LUMEN ET VIRTUS
ISSN: 2177-2789

AVALIAÇÃO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL GENERATIVA NO ENSINO DE LÓGICA DE PROGRAMAÇÃO: UMA ANÁLISE COMPARATIVA ENTRE OS MODELOS GEMMA E META LLAMA

EVALUATION OF GENERATIVE ARTIFICIAL INTELLIGENCE MODELS IN TEACHING PROGRAMMING LOGIC: A COMPARATIVE ANALYSIS BETWEEN THE GEMMA AND META LLAMA MODELS

EVALUACIÓN DE MODELOS DE INTELIGENCIA ARTIFICIAL GENERATIVA EN LA ENSEÑANZA DE LA LÓGICA DE PROGRAMACIÓN: UN ANÁLISIS COMPARATIVO ENTRE LOS MODELOS GEMMA Y METALLAMA

https://doi.org/10.56238/levv17n59-014

Data de submissão: 08/03/2026 **Data de publicação:** 08/04/2026

Henrique Augusto Santos Matos
Graduando em Sistemas de Informação
Instituição: Centro Universitário Santa Terezinha (CEST)
E-mail: augusto.h.s.matos@hotmail.com
ORCID: https://orcid.org/0009-0007-6520-7255

Jefferson Sousa Sampaio Júnior
Graduando em Sistemas de Informação
Instituição: Centro Universitário Santa Terezinha (CEST)
E-mail: jeffersosnamp2012@gmail.com
ORCID: https://orcid.org/0009-0004-1476-0323

Aline Lopes da Silva
Mestre em Ciências da Computação (Computação Móvel)
E-mail: aline.lopess@cest.edu.br
ORCID: https://orcid.org/0009-0005-3447-5471

Dadilton Bastos Melo
Especialista em Ciência de Dados e Big Data Analytics
E-mail: dadilton.melo@cest.edu.br
ORCID: https://orcid.org/0009-0000-3673-881

RESUMO

Este estudo tem como objetivo comparar o desempenho de modelos de Inteligência Artificial Generativa (IAG), especificamente o Gemma e o Meta LLaMA, no contexto do ensino de lógica de programação. A pesquisa busca analisar qual modelo apresenta melhor desempenho considerando critérios como tempo de resposta, consistência textual e adequação pedagógica. Para isso, foi utilizada uma abordagem experimental com a execução dos modelos em ambiente local por meio do software LM Studio, permitindo o acesso às suas APIs. As métricas adotadas para avaliação incluíram BLEU, ROUGE, METEOR e BERTScore, além da mensuração do tempo de resposta em milissegundos. Adicionalmente, foi realizada uma avaliação subjetiva com a participação de 30 estudantes do curso de Sistemas de Informação, que analisaram respostas geradas pelos modelos com base em critérios de

clareza, objetividade e utilidade pedagógica. Os resultados indicam que o Meta LLaMA apresentou melhor desempenho em termos de eficiência computacional e similaridade estrutural das respostas, enquanto o modelo Gemma demonstrou maior riqueza semântica e capacidade explicativa em contextos que exigem aprofundamento conceitual. A avaliação humana corroborou esses achados, evidenciando preferência pelo Meta LLaMA em questões objetivas e pelo Gemma em perguntas que demandam explicações detalhadas. Conclui-se que os modelos possuem características complementares, sendo recomendada a utilização combinada para potencializar o processo de ensino-aprendizagem em lógica de programação.

Palavras-chave: Inteligência Artificial Generativa. Ensino de Programação. Lógica de Programação. Modelos de Linguagem. Avaliação Comparativa.

ABSTRACT

This study aims to compare the performance of Generative Artificial Intelligence (GAI) models, specifically Gemma and Meta LLaMA, in the context of teaching programming logic. The research seeks to analyze which model presents better performance considering criteria such as response time, textual consistency, and pedagogical adequacy. To achieve this, an experimental approach was adopted, with the models being executed in a local environment using the LM Studio software, allowing access to their APIs. The evaluation metrics included BLEU, ROUGE, METEOR, and BERTScore, in addition to measuring response time in milliseconds. Furthermore, a subjective evaluation was conducted with the participation of 30 students from the Information Systems course, who analyzed the responses generated by the models based on criteria such as clarity, objectivity, and pedagogical usefulness. The results indicate that Meta LLaMA presented better performance in terms of computational efficiency and structural similarity of responses, while the Gemma model demonstrated greater semantic richness and explanatory capacity in contexts that require deeper conceptual understanding. The human evaluation corroborated these findings, showing a preference for Meta LLaMA in objective questions and for Gemma in questions that require more detailed explanations. It is concluded that the models have complementary characteristics, and their combined use is recommended to enhance the teaching-learning process in programming logic.

Keywords: Generative Artificial Intelligence. Programming Education. Programming Logic. Language Models. Comparative Evaluation.

RESUMEN

Este estudio tiene como objetivo comparar el rendimiento de los modelos de Inteligencia Artificial Generativa (IAG), específicamente Gemma y Meta LLaMA, en el contexto de la enseñanza de la lógica de programación. La investigación busca analizar qué modelo ofrece un mejor rendimiento considerando criterios como el tiempo de respuesta, la consistencia textual y la idoneidad pedagógica. Para ello, se utilizó un enfoque experimental, ejecutando los modelos en un entorno local mediante el software LM Studio, con acceso a sus API. Las métricas adoptadas para la evaluación incluyeron BLEU, ROUGE, METEOR y BERTScore, además de medir el tiempo de respuesta en milisegundos. Adicionalmente, se realizó una evaluación subjetiva con la participación de 30 estudiantes del curso de Sistemas de Información, quienes analizaron las respuestas generadas por los modelos según criterios de claridad, objetividad y utilidad pedagógica. Los resultados indican que Meta LLaMA tuvo un mejor rendimiento en términos de eficiencia computacional y similitud estructural de las respuestas, mientras que el modelo Gemma demostró mayor riqueza semántica y capacidad explicativa en contextos que requieren profundidad conceptual. La evaluación humana corroboró estos hallazgos, mostrando una preferencia por Meta LLaMA en preguntas objetivas y por Gemma en preguntas que requerían explicaciones detalladas. Se concluye que los modelos poseen características complementarias y se recomienda su uso combinado para mejorar el proceso de enseñanza-aprendizaje en lógica de programación.

Palabras clave: Inteligencia Artificial Generativa. Educación en Programación. Lógica de Programación. Modelos de Lenguaje. Evaluación Comparativa.

1 INTRODUÇÃO

Com o avanço da Inteligência Artificial (IA), sua aplicação dentro da educação tem se tornado cada vez mais relevante, transformando a maneira como os alunos aprendem e interagem com o conhecimento. Tecnologias baseadas em IA, especialmente os modelos generativos, têm demonstrado grande potencial para personalizar o ensino, gerar conteúdos sob demanda e oferecer feedbacks imediatos, criando experiências de aprendizagem mais dinâmicas e centradas no aluno (Luckin et al., 2016).

