Estudo Experimental sobre o Uso de LLMs na Solução de Questões de Redes de Computadores

Lucas Taylor de Sousa Aires, UFC, lucastayloraires@gmail.com
https://orcid.org/0009-0008-7544-0533

Rubens Fernandes Nunes, UFC, rubensfn@gmail.com
https://orcid.org/0000-0002-7115-8206

Arthur de Castro Callado, UFC, arthur@ufc.br
https://orcid.org/0000-0001-8354-4609

Resumo: Este estudo analisou o desempenho de Modelos de Linguagem de Grande Escala (LLMs) na resolução de questões sobre redes de computadores, considerando o nível de dificuldade das perguntas e o tipo de *prompt* utilizado. Foram avaliadas seis LLMs por meio de 30 questões categorizadas em três níveis, com um *prompt* simples e um melhor elaborado. Os resultados revelaram diferenças de comportamento entre os modelos, sendo o Sabiá o único com desempenho significativamente inferior aos demais. Indicaram também que a complexidade das questões impacta o desempenho mais do que o tipo de *prompt*. Conclui-se que, embora promissoras, os LLMs requerem uso criterioso em contextos educacionais técnicos.

Palavras-chave: Modelos de Linguagem de Grande Escala, Redes de Computadores, Desempenho de Modelos, Tecnologia Educacional.

An Experimental Study on the Use of LLMs for Solving Computer Networks Problems

Abstract: This study analyzed the performance of Large Language Models (LLMs) in solving computer network questions, considering the difficulty level of the questions and the type of prompt used. Six LLMs were evaluated using 30 questions categorized into three levels, with both a simple and a more refined prompt. The results revealed behavioral differences among the models, with Sabiá being the only one to perform significantly worse than the others. They also indicated that question complexity impacts performance more than the type of prompt. It is concluded that, although promising, LLMs require careful use in technical educational contexts.

Keywords: Large Language Models, Computer Networks, Model Performance, Educational Technology.

1. Introdução

O uso da Inteligência Artificial (IA) na educação tem se expandido de forma significativa, impactando práticas pedagógicas e estratégias de aprendizagem (Lima *et al.*, 2024). Alunos e professores têm recorrido a ferramentas baseadas em IA para tarefas como tutoria personalizada, correção automática de atividades, elaboração de conteúdos e resolução de dúvidas, revelando um novo paradigma no processo de ensino-aprendizagem, mediado por sistemas capazes de adaptar-se às necessidades cognitivas dos usuários.

Entre as inovações mais notáveis nesse campo, destacam-se os Modelos de Linguagem de Grande Escala (*Large Language Models* - LLMs), treinados com enormes

volumes de dados textuais extraídos de fontes diversas, como livros, artigos científicos e páginas da internet. Esses modelos são capazes de compreender, interpretar e gerar textos com alto grau de coerência, além de executar tarefas complexas como a produção de códigos e simulações conversacionais (IBM, 2023).

Apesar de seu potencial, o uso dos LLMs em ambientes educacionais levanta questionamentos importantes sobre sua confiabilidade, seus limites e sua efetividade em tarefas que exigem raciocínio técnico e rigor conceitual. Estudos recentes têm alertado para riscos como a superficialidade na aprendizagem, a desinformação e a ausência de critérios claros para o uso dessas ferramentas (Yan *et al.*, 2024). Tais desafios reforçam a necessidade de investigações sistemáticas sobre o uso pedagógico dos LLMs em contextos especializados.

Diante disso, este trabalho tem como objetivo analisar o desempenho de LLMs na resolução de questões sobre redes de computadores. A proposta é avaliar a efetividade e a eficiência desses modelos em tarefas que demandam conhecimento técnico, comuns do ensino e da prática profissional na área. A partir disso, formula-se a seguinte questão de pesquisa: Como diferentes Modelos de Linguagem de Grande Escala se comportam na resolução de questões de redes de computadores, levando em conta o nível de dificuldade das questões e o tipo de *prompt* utilizado?

