científicos. Entretanto, a resposta não mencionou o estágio curricular obrigatório, um dos principais mecanismos previstos no documento para aproximar o estudante da realidade profissional.

Quadro 4 – Resposta com omissão de conteúdo

Curso	Engenharia ambiental
Pergunta	O PPC apresenta estratégias para aproximar os estudantes da prática profissional? Quais?
IA utilizada	mistral-small-latest
Resposta de referencia	Sim, o PPC apresenta diversas estratégias para aproximar os estudantes da prática profissional. Destacam-se as três oficinas integradoras que simulam o trabalho em análise, projeto e gestão de engenharia. Além disso, há a obrigatoriedade do estágio curricular (160 horas), que pode ser realizado em órgãos de fiscalização, empresas de saneamento ou indústrias. Outras estratégias incluem a participação na Empresa Júnior do curso e o incentivo para que professores incluam os alunos em projetos de extensão voltados à resolução de problemas técnicos da comunidade e de setores produtivos.
Resposta da IA	Sim, o PPC apresenta várias estratégias para aproximar os estudantes da prática profissional. As principais são: 1. **Inclusão de atividades práticas nas disciplinas**: O documento destaca a necessidade de internalizar atividades práticas nos processos de avaliação dos alunos, por meio de planos de aula que contemplem conteúdos aplicados.

	2. **Oficinas interdisciplinares**: São previstas três oficinas ao longo do curso, com o objetivo de desenvolver atividades práticas e promover a interdisciplinaridade. Por exemplo, a Oficina I envolve professores de disciplinas como Saúde Ambiental, Avaliação de Impacto Ambiental e Poluição Atmosférica, onde os alunos realizam análises da qualidade ambiental de ambientes antropizados e propõem ações de intervenção. 3. **Atividades complementares**: A partir do 5º semestre, os alunos podem participar de atividades complementares (como projetos de pesquisa ou extensão), que são computadas no currículo pedagógico. Essas atividades são incentivadas pelos professores, que devem incluir pelo menos um projeto de pesquisa ou extensão em seus planos de trabalho. 4. **Participação em eventos técnico-científicos**: Os alunos podem participar de seminários, como o Seminário de Engenharia Ambiental e Sanitária, e outras atividades promovidas pelo colegiado do curso, com certificação comprovada. 5. **Infraestrutura para aulas práticas e laboratoriais**: O PPC prevê a necessidade de laboratórios e espaços específicos para a realização de aulas experimentais e práticas, como 7 laboratórios com equipamentos adequados. 6. **Enfoque na formação do egresso**: O curso busca aproximar os conteúdos do núcleo comum às Diretrizes Curriculares Nacionais dos Cursos de Graduação em Engenharia, garantindo uma formação alinhada às demandas profissionais

	Essas estratégias visam garantir que os alunos desenvolvam habilidades práticas e estejam preparados para os desafios da profissão.

Fonte: elaborado pela autora, 2026.

Embora as informações apresentadas estivessem corretas e bem organizadas, essa omissão influenciou principalmente os critérios de Aderência e Fidelidade, pois a resposta não contemplou todos os elementos esperados para responder integralmente à pergunta. Esse exemplo demonstra que respostas coerentes e bem escritas nem sempre são completas, reforçando a importância da avaliação humana para verificar se informações relevantes do documento não foram omitidas.

5 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo analisar, de forma qualitativa e comparativa, o desempenho de três modelos de linguagem de grande escala, LLaMA-3.3 70B versatile, Gemini-2.5-Flash e mistral-small-latest, na interpretação de Projetos Pedagógicos de Curso (PPCs) dos cursos de engenharia vinculados ao Bacharelado Interdisciplinar em Ciência e Tecnologia (BICT) da Universidade Federal do Maranhão (UFMA), a partir de uma rubrica analítica composta por cinco critérios de avaliação.

Para facilitar a pesquisa, foi desenvolvido um protótipo de interface web interativa utilizando o Streamlit, framework da linguagem Python, que permitiu o envio simultâneo dos prompts para os três modelos, o registro estruturado das respostas e a aplicação da rubrica de avaliação de forma sistemática. O protótipo também incorporou uma estratégia de recuperação de informação baseada em relevância (RAG), o que possibilitou recuperar trechos relevantes do documento de forma mais flexível do que uma estratégia baseada apenas em cortes fixos.

Após a coleta e avaliação de 150 respostas, 30 por curso, distribuídas entre os três modelos, os resultados indicaram que os modelos avaliados são capazes de interpretar e responder perguntas sobre PPCs com nível satisfatório de qualidade, com médias gerais acima de 3,5 em todos os critérios. O Gemini-2.5-Flash e o mistral-small-latest apresentaram desempenho superior ao LLaMA-3.3 70B versatile na maioria dos critérios, especialmente nos critérios de Aderência ao PPC e Profundidade, o que demonstra maior capacidade de se basear no conteúdo específico dos documentos e de explorar os temas com maior detalhamento.