No ensino da programação, essa ferramenta se torna particularmente promissora. A lógica de programação é uma etapa fundamental para formação de desenvolvedores e profissionais de tecnologia, ainda que represente principais desafios no ensino superior, ainda mais para estudantes iniciantes. A dificuldade em compreender abstrações, estruturas de controle e até mesmo raciocínio lógico contribui para reprovação e evasão nos cursos da área de computação. Segundo Gomes e Mendes (2007) , muitos alunos enfrentam barreiras cognitivas ao tentar compreender conceitos abstratos como variáveis, estruturas condicionais e loops, o que contribui para altos índices de reprovação e evasão.

Nesse contexto, ferramentas baseadas em Inteligência Artificial Generativa (IAG), especialmente modelos de linguagem natural, surgem como alternativas promissoras para apoiar o ensino desses conceitos. De acordo com Zawacki-Richter et al (2019), elas possibilitam a geração de exemplos, explicações personalizadas e simulações interativas que favorecem o raciocínio lógico e a fixação do conteúdo. Portanto, o questionamento que conduz a pesquisa é: **Qual dos modelos de Inteligência Artificial Generativa, Gemma e Meta LLaMA apresentam melhor desempenho em termos de tempo de resposta, consistência textual e adequação pedagógica no apoio ao ensino de lógica de programação para estudantes de Sistemas de Informação?**

A seleção dos modelos Gemma e Llama justificam-se pela posição pioneira no que se refere a modelos com pesos abertos (Google DeepMind, 2024; Dubey et al., 2024). Enquanto o LLaMa oferece uma base de conhecimento robusta devido à escala massiva do seu pré-treinamento (Dubey et al., 2024), o Gemma destaca-se pela eficiência arquitetural e desempenho superior em tarefas de raciocínio lógico em escalas menores de parâmetros. A utilização de ambos permite uma análise comparativa entre diferentes filosofias de otimização de LLMs, garantindo a reprodutibilidade dos experimentos em ambientes de computação locais (Liesenfeld et al., 2023).

Por isso, o crescimento do uso dessas tecnologias por estudantes, torna-se necessário compreender seu impacto real no processo de ensino-aprendizagem, especialmente em áreas que exigem raciocínio lógico estruturado, como a programação. Neste contexto, a pesquisa propõe uma análise comparativa entre duas ferramentas baseadas em Inteligência Artificial Generativa, Gemma e Meta LLaMA com base em critérios como tempo de resposta, consistência das respostas e avaliação

humana. Portanto, a análise busca identificar o potencial dessas ferramentas como recursos complementares no ambiente educacional, avaliando sua eficácia técnica na mediação do conhecimento entre os estudantes.

2 TRABALHOS RELACIONADOS

2.1 A INTELIGÊNCIA ARTIFICIAL GENERATIVA NO ENSINO DE PROGRAMAÇÃO

Conforme discutido por Silva et al(2024) em seu trabalho, o ensino de programação é complexo e desafiador para os alunos, no que se trata de assimilação de conceitos mais abstratos. Nesse sentido, o uso da Inteligência Artificial Generativa (IAG) tem sido explorado para oferecer recursos como suporte personalizado, feedback imediato e geração de exemplos práticos, destacando sua eficiência no que se refere a simplificar o processo de ensino.

Sendo assim, pesquisas recentes, como as de Becker et al. (2023) e Denny et al. (2024), ressaltam que as ferramentas de IAG oferecem diversas alternativas no suporte no ensino de programação aos alunos, como: ferramentas que podem gerar exemplos e testes de código, a apresentação de diferentes maneiras de resolver um problema, ferramentas que podem criar um exercício a partir de um único exemplo e construção de prompts inserido nas ferramentas que influenciam seu desenvolvimento e entre outros.

Além disso, Becker et al. (2023) destacam que tais ferramentas de IA também geram preocupações relacionadas à integridade acadêmica e à reutilização de código, isso inclui: a utilização da IAG por alunos para superar provas e desafios a fim de gerar código, intensificando desonestidade acadêmica, a falta de segurança da geração do código pela IAG, o excesso do uso levando a dependência extrema na ferramenta e entre outros.

2.2 ANÁLISE COMPARATIVA DE MODELOS DE INTELIGÊNCIA ARTIFICIAL GENERATIVA APLICADOS AO ENSINO

O estudo de Joyce Silva (2024) analisou comparativamente modelos de IA como ChatGPT, Copilot e Gemini no ensino introdutório de programação. A autora observou que o ChatGPT apresentou as explicações mais claras e coerentes, sendo eficaz para alunos iniciantes, enquanto o Copilot se destacou na geração automática de código, embora com menor foco pedagógico e contextual. Essa análise reforça a importância de avaliar o potencial didático das ferramentas e não apenas seu desempenho técnico.

De modo complementar, Marques (2024) investigou o uso do Gemini no ensino de matemática, evidenciando que seu diferencial está na interação multimodal e na contextualização visual de conceitos, o que favorece a compreensão de conteúdos abstratos. Contudo, a autora também apontou

limitações quanto à precisão lógica, especialmente em tarefas que exigem raciocínio estruturado, o que sugere a necessidade de acompanhamento docente no uso dessas tecnologias.

No caso do Gemma, desenvolvido pela Google DeepMind, estudos apontam que o modelo combina eficiência e leveza, sendo ideal para contextos educacionais com infraestrutura limitada (Google AI, 2024; DeepMind, 2024). Seu desempenho equilibrado entre clareza textual e rapidez o torna uma opção promissora para o ensino de lógica de programação, permitindo explicações passo a passo e geração de exercícios adaptados ao nível do aluno.

Já o Meta LLaMA se destaca pela flexibilidade e personalização, com versões abertas e multimodais que possibilitam integração em plataformas educacionais e sistemas de avaliação automática (Meta AI, 2024; 2025). Seu diferencial está na capacidade de gerar exemplos em múltiplas linguagens e corrigir erros de lógica, o que favorece a aprendizagem prática. Contudo, sua complexidade técnica exige mediação docente para garantir uso ético e orientado. Além disso o Meta LLaMA, conforme Brilhante (2025), apresenta como principal vantagem a execução local de modelos de linguagem, oferecendo privacidade e controle de dados. Essa característica o torna atrativo para ambientes educacionais que buscam independência de nuvem. Porém, sua configuração técnica pode representar um desafio para docentes com pouca familiaridade em infraestrutura computacional.

O Copilot, segundo Silva (2024), mostra-se eficiente para usuários com algum domínio em programação, oferecendo suporte direto na escrita e depuração de códigos. No entanto, seu caráter mais voltado ao desenvolvimento profissional o torna menos indicado para iniciantes, uma vez que não fornece explicações pedagógicas detalhadas.

O Google Studio AI destaca-se por sua interface intuitiva e integração com o ecossistema Google, permitindo criar fluxos interativos e atividades práticas com IAs personalizadas (Silva, 2024). Essa ferramenta é especialmente útil na criação de avaliações formativas e feedbacks automáticos, promovendo um aprendizado mais ativo e dinâmico.