Para responder a essa questão, foi conduzido um estudo experimental com um total de 30 questões, distribuídas entre os níveis básico, intermediário e avançado, aplicadas a diferentes LLMs mediante dois tipos de *prompt* (genérico e estruturado). A análise dos resultados visa compreender como esses modelos se comportam frente a desafios técnicos e qual o grau de confiabilidade que oferecem em tarefas desse tipo.

O restante deste artigo está organizado da seguinte forma: na Seção 2, apresenta-se a fundamentação teórica; na Seção 3, os trabalhos relacionados; na Seção 4, a metodologia; na Seção 5, os resultados; na seção 6, as considerações finais; e, por fim, os agradecimentos e as referências.

2. Fundamentação Teórica

A Inteligência Artificial (IA) refere-se à capacidade das máquinas de realizar tarefas que tradicionalmente exigiriam inteligência humana, como raciocínio, resolução de problemas, percepção e processamento de linguagem natural (Barbosa e Portes, 2023). A depender de seu grau de autonomia e capacidade cognitiva, a IA pode ser classificada em diferentes categorias, desde sistemas reativos até modelos de superinteligência, capazes de realizar análises complexas e tomar decisões autônomas em diversos contextos (Morandín-Ahuerma, 2022). A abrangência da IA não se limita a uma única função, mas contempla múltiplas tecnologias orientadas ao raciocínio lógico, representação do conhecimento, aprendizado adaptativo e interação linguística, impactando significativamente setores como saúde, finanças, segurança cibernética e, cada vez mais, a educação (Ram e Verma, 2023).

Entre as vertentes mais relevantes da inteligência artificial na atualidade, destacam-se os LLMs, sistemas treinados com volumes massivos de dados textuais extraídos da internet, livros, artigos e outras fontes. Esses modelos são capazes de compreender, gerar e interpretar textos, bem como produzir outros tipos de conteúdo, como imagens, códigos e conversações, de forma contextualizada e coerente (IBM, 2023).

Essas arquiteturas demonstram desempenho notável em tarefas de processamento de linguagem natural, como tradução, resposta a perguntas e síntese de informações, aproximando-se das competências comunicativas humanas (Zangana; Mustafa e Li, 2024). Desde o advento de modelos como o ChatGPT, no final de 2022, os LLMs

vêm revolucionando a interação homem-máquina, tornando-se ferramentas centrais em aplicações que demandam compreensão e geração de linguagem.

No campo educacional, os LLMs vêm sendo utilizados como ferramentas de apoio ao ensino e à aprendizagem, oferecendo respostas personalizadas, promovendo sistemas inteligentes de tutoria e estimulando práticas pedagógicas orientadas pela curiosidade e pela autonomia dos estudantes (Shen, 2024). Esses modelos processam consultas, sintetizam informações e geram conteúdos sob demanda, ampliando as possibilidades de personalização no processo educativo e permitindo suporte imediato em diversas áreas do conhecimento.

Apesar do potencial, seu uso na educação levanta desafios importantes. Como destacam (Yan *et al.*, 2024), surgem questões éticas e práticas relacionadas à transparência dos sistemas, aos riscos à privacidade, aos vieses nos dados e à dependência excessiva de soluções automatizadas. Esses fatores exigem uma análise cuidadosa sobre a forma como essas tecnologias devem ser integradas aos ambientes de aprendizagem de maneira ética, responsável e segura.

No que diz respeito às redes de computadores, elas podem ser definidas como as conexões entre dispositivos que trocam informações entre si, como computadores, celulares, servidores e *tablets*. Essa comunicação ocorre por meio de enlaces físicos, como cabos e sinais de rádio, e de equipamentos de rede, como roteadores e *switches*, permitindo o compartilhamento de dados e recursos (Forouzan e Mosharraf, 2013).