O critério de Profundidade foi constantemente o mais baixo entre os três modelos, revelando uma limitação comum dos LLMs quando lidam com perguntas que exigem interpretação mais profunda de documentos extensos. Embora os modelos consigam identificar e reproduzir informações presentes no texto, eles apresentam maior dificuldade em sintetizar e aprofundar as informações implícitas ou estruturais do PPC. Essa limitação deve ser considerada em aplicações práticas que demandem análises mais críticas e contextualizadas dos documentos pedagógicos.

A análise por curso revelou que a estrutura do PPC e a objetividade influenciam diretamente no desempenho dos modelos. O PPC de Engenharia Mecânica, por exemplo, obteve as maiores médias entre todos os modelos, enquanto o de Engenharia da Computação apresentou os resultados mais baixos para o LLaMA-3.3

70B versatile. Esses resultados indicam que a qualidade da elaboração dos documentos pedagógicos pode potencializar ou limitar o uso de ferramentas de inteligência artificial para sua interpretação, sendo relevante para gestores e coordenadores de cursos que considerem incorporar essas tecnologias em processos de análise institucional.

A similaridade observada entre Gemini-2.5-flash e mistral-small-latest em vários critérios e cursos mostra que modelos de diferentes empresas, quando apresentados ao mesmo contexto de entrada, tendem a mostrar comportamentos similares em tarefas de interpretação de documentos. Esses resultados sugerem que, nas condições avaliadas nesta pesquisa, a organização do documento e a formulação das perguntas exerceram influência relevante sobre a qualidade das respostas geradas.

Em resumo, esta pesquisa demonstrou que os modelos de linguagem de grande escala demonstram potencial para auxiliar na interpretação e análise de documentos pedagógicos institucionais, como os PPCs, podendo contribuir para processos de avaliação curricular, autoavaliação institucional e suporte à gestão acadêmica. No entanto, a utilização desses modelos deve ser acompanhada também de avaliação humana criteriosa, especialmente quando os critérios exigem profundidade analítica e fidelidade às informações do documento.

Este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos resultados. A avaliação das respostas foi realizada por um único avaliador, o que pode introduzir vieses subjetivos na atribuição das notas, ainda que a rubrica tenha sido elaborada com descritores detalhados para cada nível de desempenho. Além disso, a estratégia de RAG adotada baseia-se em busca por palavras-chave, sem o uso de embeddings semânticos, o que pode limitar a recuperação de trechos relevantes quando o documento utiliza termos diferentes ou sinônimos para representar o mesmo conceito. Por fim, os resultados obtidos referem-se especificamente aos modelos e versões utilizados no período de coleta, podendo variar com atualizações futuras.

Com base nos achados desta pesquisa, sugerem-se que, em trabalhos futuros, sejam ampliadas as análises para outros cursos da UFMA, permitindo comparações mais abrangentes. Também é recomendada a inclusão de modelos mais recentes, como GPT-4o e Gemini 3, para verificar se os resultados observados se mantêm em versões mais avançadas. Além disso, podem ser utilizadas métricas automáticas de

avaliação da qualidade das respostas, como BLEU e ROUGE, complementando a avaliação realizada por meio da rubrica. Outra possibilidade consiste na implementação de uma estratégia de RAG semântico baseada em *embeddings*, buscando melhorar a recuperação do contexto e, consequentemente, a aderência e a profundidade das respostas.

Espera-se que os resultados desta pesquisa contribuam para ampliar as discussões sobre o uso de modelos de linguagem na interpretação de documentos institucionais, fornecendo subsídios para futuras pesquisas e para o desenvolvimento de ferramentas de apoio à gestão e à avaliação no ensino superior.

REFERENCIAS

ANJOS, Juliana Rodrigues dos *et al*. Uma análise do uso de IA generativa como investigadora em pesquisas qualitativas em educação científica. **Revista Pesquisa Qualitativa**, v. 12, n. 30, p. 1-29, 2024.

ASSUMPÇÃO, Heitor Dutra de; BRITO, Adriana Camargo de. Seleção de modelo de linguagem de redações do Enem. Estratégias & Soluções, dez., 2025. Disponível em: https://ipt.br/2026/02/03/selecao-de-modelo-de-linguagem-de-redacoes-do-enem/. Acesso em: 12 jul. 2026.

BAO, Y. *et al*. FaithBench: benchmarking LLM faithfulness in RAG. In: Proceedings of EMNLP Industry Track, 2025.

CASSAR, Peter; GALEA, Francesca; FERRY, Peter. Evaluating the Clinical Decision-Making Accuracy of Artificial Intelligence in Common Geriatric Syndromes Using Evidence-Based Guidelines. Cureus, 2026.