Tabela 1 – Quadro Comparativo

Modelos de Linguagem	Interação Natural	Geração de Código	Personalização	Interface Visual	Uso Offline	Acessibilidade	Fonte
ChatGPT	Alta	Média	Média	Média	Não	Alta	Silva (2024, p. 48)
Gemini	Alta	Média	Média	Alta	Não	Alta	Silva (2024, p. 48); Simões (2024, p.5)
Copilot	Média	Alta	Baixa	Baixa	Parcial	Média	Silva (2024, p. 48)
Meta LLaMA	Alta	Alta	Alta	Alta (em versões multimodal)	Sim	Alta	Meta LLaMA blog / model cards; Hugging Face.
Gemma	Alta	Alta	Média	Alta	Sim	Alta	Gemma model card / technical report / DeepMind blog.
Ollama	Média	Alta	Alta	Baixa	Sim	Baixa	Brilhante (2025, p. 24-25)
Google Studio AI	Alta	Alta	Alta	Alta	Não	Média	Silva (2024, p. 5)

Fonte: Autoria Própria

De acordo com a Tabela 1, a Inteligência Artificial Generativa (IAG) tem se consolidado como uma tecnologia promissora para apoiar o ensino, especialmente na área de computação. Modelos como ChatGPT, Gemini, Meta LLaMA e Google Studio AI têm sido aplicados para objetivos variados, que incluem desde a geração automática de código até a personalização de conteúdos e o fornecimento de feedback educacional. Estudos recentes indicam um crescimento significativo no uso de agentes de IA, chatbots e ferramentas generativas no ensino de programação, com foco no suporte ao aprendizado, na melhoria do desempenho discente e na adaptação do ensino às necessidades individuais dos estudantes (ELNAFFAR et al., 2025).

Tabela 2 – Quadro Comparativo Meta LLaMA x Gemma

Critério	Meta Llama	Gemma
Tipo de Execução	Local ou em nuvem; pesos públicos disponíveis	Local, eficiente em dispositivos leves ou cloud Google
Privacidade e Controle	Alto, conforme licença	Alto, com foco em execução segura e leve
Integração com Aplicações	Alta (Hugging Face, ferramentas open-source)	Alta (Google AI Studio e ecossistema Google)
Modelo Utilizado	LLaMA (8B, 70B, 405B)	Gemma 3 (1B, 4B, 12B, 27B)
Objetivo Principal	Modelo de uso geral, pesquisa e IA aberta	Modelo leve e multimodal com foco em eficiência
Interface e Facilidade	Média-Alta, depende de suporte técnico	Alta, com guias e APIs próprias do Google
Multimodalidade	Alta (versões recentes com visão)	Alta (texto + imagem nativo)
Atualizações e Limitações	Atualizações frequentes, licença discutível	Em atualização ativa, pouco explorado pedagogicamente
Customização e Treinamento	Alta, com suporte a fine-tuning	Média-Alta, conforme infraestrutura

Fonte: Autoria Própia

A comparação apresentada na Tabela 2 revela que nenhuma ferramenta é completa em todos os aspectos. Enquanto o Meta LLaMA se destaca pela ampla customização, suporte a fine-tuning e multimodalidade, a Gemma oferece execução eficiente em dispositivos leves, integração fácil com o ecossistema Google e interface amigável. Assim, a escolha da ferramenta mais adequada dependerá do objetivo educacional, da infraestrutura disponível e do perfil dos usuários. Em alguns casos, a combinação dessas ferramentas pode ser vantajosa, permitindo aproveitar os pontos fortes de cada uma para suprir limitações específicas.

3 MATERIAIS E MÉTODOS

Para o desenvolvimento deste trabalho, realizou-se a comparação entre dois modelos de linguagem de grande porte (Large Language Models – LLMs): Gemma e Meta LLaMA., a partir do software LM Studio que permite o uso da Application Programming Interface (API) para manipulação com os modelos a fim de entender qual dos modelos apresenta o maior potencial como ferramenta de ensino e aprendizagem, utilizando métricas como tempo de resposta em milissegundos, consistência textual com ROUGE (LIN, 2004), BLEU (PAPINENI et al., 2002), METEOR (BANERJEE; LAVIE, 2005) e BERTScore (ZHANG et al., 2019).

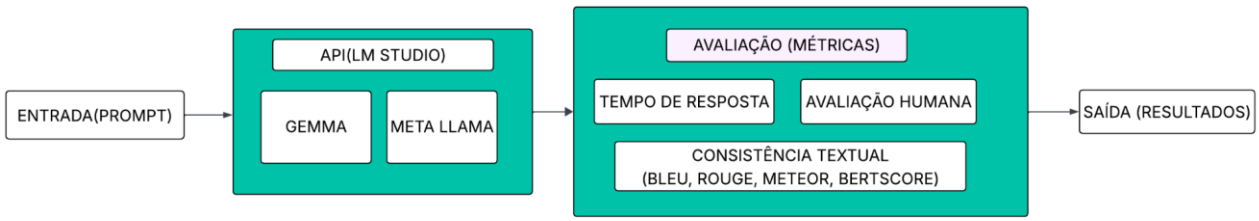
Do mesmo modo com métrica de avaliação humana, a partir da criação de um questionário eletrônico, com 30 estudantes de sistemas de informação, contendo cinco questões com apenas 2 alternativas cada, que atende a uma análise subjetiva em cima das resposta geradas pelos modelos,

com o objetivo de entender qual modelo, a partir das respostas geradas, é mais adequado e que satisfaça cada questão.

Inicialmente, para atender às métricas estabelecidas, especialmente aquelas relacionadas ao tempo de resposta e à consistência, foi realizada a utilização do LM Studio para escolha dos modelos e acesso à sua API. No entanto, para o estudo as LLMs escolhidas foram o google/gemma-2-9b e meta-llama-3.1-8b-instruct, considerando sua performance fluída e capacidade de gerar respostas assertivas.

Soma-se a isso o ambiente utilizado para execução dos modelos, o hardware composto por processador Intel Core i7-4700MQ com frequência base de 2.40 GHz, 16 GB de memória RAM DDR3 e sistema operacional de 64 bits, além de uma execução com suporte de GPU dedicada NVIDIA GeForce GT 740M com 2 GB de memória, além da GPU integrada Intel HD Graphics 4600. E também uso dos modelos Meta-Llama-3.1-8B-Instruct, na versão quantizada Q4_K_S (4 bits) de 4.69 GB, e Google Gemma-2-9B, na versão quantizada Q3_K_L(3 bits) de 5.13 GB, conforme mostrada na Figura 1.

Figura 1 - Diagrama de Arquitetura do Software



Fonte: Autoria própria

Em seguida na escolha das métricas dentro de consistência, o estudo explorou algoritmos como ROUGE (LIN, 2004), BLEU (PAPINENI et al., 2002), METEOR (BANERJEE; LAVIE, 2005) e BERTScore (ZHANG et al., 2019) que foram de suma importância para mensurar a consistência e verificação da precisão das respostas geradas. Com base nisso, para avaliação das respostas do modelo com uso das métricas foi necessário um gabarito com respostas pré-estabelecidas e criação de diferentes prompts.

Sendo assim, a construção de prompts seguiu as categorias pedagógicas sugeridas (Conceitual, Aplicação, Explicação e Generalização) e foi baseou-se no estudo de Wang et al. (2024), conforme a Tabela 3, cujo objetivo é a elaboração de estratégias de engenharia de prompt que variam do nível mais básico ao mais específico, a fim de reproduzir a perspectiva do aluno na formulação de perguntas e analisar resposta dos modelos as questões.