Mapeamentos sistemáticos recentes apontam que tecnologias como jogos, simuladores e ferramentas interativas têm sido amplamente utilizadas no ensino de redes de computadores, com destaque para sua capacidade de tornar o aprendizado mais dinâmico e acessível, especialmente frente às limitações de infraestrutura física (Júnior *et al.*, 2019; Santos *et al.*, 2020). Tais recursos podem ser ainda mais eficazes quando integrados a LLMs, que permitem a personalização do ensino, o esclarecimento automático de dúvidas e a adaptação dos conteúdos conforme o perfil do estudante, ampliando significativamente o potencial pedagógico dessas abordagens.

Diante disso, este estudo busca analisar como os LLMs se comportam na resolução de atividades da área de redes de computadores. O objetivo é avaliar seu desempenho, compreender seus limites e estimar seu nível de confiança no contexto educacional.

3. Trabalhos Relacionados

A avaliação de LLMs em tarefas técnicas tem ganhado destaque na literatura, especialmente em áreas que demandam alta precisão, eficiência e capacidade de raciocínio lógico. Estudos têm explorado o desempenho desses modelos em atividades específicas, desde geração de código até solução de problemas complexos, evidenciando avanços e limitações significativas. A seguir, são apresentados três trabalhos que abordam diferentes perspectivas sobre a aplicação de LLMs em contextos técnicos e educacionais.

(Coignion; Quinton e Rouvoy, 2024) realizaram um estudo abrangente para avaliar a eficiência de código gerado por LLMs em comparação com soluções humanas, utilizando problemas do Leetcode*como base. Foram analisados 18 modelos distintos, levando em conta métricas como tempo de execução e taxa de sucesso. Os autores concluíram que os LLMs conseguem gerar código com desempenho semelhante, e em alguns casos superior, ao produzido por humanos. No entanto, o estudo também alerta para desafios persistentes, como possíveis contaminações nos dados de treinamento e a necessidade de métodos de avaliação mais robustos.

*⟨https://leetcode.com/⟩

(Han *et al.*, 2025) investigaram a capacidade de LLMs em raciocinar logicamente e seguir regras estritas, utilizando quebra-cabeças de word ladder e cenários envolvendo a conformidade com a HIPAA como estudos de caso. Os resultados revelaram que os modelos frequentemente falham em aderir a regras explícitas, priorizando a conclusão da tarefa em detrimento da precisão lógica ou ética. Esse estudo ressalta uma limitação fundamental dos LLMs em contextos que exigem rigor técnico.

(Agarwal *et al.*, 2024) focaram na aplicação de LLMs em tarefas comuns a estudantes de Ciência da Computação, incluindo explicação de código, resolução de exercícios e preparação para entrevistas técnicas. Por meio de uma avaliação conduzida por alunos, os autores identificaram que a efetividade dos modelos varia significativamente dependendo da tarefa, com desempenho superior em atividades estruturadas e limitações em cenários que exigem aprendizado de conceitos novos.

Enquanto os trabalhos discutidos exploram diferentes vertentes do desempenho de LLMs, como eficiência de código, raciocínio lógico e adequação a tarefas educacionais, nossa pesquisa avança nessa linha ao propor uma avaliação sistêmica de modelos em tarefas de redes de computadores. Este estudo busca oferecer um panorama abrangente sobre a viabilidade de LLMs como ferramentas de apoio para profissionais e educadores da área, considerando tanto seus aspectos funcionais quanto sua aplicabilidade prática.