CARDOSO, Fábio Santos *et al*. O uso da Inteligência Artificial na Educação e seus benefícios: uma revisão exploratória e bibliográfica. **Revista Ciência em Evidência**, v. 4, n. FC, p. e023002-e023002, 2023.

CARVALHO, Sidartha A. L. de *et al*. QUIA: uma ferramenta de suporte ao conhecimento baseada em inteligência artificial. Simpósio Brasileiro de Informática na Educação (SBIE). 2024. p. 1617-1628.

CHANG, Yupeng *et al*. **A Survey on Evaluation of Large Language Models**. *ACM Transactions on Intelligent Systems and Technology*, v. 15, n. 3, artigo 39, p. 1–45, 2024

CIEB -CENTRO DE INOVAÇÃO PARA A EDUCAÇÃO BRASILEIRA. Notas técnicas #16: inteligência artificial na educação. São Paulo: CIEB, 2019. Disponível em: https://cieb.net.br/inteligencia-artificial-na-educacao/.Acesso em: 25 ago. 2025.

COPAC. Projetos pedagógicos de cursos (ppc) e curricularização da extensão. 2022.

GENTIL, Gabriel Felipe Baia; OLIVEIRA, Samuel Vinícius Pereira de. Proposição metodológica para comparação entre Engenharia de Prompt e Engenharia de Contexto na administração pública brasileira. **RevistaFT**, v. 29, ed. 152, nov. 2025.

INEP. Glossário dos Instrumentos de Avaliação Externa. 2018. Disponível em <https://download.inep.gov.br/educacao_superior/avaliacao_institucional/apresentacao/glossario_3_edicao.pdf.> Acesso em 31 de julho de 2025.

JIANG, Albert Q. *et al*. Mistral 7B. arXiv preprint arXiv:2310.06825, 2023.

KLOSS, Steffanie; CORDOVEZ-FERNÁNDEZ, Maximiliano; BUSTAMANTE, Cristóbal. Avaliação da qualidade da escrita em ensaios argumentativos: comparação entre Modelos Linguísticos Amplos (LLM) e avaliações humanas baseadas em uma rubrica. Texto Livre, Belo Horizonte-MG, v. 19, p. e63123, 2026.

LEWIS, Patrick *et al*. Retrieval-augmented generation for knowledge-intensive NLP tasks. **Advances in Neural Information Processing Systems**, v. 33, p. 9459-9474, 2020.

LIMA, Cleosanice Barbosa; SERRANO, Agostinho. Inteligência artificial generativa e ChatGPT: uma investigação sobre seu potencial na Educação. Transinformação, Campinas, v. 36, e2410839, p. 1-12, 2024.

LIU, Tongxi; YE, Luyao; YAN, Wei. [Linguistic features and domain classification in LLM writing]. ***Computers and Education: Artificial Intelligence***, v. 10, 2026.

MEDEIROS, Cláudio Bezerra de. Projeto Pedagógico de Curso – PPC. Brasília: Decanato de Ensino de Graduação (DEG/UnB), 4 set. 2024. Disponível em: https://deg.unb.br/faq_ppc/. Acesso em: 01 jul. 2026.

MIRANDA, Gina Magali Horvath; ROSINI, Alessandro Marco; MACEDO**,** Elida Pereira. Potencializando a aprendizagem com IA: ChatGPT como suporte ao professor universitário na era dos grandes desafios avaliativos. **REPAE – Revista de Ensino e Pesquisa em Administração e Engenharia**, v. 11, n. 1, p. 52-64, 2025.

MONTES, Cristina Martinez *et al*. Large Language Models in Thematic Analysis: Prompt Engineering, Evaluation, and Guidelines for Qualitative Software Engineering Research. Gothenburg: Chalmers University of Technology, 2025.

MUGAANYI, Joseph *et al.* Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. **Journal of Medical Internet Research**, v. 26, e52935, 2024.

MÜLLER, William Henrique. A evolução e a regulamentação da inteligência artificial no brasil. **Revista InterCiência-IMES Catanduva**, v. 1, n. 11, p. 2-2, 2023.

OLIVEIRA, Laize Almeida de *et al*. Inteligência Artificial na Educação: uma revisão integrativa da literatura. **Peer Review**, v. 5, n. 24, p. 249–268, 2023.

PEREIRA, Ana Lúcia; BONIN, Carlos Alberto; SOARES, Carlos Eduardo Krassinski. Utilizando a Inteligência Artificial como apoio na análise de conteúdo: algumas considerações. v. 9, n. 25, p. 258-282, 2024.