Tabela 3 - Prompts para avaliação dos modelos

Categoria	Tipo de Prompt	Exemplo de Prompt Reformulado
Conceitual	Simples e direto	“O que é uma variável em programação?”
Generalização	Médio, com solicitação de exemplo	“Explique o que é uma variável e mostre um exemplo em Python.”
Aplicação	Focado em lógica aplicada	“Como posso usar uma variável para armazenar a soma de dois números em Python?”
Avaliação de Ambiguidade	Vago	“Como usa variável?”
Explicação	Detalhado	“Explique o que é uma variável em Python, para que ela serve, e mostre um exemplo com comentários explicando cada linha.”

Fonte: Adaptação de Wang et al.(2024)

Para atender os objetivos do estudo, foi implementado um algoritmo em Python que permite automatizar o processo de geração de respostas e cálculos de métricas de correspondência com o gabarito pré-estabelecido. Dessa forma, o algoritmo importa bibliotecas como LM studio para comunicação com modelo, mas também a biblioteca Time para cálculo de tempo em milissegundos, além da biblioteca Natural Language Toolkit (NTLK) para processamento de linguagem natural e cálculo de métricas com algoritmos citados anteriormente.

Portanto, o script lê a lista de prompts e suas respectivas respostas em um laço de repetição e faz o cálculo das métricas comentadas anteriormente na qual é executado em 30 vezes, registrando-se as informações como tempo de resposta em milissegundos e consistência textual, como mostrado na Figura 2.

Figura 2 - Fluxo de Avaliação dos Modelos

```
para x de 1 até 30 faça
   inicio ← marcarTempo()  // iniciar cronômetro(captura tempo inicial)

   resposta ← gerarRespostaDoModelo(prompt)

   fim ← marcarTempo()              // para cronômetro (captura tempo final)
   tempoResposta ← (fim - inicio) * 1000  // tempo em milissegundos

   referencia ← gabarito[x]         // usa o índice do laço

   resposta_tokens ← separarPorPalavras(resposta) // Separa cada palavra das respostas do modelo
   referencia_tokens ← separarPorPalavras(referencia) // Separa cada palavra da referencia

  // Cálculo de Métricas
   aplicarSuavizacaoBLEU()
   valorBLEU ← calcularBLEU([referencia_tokens], resposta_tokens,
   pesos=(0.25,0.25,0.25,0.25))
   valorMETEOR ← calcularMETEOR([referencia_tokens], resposta_tokens)
   valorROUGE ← calcularROUGE(referencia, resposta)

   valorBERT ← calcularBERTScore([resposta], [referencia], idioma="pt")

   registrarResultado(x, tempoResposta, valorBLEU, valorMETEOR, valorROUGE, valorBERT)

fim-para
```

Fonte: Autoria própria

4 RESULTADOS

A análise dos resultados dos modelos Gemma e Meta LLaMA, permitiu avaliar o desempenho de cada LLM sob duas perspectivas: (1) métricas objetivas de desempenho computacional e consistência textual, no item 4.1, e (2) avaliação subjetiva feita por estudantes de Sistema de Informação, conforme o item 4.2.

4.1 QUADROS DE ANÁLISE DAS RESPOSTAS SUBMETIDAS ÀS IAGS

Dado aos resultados obtidos a partir dos experimentos com os modelos conforme a Figura 3, mostram que o Meta LLaMA apresentou tempo médio de resposta de 1,98 minutos, sendo aproximadamente 41,5% mais rápido que o Gemma (3,39 minutos). Essa diferença demonstra maior eficiência computacional e potencial de aplicação em ambientes educacionais que exigem respostas imediatas.

Figura 3 – Resultados Finais dos Experimentos

GEMMA								
QUESTOES	Tempo de Resposta (ms)	Tempo de Resposta (minutos)	BLEU	METEOR	ROUGE1	ROUGE2	ROUGEL	BERTScore_F1
1ª PERGUNTA	217190,14	3,62	0,0095	0,2065	0,2184	0,0802	0,1335	0,6887
2ª PERGUNTA	175887,49	2,93	0,0089	**0,1861**	**0,1051**	**0,0382**	**0,0861**	0,6153
3ª PERGUNTA	147719,55	2,46	0,0339	0,2957	0,2399	0,1211	0,1791	0,6742
4ª PERGUNTA	205038,05	3,42	**0,0060**	**0,1613**	**0,1114**	**0,0331**	**0,0827**	**0,6592**
5ª PERGUNTA	270793,21	4,51	0,0109	**0,2146**	**0,2748**	0,0678	0,1669	0,6740
TOTAL GEMINI	203325,69	3,39	0,0138	0,2129	**0,1899**	0,0681	0,1296	0,6623

LLAMA								
QUESTOES	Tempo de Resposta (ms)	Tempo de Resposta (minutos)	BLEU	METEOR	ROUGE1	ROUGE2	ROUGEL	BERTScore_F1
1ª PERGUNTA	106964,63	1,78	**0,0280**	**0,2554**	**0,2352**	**0,1120**	**0,1656**	**0,7161**
2ª PERGUNTA	122367,44	2,04	**0,0115**	0,1660	0,0942	0,0369	0,0755	**0,6315**
3ª PERGUNTA	87255,81	1,45	**0,0496**	**0,3121**	**0,2476**	**0,1452**	**0,2061**	**0,6982**
4ª PERGUNTA	141369,09	2,36	0,0056	0,1266	0,0992	0,0275	0,0766	0,6546
5ª PERGUNTA	137062,84	2,28	**0,0192**	0,2117	0,2589	**0,0742**	**0,1691**	**0,6799**
TOTAL LHAMA	**119003,96**	**1,98**	**0,0228**	**0,2144**	0,1870	**0,0792**	**0,1386**	**0,6761**

Fonte: Autoria própria

Além disso, nas métricas de consistência textual, o Meta LLaMA obteve médias superiores em BLEU (0,0228) e ROUGE-2 (0,0792), enquanto o Gemma apresentou maior desempenho, como apresentado na Tabela 4, em METEOR (0,2129) e ROUGE-L (0,1296), o que indica maior clareza semântica e fluidez textual.

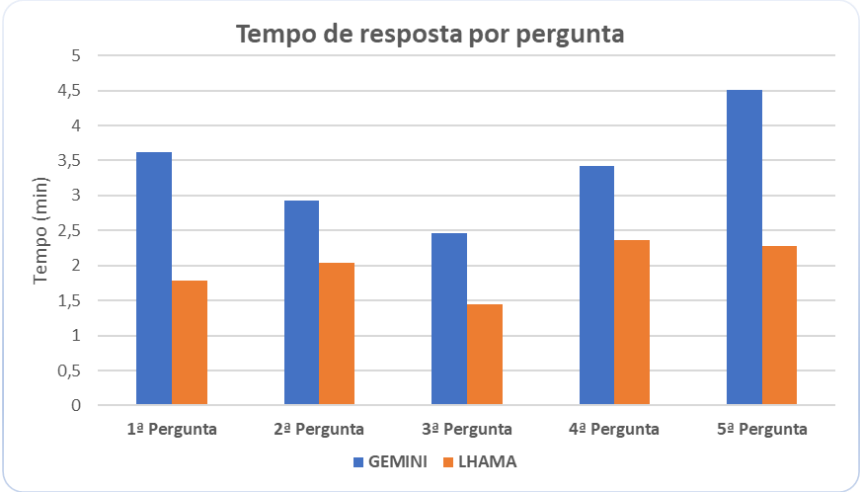
Tabela 4 - Médias das métricas de desempenho dos modelos Gemma e Meta LlaMA