Tabela 1. Comparação entre os trabalhos relacionados e o presente estudo

Trabalho	Foco	Metodologia	Conclusões
- (Coignion; Quinton e Rouvoy, 2024)	Desempenho de código gerado por LLMs	Testes com 18 LLMs em problemas do Leetcode	LLMs geram código com desempenho igual ou superior ao humano
- (Han *et al.*, 2025)	Raciocínio lógico e cumprimento de regras por LLMs	Testes com puzzles e avaliação sobre regras de privacidade usando 5 LLMs	LLMs falham em seguir regras e raciocinar; oferecem riscos em tarefas críticas
- (Agarwal *et al.*, 2024)	Efetividade dos LLMs em tarefas de estudantes de Computação	Avaliação prática de LLMs em atividades acadêmicas e profissionais	Cada LLM possui forças e fraquezas; oferece recomendações de uso educacional
- Este trabalho	Avaliação do desempenho de LLMs em tarefas de redes de computadores	Experimento com 30 questões aplicadas a diferentes LLMs	LLMs mostram desempenho variável, mas potencial no ensino técnico

4. Metodologia

Este estudo caracteriza-se como experimental e qualiquantitativo, com o objetivo de analisar o desempenho de Modelos de Linguagem de Grande Escala (LLMs) na resolução de atividades sobre redes de computadores. Busca-se avaliar a efetividade e a eficiência desses modelos na execução de tarefas que demandam conhecimento técnico, comuns do ensino e da prática profissional na área. A investigação concentra-se no

comportamento dos modelos e na precisão das respostas fornecidas, oferecendo subsídios para compreender as potencialidades e limitações dos LLMs nesse domínio.

4.1. Questões e LLMs

Para compor o instrumento de avaliação, foram selecionadas 30 questões sobre redes de computadores, extraídas da base da plataforma TecConcursos[†]. As questões foram previamente verificadas quanto à correção e adaptadas para fins de padronização e controle experimental, mantendo o conteúdo técnico essencial. A seleção seguiu uma abordagem não probabilística, com o objetivo de contemplar diferentes níveis de complexidade e representatividade temática dentro da área.

As 30 questões foram organizadas de forma equitativa em três níveis de dificuldade: 10 questões de nível básico, 10 de nível intermediário e 10 de nível avançado. Essa estratificação teve como propósito avaliar o comportamento dos modelos diante de diferentes graus de exigência cognitiva e técnica, favorecendo uma análise comparativa mais abrangente e robusta. Cada questão foi submetida aos modelos acompanhada por dois tipos distintos de *prompt*: *Prompt* Genérico, com instruções diretas e objetivas; e *Prompt* Estruturado, com linguagem mais contextualizada, simulando uma solicitação mais próxima do uso real por parte de estudantes ou profissionais. Entende-se como *prompt* uma instrução ou pergunta apresentada ao LLM com o objetivo de orientar a geração de sua resposta. As referidas questões e *prompts* estão disponíveis aqui.

Foram avaliados seis LLMs amplamente utilizados e disponíveis publicamente até o momento da pesquisa: (1) ChatGPT-4o (OpenAI), (2) Gemini 2.5 Flash (Google), (3) DeepSeek-V3 (DeepSeek), (4) Sabiá-3.1 (Maritaca AI), (5) Microsoft Copilot (integrado ao Bing Chat), e (6) Meta AI (baseado no modelo Llama 4). Os modelos foram acessados em suas respectivas interfaces oficiais, garantindo condições controladas de aplicação.

Os experimentos foram conduzidos entre os dias 05 e 08 de julho de 2025. Todas as respostas foram analisadas individualmente quanto à correção com base no gabarito verificado previamente, e os resultados foram registrados para posterior análise estatística e qualitativa. Essa etapa visou assegurar a fidelidade dos dados coletados e a validade das comparações entre os diferentes modelos e tipos de *prompt*.

4.2. Variáveis e Hipóteses

A definição das variáveis é fundamental para orientar a estrutura do experimento e a análise dos resultados. Neste estudo, foram consideradas duas variáveis independentes: o nível de complexidade das questões (básica, intermediária e avançada) e o tipo de *prompt* (genérico e estruturado). Ambas foram manipuladas com o objetivo de investigar seu impacto sobre o desempenho dos LLMs.

As variáveis dependentes referem-se ao comportamento dos modelos e à precisão das respostas. O comportamento foi analisado de forma qualitativa, considerando aspectos discursivos e estruturais das respostas. Já a precisão foi avaliada quantitativamente, por análises estatísticas baseadas na correção das soluções fornecidas.