RODRIGUES, Carlos Gabriel de Castro. Um benchmark para avaliação de grandes modelos de linguagem no português usando questões do exame nacional do ensino médio. 2025. TCC — Universidade Federal do Ceará, Fortaleza, 2025.

SAMPAIO, Rafael Cardoso; SABBATINI, Marcelo; LIMONGI, Ricardo. Diretrizes para o uso ético e responsável da Inteligência Artificial Generativa: um guia prático para pesquisadores. São Paulo: Intercom, 2024.

SANTANA, Fernanda Pereira; MAGALHÃES, Leandro Corrêa. Aplicações do processamento de linguagem natural no ambiente educacional: Uma revisão sistemática da literatura**. Revista Foco**, v. 17, n. 1, p. e3921-e3921, 2024.

SÁNCHEZ, Francisco Jiménez. Desarrollo de una aplicación de inteligencia artificial corporativa para desarrolladores. 2024.

SILVA, Keila Ramos da *et al*. Inteligência artificial e seus impactos na educação: uma revisão sistemática. **RECIMA21-Revista Científica Multidisciplinar-ISSN 2675-6218**, v. 4, n. 11, p. e4114353-e4114353, 2023.

SILVA, Maria Juliana Farias; PAULA, Marlúbia Corrêa de. Perspectivas da inteligência artificial como ferramenta de apoio para análise textual discursiva. ***Revista Pesquisa Qualitativa***, v. 12, n. 30, p. 1-26, 2024.

SILVA, Vinícius Facin da. Assistente Virtual para Coordenador de Curso: Aplicação de LLMs à Gestão de Perguntas Frequentes em um Curso de Graduação. 2025. 71 f. TCC (Graduação em Engenharia de Controle e Automação) - Universidade Federal de Santa Catarina, Blumenau, 2025.

SILVA, Luís Felipe Mattos da. Análise e Visualização de Trabalhos de Conclusão de Curso utilizando Processamento de Linguagem Natural: Um Método Baseado Grandes Modelos de Linguagem. 2025.

SHUKLA, Mukesh *et al*. Evaluating the accuracy and explanatory quality of large language models ChatGPT, Claude, DeepSeek, Gemini, Grok, and Le Chat in statistical test selection for hypothesis testing decisions. Cureus, v. 17, n. 10, e94949, 2025.

STRYKKER, Cole; KAVLAKOGLU, Eda. O que é inteligência artificial? 2024. Disponível em <https://www.ibm.com/br-pt/think/topics/artificial-intelligence> Acesso em 30 de julho de 2025.

UNIVERSIDADE FEDERAL DO CEARÁ. Pró-Reitoria de Graduação. Coordenadoria de Projetos e Acompanhamento Curricular (COPAC). Documento orientador para elaboração de PPC – Projeto Pedagógico de Curso. Fortaleza: UFC, abr. 2022.

Disponível em: https://prograd.ufc.br/wp-content/uploads/2022/04/documento-orientador-ppc.pdf. Acesso em: 01 jul. 2026.

UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL. Pró-Reitoria de Graduação. Projetos Pedagógicos. Campo Grande: UFMS, 2021. Disponível em: https://prograd.ufms.br/projetos-pedagogicos-cursos/. Acesso em: 01 jul. 2026.

WANG, Yuehan *et al*. Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. ***Computers and Education: Artificial Intelligence***, v. 9, 2025.

YAVUZ, Fatih; ÇELIK, Özgür; ÇELIK, Gamze Yavaş. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. ***British Journal of Educational Technology***, v. 56, p. 150-166, 2024.

ZHANG, Ting *et al*. Evaluating generative AI chatbots for largescale assessment data: comparing LLM-as-a-judge and human ratings. ***Large-scale Assessments in Education***, v. 14, 2026.

APÊNDICE A – Perguntas utilizadas na avaliação dos modelos de linguagem

Nº	Categoria	Pergunta
1	Curso e objetivos	Qual é a principal proposta deste curso de engenharia?
2	Curso e objetivos	O que o curso espera que os alunos saibam ao se formar?
3	Curso e objetivos	De que maneira o curso busca equilibrar teoria e prática na formação do estudante?
4	Curso e objetivos	Qual é a importância do curso para a sociedade, segundo o PPC?
5	Perfil do egresso	Como é o profissional que o curso quer formar?
6	Perfil do egresso	Como o estudante é incentivado a se desenvolver durante o curso?
7	Metodologia e avaliação	As metodologias de ensino propostas no PPC incentivam a autonomia e o pensamento crítico dos alunos?
8	Metodologia e avaliação	O PPC apresenta estratégias para aproximar os estudantes da prática profissional? Quais?
9	Impacto social e valores	O aluno é incentivado a pensar em problemas reais?
10	Estrutura curricular	Como a estrutura curricular está organizada ao longo da formação do estudante?

Fonte: elaborada pela autora, 2026.