Métrica	Gemma	Meta Llama
Tempo de Resposta(min)	3,39	1,98
BLEU	0,0138	0,0228
METEOR	0,2129	0,2144
ROUGE-1	0,1899	0,1870
ROUGE-2	0,0681	0,0792
ROUGE-L	0,1296	0,1386
BERTscore-F1	0,6623	0,6761

Fonte: Autoria própria

Dado a análise quantitativa entre a diferença dos modelos, tem-se a análise dos gráficos com as seguintes métricas:

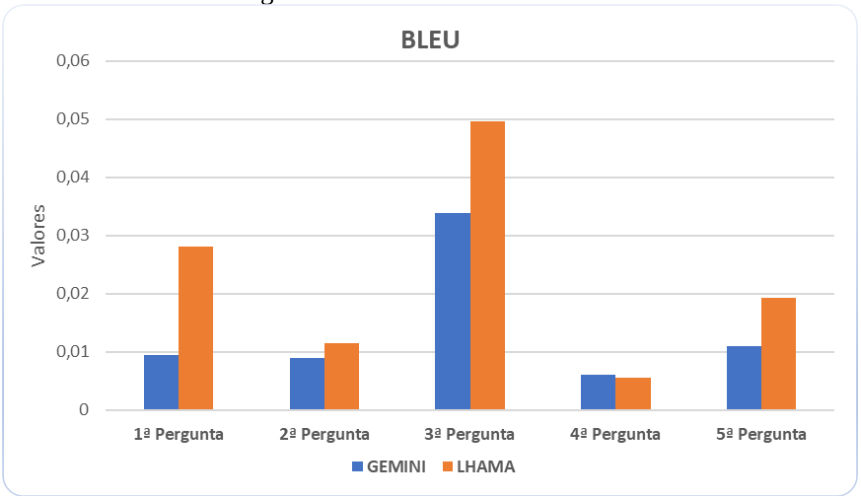
Figura 4 – Análise do Gráfico: Tempo de Resposta



Fonte: Autoria própria

O tempo de resposta, ilustrado na Figura 4, demonstra que o Meta LLaMA apresentou desempenho superior em todas as questões avaliadas, com tempos significativamente menores quando comparado ao Gemma. Esse resultado indica maior eficiência computacional do Meta LLaMA, tornando-o mais adequado para cenários educacionais que demandam respostas rápidas, como atividades síncronas ou suporte imediato ao aluno.

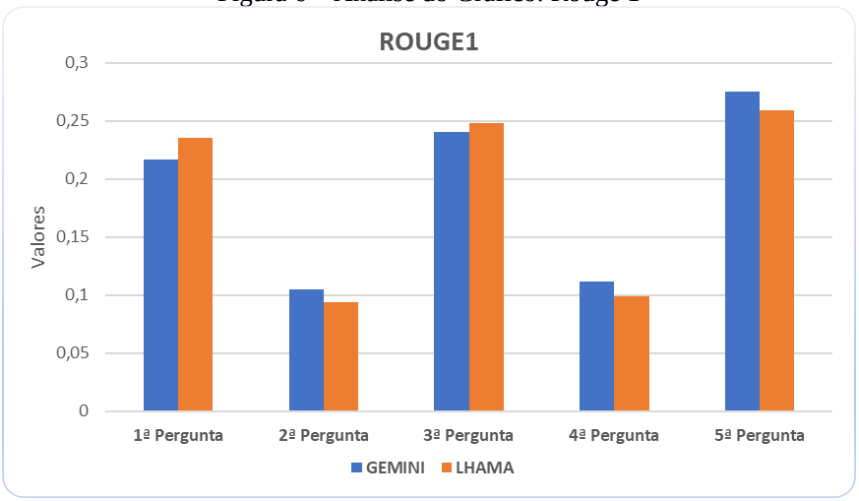
Figura 5 – Análise do Gráfico: Bleu



Fonte: Autoria própria

No que se refere à métrica BLEU, apresentado na Figura 5, observou-se que o Meta LLaMA obteve valores mais elevados na maioria das questões, indicando maior similaridade lexical entre as respostas geradas e o gabarito de referência. Esse comportamento sugere que o modelo tende a produzir respostas mais próximas do texto esperado, com maior precisão terminológica e estrutural.

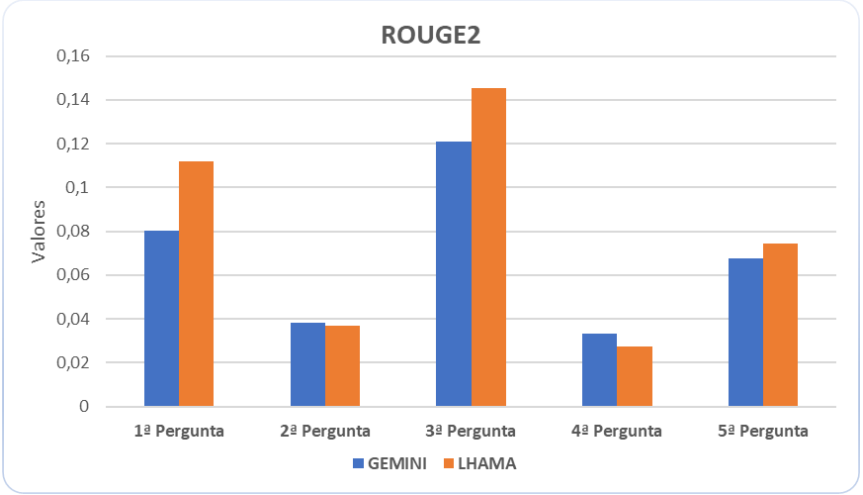
Figura 6 – Análise do Gráfico: Rouge 1



Fonte: Autoria própria

A análise do ROUGE-1, apresentado na Figura 6, revelou desempenho semelhante entre os dois modelos, indicando que ambos conseguem capturar adequadamente os conceitos centrais das respostas esperadas.