Com base nesse delineamento experimental, foram formuladas hipóteses para investigar se as variáveis independentes influenciam significativamente a variável dependente de precisão. As hipóteses foram organizadas em pares, nula (H_0) e alternativa (H_1), conforme descrito a seguir:

- **Hipótese** H_01**:** A média de desempenho dos LLMs na resolução das atividades é igual entre todos os modelos avaliados.
- **Hipótese** H_11**:** Pelo menos uma das médias de desempenho dos LLMs difere

[†] ⟨https://www.tecconcursos.com.br/⟩

significativamente das demais.

- **Hipótese** $H_0 2$**:** A média de desempenho dos LLMs não varia significativamente conforme o nível de complexidade das questões.
- **Hipótese** $H_1 2$**:** Pelo menos uma das médias de desempenho dos LLMs varia significativamente conforme o nível de complexidade das questões.
- **Hipótese** $H_0 3$**:** A média de desempenho dos LLMs não difere entre os tipos *prompt*.
- **Hipótese** $H_1 3$**:** A média de desempenho dos LLMs difere significativamente entre os tipos de *prompt*.

Essas hipóteses foram elaboradas com o objetivo de verificar o impacto dos fatores experimentais sobre a precisão das respostas geradas pelos LLMs, permitindo identificar possíveis variações de desempenho em função dos diferentes níveis de complexidade das questões e do tipo de *prompt* empregado. A distinção entre hipóteses nulas e alternativas foi importante para a condução da análise estatística e interpretação dos resultados.

4.3. Métricas de Avaliação

A avaliação do desempenho dos LLMs foi conduzida por meio de uma abordagem combinada, envolvendo análise qualitativa e testes estatísticos. Inicialmente, as respostas dos modelos foram analisadas qualitativamente, considerando critérios como correção técnica e completude. Essa etapa permitiu observar tendências, padrões e fragilidades específicas nos diferentes LLMs.

Para a análise quantitativa, foi utilizada a Análise de Variância (ANOVA), conforme descrito por (Kishore; Sanghadasa e Priya, 2017), um teste estatístico que avalia se há diferenças significativas entre as médias de três ou mais grupos, com base na variância intra e entre grupos. A ANOVA é empregada para determinar se os resultados observados em uma análise são estatisticamente significativos para rejeitar a hipótese nula (Chen *et al.*, 2022). Posteriormente, utilizou-se o Teste de Tukey HSD (*Honestly Significant Difference*) para comparações múltiplas entre os pares de modelos. Segundo (Nanda *et al.*, 2021), o Teste de Tukey é uma abordagem estatística robusta que permite identificar quais diferenças entre os grupos são, de fato, significativas, oferecendo uma visão mais detalhada após a ANOVA. Ambas métricas foram executadas por meio de códigos na linguagem de programação Python.

A significância na ANOVA é verificada por meio do valor de p, que representa a probabilidade de que as diferenças observadas entre os grupos tenham ocorrido ao acaso. Quando esse valor é inferior a um nível de significância previamente definido, geralmente 0,05, conclui-se que há diferença estatística entre os grupos. No teste de Tukey, a significância também é indicada quando o valor ajustado de p (p-*adj*) é inferior ao limite adotado (0,05) ou quando a coluna *reject* retorna *True*, apontando que há diferença significativa entre os grupos comparados.

5. Resultados

5.1. Análise Qualitativa

A análise qualitativa focou em observar o comportamento geral dos LLMs durante a execução da tarefa, considerando fatores como clareza das respostas, estabilidade entre os tipos de *prompt*, capacidade de lidar com imagens e eventuais dificuldades operacionais. A seguir, descrevemos os principais achados por modelo avaliado.

O ChatGPT-4o demonstrou desempenho consistente em ambos os cenários de *prompt*. O modelo respondeu às questões com simplicidade, clareza e organização, sem apresentar qualquer tipo de instabilidade ou ambiguidade. Suas respostas permaneceram

inalteradas entre os dois tipos de *prompt*, evidenciando robustez e previsibilidade em seu comportamento. Essa estabilidade também foi observada no modelo da MetaAI, que mesmo não aceitando anexos de imagem, impossibilitando a resolução de questões visuais, apresentou um padrão de resposta claro e direto, semelhante ao do ChatGPT. Em ambas as abordagens (*Prompt* Genérico e *Prompt* Estruturado), o modelo manteve coerência e organização, indicando boa capacidade de interpretação textual, embora com limitação em tarefas multimodais.

O Gemini 2.5 Flash também obteve bom desempenho, especialmente em termos de clareza e adaptação ao tipo de *prompt*. Com *Prompt* Genérico, limitou-se a indicar diretamente as alternativas corretas, sem justificativas. Já sob *Prompt* Elaborado, passou a fornecer explicações detalhadas para cada resposta, mostrando maior sensibilidade ao contexto e ao estilo de comando recebido. Essa variação de comportamento conforme o tipo de entrada destaca a flexibilidade do modelo, sem prejuízo à qualidade das respostas.

O modelo DeepSeek-V3 apresentou respostas claras, simples e diretas, de modo similar ao ChatGPT e MetaAI. Não ocorreram inconsistências relevantes e o modelo se manteve estável em ambos os tipos de *prompt*. Sua previsibilidade e clareza tornam-no um modelo funcional para aplicações que exigem objetividade e baixo risco de alucinação.

Em contraste, o modelo Sabiá-3.1, da Maritaca AI, apresentou instabilidades no teste com *prompt* genérico. Foram registradas respostas contraditórias, formatos variados e até múltiplos gabaritos diferentes enviados em sequência, gerando dúvidas quanto à resposta final. Para mitigar essa ambiguidade, foi necessário o envio de um segundo *prompt* solicitando explicitamente apenas o gabarito, sem justificativas. Após essa intervenção, o modelo respondeu conforme o esperado. No cenário com *prompt* estruturado, o Sabiá se comportou de maneira estável e clara, o que indica sensibilidade significativa à forma do comando recebido.

Já o modelo Microsoft Copilot demandou mais intervenções ao longo do teste. Devido a restrições da ferramenta, o questionário foi dividido em três partes, respeitando o limite de 10.240 caracteres por *prompt*. Além disso, a restrição de apenas uma imagem por *prompt* exigiu uma organização cuidadosa das questões que dependiam de recursos visuais para sua resolução. Com o *prompt* genérico, o Copilot interpretou erroneamente a atividade como uma solicitação para montar um *quiz* interativo (Figura 1).

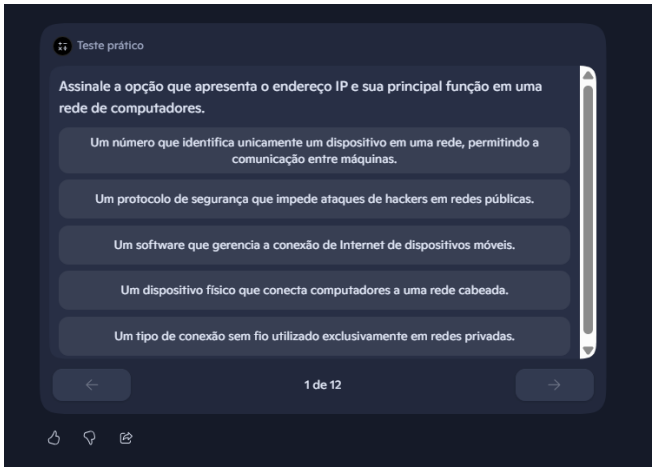

Figura 1. Quiz interativo desenvolvido pelo Microsoft Copilot

Foi necessário um segundo *prompt* esclarecendo a real intenção da tarefa. Após o ajuste, o modelo respondeu de forma adequada, embora não tenha conseguido interpretar corretamente uma das imagens anexadas (questão 4), respondendo com uma

mensagem de ausência de conteúdo. Curiosamente, conseguiu compreender e responder corretamente à questão 29, que também dependia de uma imagem. Já com o *prompt* estruturado, o modelo interpretou corretamente a tarefa e apresentou respostas mais completas, além de conseguir identificar a presença da imagem da questão 4.