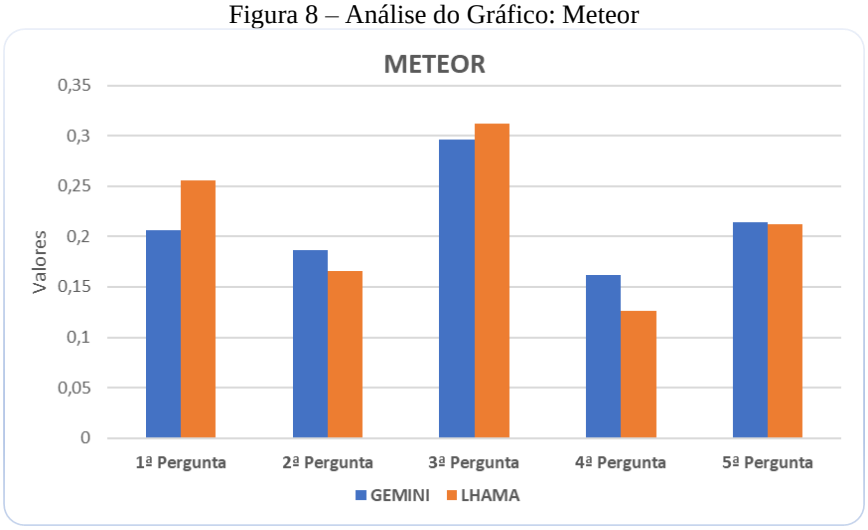
Figura 7 – Análise do Gráfico: Rouge 2



Fonte: Autoria própria

Entretanto, ao observar a métrica ROUGE-2 conforme apresentado na Figura 7, que avalia a correspondência de bigramas, o Meta LLaMA apresentou resultados superiores de forma consistente, evidenciando maior coerência na organização das frases e na relação entre os termos utilizados.

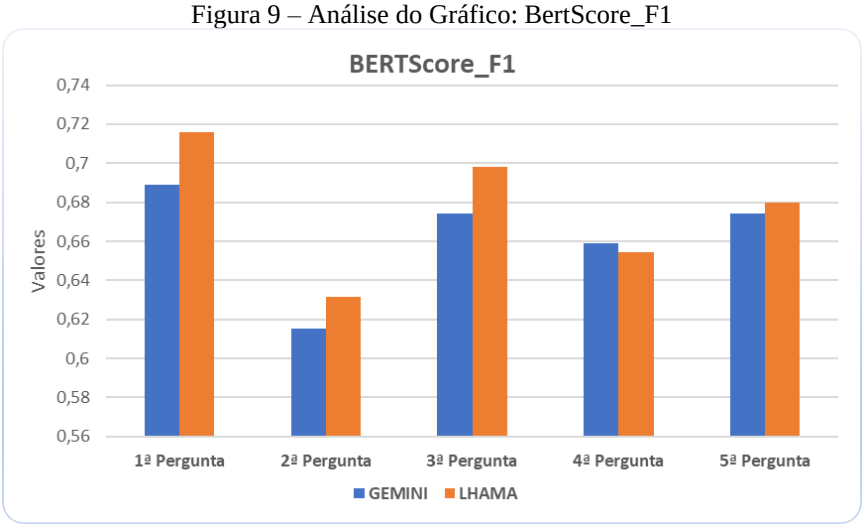
Figura 8 – Análise do Gráfico: Meteor



Fonte: Autoria própria

Em relação à métrica METEOR, conforme apresentado na Figura 8, que considera aspectos semânticos mais amplos, como sinônimos e variações linguísticas, o Gemma apresentou desempenho superior em algumas questões. Esse resultado indica maior riqueza semântica e capacidade explicativa, características desejáveis em contextos educacionais voltados à compreensão conceitual e ao aprofundamento do conteúdo.

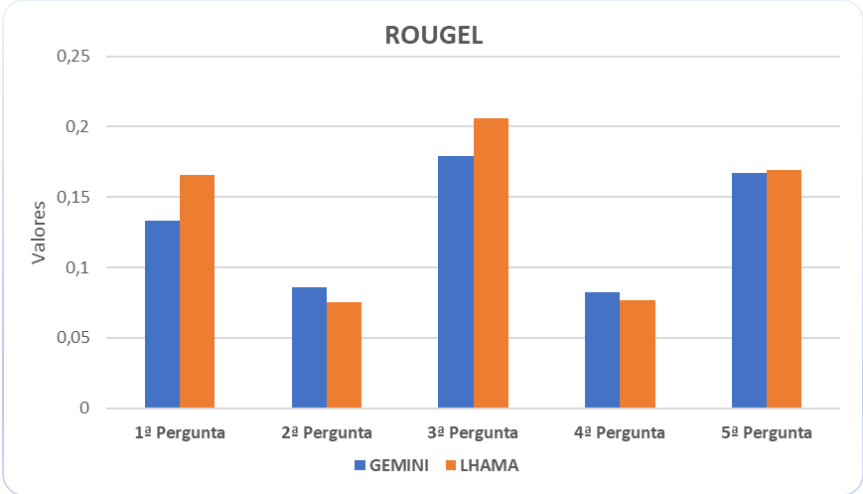
Figura 9 – Análise do Gráfico: BertScore_F1



Fonte: Autoria própria

A métrica BERTScore_F1, conforme a Figura 9, apresentou valores elevados e muito próximos para ambos os modelos, demonstrando que, apesar das diferenças lexicais e estruturais, tanto o Gemma quanto o Meta LLaMA conseguem preservar o significado essencial das respostas, mantendo alta similaridade semântica com o gabarito.

Figura 10 – Análise do Gráfico: RougeL



Fonte: Autoria própria

Por fim, o ROUGE-L conforme a Figura 10, reforça os resultados observados anteriormente, indicando leve vantagem do Meta LLaMA quanto à fluidez textual e à organização lógica das respostas. Ainda assim, o Gemma manteve desempenho consistente, evidenciando equilíbrio entre clareza conceitual e estrutura textual.

De forma geral, os resultados indicam que o Meta LLaMA se destaca em eficiência computacional e consistência estrutural, enquanto o Gemma apresenta maior potencial pedagógico em explicações detalhadas e semanticamente ricas. Esses achados reforçam a ideia de que ambos os modelos possuem características complementares.

4.2 RESULTADOS DA AVALIAÇÃO HUMANA ENTRE AS RESPOSTAS DAS IAG

Figura 11 - Respostas do Formulário: Questão 1

Fonte: Autoria Própria

Este gráfico, conforme Figura 11, apresenta a distribuição da preferência dos estudantes em relação às respostas geradas pelos modelos Gemma e Meta LLaMA para a Questão 1. Observa-se maior aceitação das respostas produzidas pelo Meta LLaMA, indicando que, para essa pergunta

introdutória, o modelo apresentou respostas consideradas mais adequadas pelos avaliadores, possivelmente em função de sua objetividade e clareza técnica.

Figura 12 - Respostas do Formulário: Questão 2

Fonte: Autoria Própria

O gráfico referente à Questão 2, apresentado na Figura 12, evidencia uma predominância ainda mais expressiva da preferência pelo Meta LLaMA. Esse resultado sugere que, em questões que exigem maior precisão conceitual ou explicações mais diretas, o modelo foi percebido como mais eficiente pelos estudantes, reforçando seu desempenho superior em respostas técnicas e estruturadas.