Em síntese, os resultados qualitativos revelam diferenças relevantes no comportamento dos LLMs testados. Enquanto alguns modelos demonstraram alta estabilidade e consistência entre tipos de *prompt*, outros apresentaram variações significativas, necessidade de intervenções adicionais e sensibilidade ao formato de entrada. Esses achados ajudam a delimitar o uso prático de cada modelo e orientam sua aplicação em contextos educacionais que demandam precisão e compreensão multimodal.

5.2. Análise Estatística

A Tabela 2 apresenta o número de respostas corretas por nível de questão (básica, intermediária e avançada), considerando o uso de *prompt* genérico. Já a Tabela 3 mostra os resultados com o uso de *prompt* estruturado. Observa-se que nenhum dos LLMs alcançou 100% de acertos, mesmo nas questões básicas, sendo que a taxa máxima foi de 90%. Isso evidencia que, embora apresentem desempenho satisfatório, ainda há limitações na precisão dessas ferramentas.

Tabela 2. Número de questões certas por nível de dificuldade - Prompt Genérico

LLM	Básicas	Intermediárias	Avançadas	Total
ChatGPT	10	9	8	27
Gemini	10	9	7	26
DeepSeek	10	10	5	25
Sabiá	4	5	3	12
Microsoft Copilot	9	9	8	26
Meta AI	9	10	8	27

Tabela 3. Número de questões certas por nível de dificuldade - Prompt Estruturado

LLM	Básicas	Intermediárias	Avançadas	Total
ChatGPT	10	9	8	27
Gemini	10	9	6	25
DeepSeek	10	10	6	26
Sabiá	8	4	5	17
Microsoft Copilot	10	9	7	26
Meta AI	9	10	8	27

Com base no total de acertos por LLM, foi realizada a ANOVA, cujo valor de p foi de 0,000259, indicando que há, pelo menos, um LLM com desempenho significativamente diferente das demais. Com isso, a hipótese $H_1 1$ foi validada e a hipótese $H_0 1$, rejeitada. O Teste de Tukey (Tabela 4) mostrou que todas as comparações envolvendo o LLM Sabiá apresentaram diferenças estatisticamente significativas, indicando seu desempenho inferior. Já nas comparações entre os demais LLMs, não houve rejeição da hipótese nula, sugerindo desempenhos semelhantes.

Na sequência, foi analisado o impacto da complexidade das questões no desempenho dos LLMs. A ANOVA retornou $p = 0{,}0038$, sugerindo efeito estatisticamente significativo da complexidade sobre os resultados, o que valida a hipótese $H_1 2$ e invalida a $H_0 2$. O Teste de Tukey apontou diferenças significativas entre o desempenho nas questões

de nível avançado em relação às de nível básico ($p = 0{,}0045$) e intermediário ($p = 0{,}0253$). A diferença entre os níveis básico e intermediário não foi significativa ($p = 0{,}7724$).

Por fim, foi avaliado o impacto do tipo de *prompt* nas respostas dos LLMs. Embora a análise qualitativa tenha indicado variações no comportamento, o teste ANOVA resultou em $p = 0{,}6902$, sem diferença estatística significativa nas respostas finais. Isso sugere que o tipo de *prompt* influencia o processo, mas tem pouco efeito no conteúdo gerado. Assim, a hipótese $H_0 3$ foi aceita e a $H_1 3$, rejeitada.