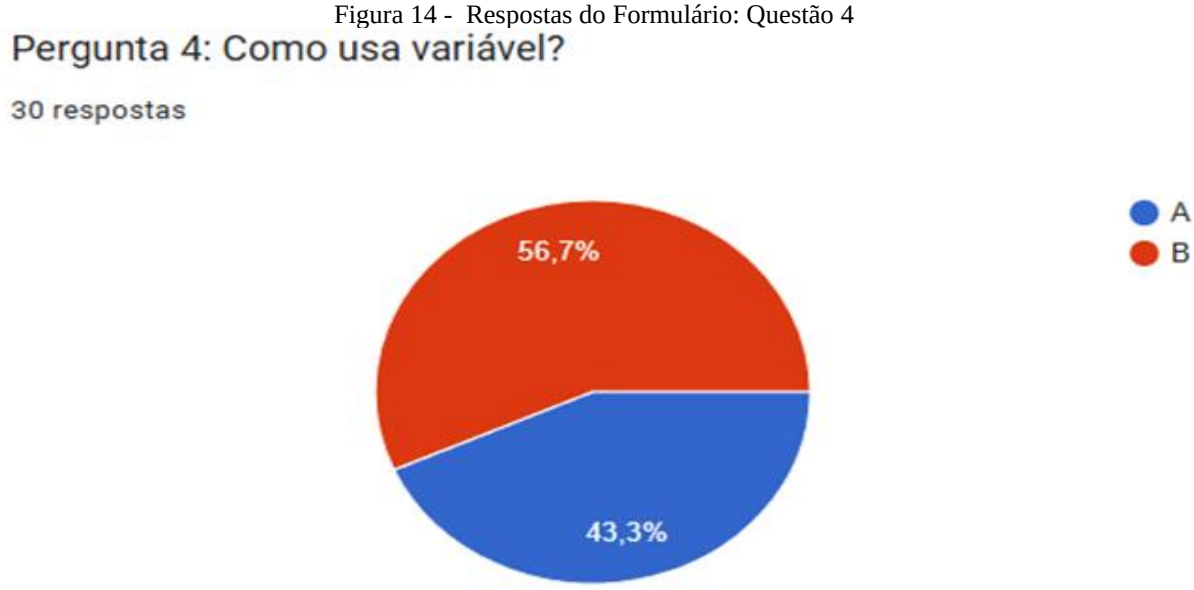
Figura 13 - Respostas do Formulário: Questão 3

Fonte: Autoria Própria

Na Questão 3, conforme a Figura 13, embora o Meta LLaMA continue apresentando maior percentual de preferência, observa-se uma distribuição mais equilibrada quando comparada às questões anteriores. Esse comportamento indica que ambos os modelos foram capazes de atender às

expectativas dos estudantes, ainda que o Meta LLaMA tenha se destacado levemente em termos de clareza e organização lógica da respost

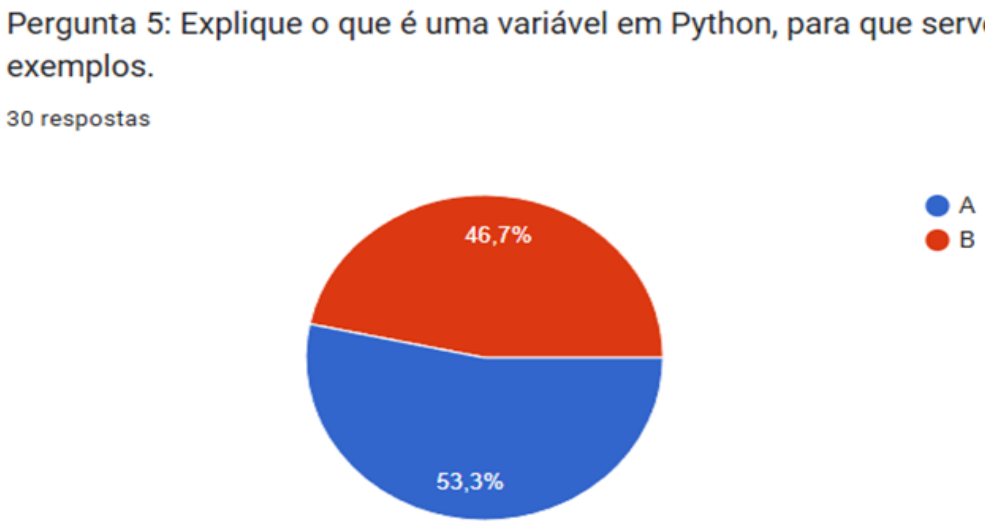
Figura 14 - Respostas do Formulário: Questão 4



Fonte: Autoria Própria

O gráfico da Questão 4 na Figura 14, demonstra novamente maior preferência pelas respostas do Meta LLaMA. Esse resultado reforça a consistência do modelo em fornecer respostas consideradas mais adequadas em atividades relacionadas à lógica aplicada, sugerindo maior alinhamento com o raciocínio esperado pelos estudantes nessa etapa.

Figura 15 - Respostas do Formulário: Questão 5



Fonte: Autoria Própria

Diferentemente das questões anteriores, o gráfico da Questão 5, conforme Figura 15, revela maior preferência pelas respostas geradas pelo modelo Gemma. Esse resultado indica que, em

perguntas que exigem explicações mais detalhadas e maior aprofundamento conceitual, o Gemma apresentou desempenho superior sob a perspectiva pedagógica, sendo considerado mais claro e didático pelos avaliadores.

A análise subjetiva, realizada com 30 estudantes do curso de Sistema de Informação, complementou os dados técnicos. Cada participante avaliou uma amostra de respostas geradas por ambos os modelos e indicou qual considerava mais adequada. Os resultados revelaram predominância da preferência pelo Meta LLaMA nas perguntas 1, 2, 3 e 4 (60%, 90%, 66,7% e 56,7% respectivamente), enquanto o Gemma obteve maior aceitação na pergunta 5 (53,3%). Logo, essa variação demonstra que o Meta LLaMA é mais eficiente em respostas diretas e técnicas, enquanto o Gemma destaca-se em explicações detalhadas e de maior valor pedagógico.

5 CONCLUSÃO

Este estudo teve como objetivo comparar o desempenho dos modelos de Inteligência Artificial Generativa Gemma e Meta LLaMA no contexto do ensino de lógica de programação, considerando métricas objetivas de avaliação textual, desempenho computacional e percepção humana dos estudantes. Em relação à análise integrada dos dados obtidos, as métricas quantitativas demonstraram que o Meta LLaMA apresentou desempenho superior em métricas relacionadas à similaridade lexical e estrutural, como BLEU, ROUGE-2 e ROUGE-L, além de tempos de resposta consistentemente menores. Esses achados indicam maior eficiência computacional e maior precisão na geração de respostas alinhadas ao gabarito, tornando o modelo particularmente adequado para cenários educacionais que demandam rapidez e objetividade, como atividades práticas e ambientes de aprendizagem síncronos.

Tabela 5 - Aspectos Avaliados nos Modelos

Critério	Gemma	Meta LLaMa
Métricas semânticas	Desempenho semelhante ou superior	Desempenho comparável, mas não destacado como superior
Tipo de resposta	Maior riqueza explicativa e flexibilidade semântica	Mais voltado para objetividade e tecnicidade
Uso pedagógico	Melhor para compreensão conceitual e explicações detalhadas	Melhor para perguntas objetivas e técnicas
Preferência dos estudantes	Preferido em questões que exigem explicações detalhadas e aprofundamento conceitual	Majoritariamente preferido em perguntas mais objetivas e técnicas

Fonte:Autoria Própria

Em relação a avaliação geral do modelo, visto na Tabela 5, pode-se afirmar que o Gemma, em relação ao Meta LLaMA, possui um teor mais diversificado por seu detalhamento e generalização em

geração de respostas, sendo mais preferível em contextos de aprofundamento de conteúdo. No entanto o Meta LLaMA, possui maior preferência dos estudantes no formulário, indicando favoritismo dos estudantes por respostas mais diretas e objetivas em contextos mais precisos. Logo, percebe-se que a preferência dos estudante não se concentra em um único modelo, variando de acordo com o tipo de questão, o que evidencia uma complementaridade entre os modelos analisados

Portanto, considerando os critérios definidos na questão de pesquisa: tempo de resposta, consistência textual e adequação pedagógica, o Meta LLaMA apresentou melhor desempenho global, especialmente em eficiência computacional e alinhamento estrutural ao gabarito. No entanto, Gemma demonstrou maior potencial explicativo em questões que exigiam aprofundamento conceitual, dessa forma, conclui-se que não há um modelo universalmente superior para todos os cenários educacionais. Ao contrário, os resultados indicam um caráter complementar entre Gemma e Meta LLaMA, sugerindo que a adoção estratégica de múltiplos modelos pode potencializar o processo de ensino-aprendizagem em lógica de programação. Como trabalhos futuros, recomenda-se a ampliação do número de participantes na avaliação humana, a inclusão de outros modelos de IA generativa e a aplicação da metodologia em diferentes disciplinas da área de computação, a fim de validar e expandir os achados deste estudo.

REFERÊNCIAS

BANERJEE, Satanjeev; LAVIE, Alon. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**.In: WORKSHOP ON INTRINSIC AND EXTRINSIC EVALUATION MEASURES FOR MACHINE TRANSLATION AND/OR SUMMARIZATION, 2005, Ann Arbor.Proceedings […]. Ann Arbor: Association for Computational Linguistics, 2005.

BECKER, B. A., DENNY, P. AND FINNIE-ANSLEY, J. et al. (2023) "**Programming Is Hard - Or at Least It Used to Be: Educational Opportunities and Challenges of AI Code Generation**". In Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1. . ACM. https://doi/10.1145/3545945.3569759.

BRILHANTE, F. J. P. TellMe: uma ponte de comunicação simplificada entre o aluno e a instituição de ensino. 2025. Trabalho de Conclusão de Curso (Tecnologia em Análise e Desenvolvimento de Sistemas) – Instituto Federal da Paraíba, Cajazeiras, 2025.

CASTILHO, Gustavo Uruguay; RODRIGUEZ, Carla Lopes; HERRERA, Victoria Alejandra Salazar. **Um relato de experiência de aplicação de engenharia de prompt no ensino superior em STEM**. In: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (CBIE 2024); WORKSHOP EM ESTRATÉGIAS TRANSFORMADORAS E INOVAÇÃO NA EDUCAÇÃO (WETIE 2024), 2024, Santo André – SP. Anais [...]. Santo André: UFABC, 2024.

DENNY, P., PRATHER, J., BECKER, B. A., FINNIE-ANSLEY, J., HELLAS, A., LEINONEN, J., LUXTON-REILLY, A., REEVES, B. N., SANTOS, E. A., & SARSA, S. (2024). "**Computing Education in the Era of Generative AI**". Communications of the ACM, 67(2), 56–67. https://doi.org/10.1145/3624720.

DUBEY, A., et al. (2024). **The Llama 3 Herd of Models**. arXiv preprint arXiv:2407.21783.

ELNAFFAR, Said et al. **Teaching with AI: A systematic review of chatbots, generative tools, and tutoring systems in programming education**. *arXiv preprint*, 2025. Disponível em: https://arxiv.org/abs/2510.03884. Acesso em: 15 jan. 2026.

GEG BRASIL. **Educação 5.0: Pedagogia de Prompt Vs Engenharia de Prompt em Sala de Aula.** 2024. Disponível em:https://comunidadegegbrasil.blogspot.com/2024/04/educacao-50-pedagogia-de-prompt-vs.html. Acesso em: 03/04/2025.

GOOGLE DEEPMIND. (2024). **Gemma: Open Models Based on Gemini Research and Technology**. arXiv preprint arXiv:2403.08295.

GOMES, A.; MENDES, A. J. Learning to program – difficulties and solutions. In: INTERNATIONAL CONFERENCE ON ENGINEERING EDUCATION, 2007, Coimbra. Proceedings [...]. Coimbra: ICEE, 2007.

LIESENFELD, A., LOPEZ, A., & DINGEMANSE, M. (2023). **Opening up ChatGPT: Tracking openness of instruction-tuned LLMs**. Proceedings of the 5th International Conference on Conversational User Interfaces.

LIN, Chin-Yew. **ROUGE: A package for automatic evaluation of summaries**.In: WORKSHOP ON TEXT SUMMARIZATION BRANCHES OUT, 2004, Barcelona. Proceedings […]. Barcelona: Association for Computational Linguistics, 2004.

LUCKIN, R. et al. **Intelligence Unleashed: An argument for AI in Education**. London: Pearson Education, 2016.

MARQUES, T. M.; SANT'ANA, C. C. A inteligência artificial como recurso para o ensino de matemática: comparativo entre ChatGPT e Gemini. 2024. Disponível em: [PDF].

PAPINENI, Kishore et al. **BLEU: a method for automatic evaluation of machine translation**.In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 40., 2002, Philadelphia. Proceedings […]. Philadelphia: Association for Computational Linguistics, 2002.

SILVA, J. Estudo exploratório e análise comparativa de ferramentas de inteligência artificial generativa para o ensino de computação. 2024. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Universidade Federal de Uberlândia, Uberlândia, 2024.

SILVA, RODRIGO MERO SARMENTO DA; SILVA, JANIEL DOS SANTOS. **Prompt Engineering na Educação em Engenharia: Potencializando a Experiência dos Alunos em Sala de Aula.** In: CONGRESSO BRASILEIRO DE EDUCAÇÃO EM ENGENHARIA, 52., 2024, Maceió. Anais [...]. Maceió: ABENGE, 2024. Disponível em: https://abenge.org.br/sis_artigo_com_capa.php/?cod_trab=4897. Acesso em: 03/04/2025

SILVA, Teresinha Letícia da; VIDOTTO, Kajiana Nuernberg Sartor; TAROUCO, Liane Margarida Rockenbach; SILVA, Patrícia Fernanda da. **Potencialidades do uso de Inteligência Artificial Generativa como apoio ao Ensino de Programação**. In: **CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (CBIE 2024); SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE 2024)**, 2024, Porto Alegre. **Anais [...].** Porto Alegre: UFRGS, 2024.

WANG, Tianyu; ZHOU, Nianjun; CHEN, Zhixiong. **Enhancing Computer Programming Education with LLMs: A Study on Effective Prompt Engineering for Python Code Generation**. *A Preprint*. Mercy University; IBM Research, 2024. Disponível em: https://arxiv.org/abs/2407.05437. Acesso em: 10 out. 2025.

ZHANG, Tianyi et al. **BERTScore: Evaluating text generation with BERT**.In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 2019.

ZAWACKI-RICHTER, O. et al. **Systematic review of research on artificial intelligence applications in higher education: Where are the educators?** *International Journal of Educational Technology in Higher Education*, [S.l.], v. 16, n. 1, p. 1–27, 2019. DOI: https://doi.org/10.1186/s41239-019-0171-0. Acesso em: 17 abr. 2025.