Tabela 4. Resultados do Teste de Tukey (comparações múltiplas)

LLM1	LLM2	meandiff	p-adj	reject
cgpt	copi	-0.3333	0.9989	False
cgpt	dese	-0.5	0.9923	False
cgpt	gemi	-0.5	0.9923	False
cgpt	meta	0.0	1.0	False
cgpt	sabi	-4.1667	0.0006	True
copi	dese	-0.1667	1.0	False
copi	gemi	-0.1667	1.0	False
copi	meta	0.3333	0.9989	False
copi	sabi	-3.8333	0.0017	True
dese	gemi	0.0	1.0	False
dese	meta	0.5	0.9923	False
dese	sabi	-3.6667	0.0029	True
gemi	meta	0.5	0.9923	False
gemi	sabi	-3.6667	0.0029	True
meta	sabi	-4.1667	0.0006	True

6. Considerações Finais

Este trabalho investigou a efetividade de diferentes LLMs na resolução de questões técnicas sobre redes de computadores, com foco na influência dos níveis de complexidade dessas questões e dos tipos de *prompt*. A abordagem experimental e qualiquantitativa permitiu observar padrões de desempenho e estabilidade entre os modelos avaliados. Os resultados mostraram que ChatGPT, MetaAI e DeepSeek se destacaram em precisão e consistência, enquanto a Sabiá apresentou instabilidades significativas. A análise estatística confirmou que a complexidade das questões impacta diretamente o desempenho dos LLMs, ao passo que o tipo de *prompt* não teve efeito significativo. Conclui-se que os LLMs têm potencial para apoiar a aprendizagem em áreas técnicas, como a de redes de computadores, mas sua adoção requer cuidados quanto à escolha do modelo e ao contexto de uso. Futuras pesquisas podem ampliar a base de questões e robustez do experimento, explorar a aplicação dessas ferramentas em outras áreas da computação, bem como o desenvolvimento de estratégias pedagógicas que integrem IA com práticas avaliativas confiáveis.

7. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

Referências

Agarwal, V.; Garg, M. K.; Dharmavaram, S.; Kumar, D. “which llm should i use?”:

Evaluating llms for tasks performed by undergraduate computer science students. **arXiv preprint arXiv:2402.01687**, 2024.

Barbosa, L. M.; Portes, L. A. F. A inteligência artificial. **Revista Tecnologia Educacional [on line], Rio de Janeiro**, n. 236, p. 16–27, 2023.

Chen, W.-H. *et al.* A comprehensive review of thermoelectric generation optimization by statistical approach: Taguchi method, analysis of variance (anova), and response surface methodology (rsm). **Renewable and Sustainable Energy Reviews**, Elsevier, v. 169, p. 112917, 2022.

Coignion, T.; Quinton, C.; Rouvoy, R. A performance study of llm-generated code on leetcode. In: **Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering**. [S.l.: s.n.], 2024. p. 79–89.

Forouzan, B. A.; Mosharraf, F. **Redes de computadores: uma abordagem top-down**. [S.l.]: AMGH Editora, 2013.

Han, Z.; Battaglia, F.; Mansuria, K.; Heyman, Y.; Terlecky, S. R. Beyond text generation: Assessing large language models' ability to reason logically and follow strict rules. **AI**, Multidisciplinary Digital Publishing Institute, v. 6, n. 1, p. 12, 2025.

IBM. **O que é LLM (large language models)?** 2023. Acesso em: 20 jun. 2025. Disponível em: ⟨https://www.ibm.com/br-pt/think/topics/large-language-models⟩.

Júnior, G. da S.; Medina, R. D.; Lima, P. R. B. D.; Mazzuco, A. E. da R. Mapeamento sistemático sobre o desenvolvimento e a utilização de jogos e simuladores no ensino de redes de computadores. **Revista Novas Tecnologias na Educação**, v. 17, n. 3, p. 152–162, 2019.

Kishore, R. A.; Sanghadasa, M.; Priya, S. Optimization of segmented thermoelectric generator using taguchi and anova techniques. **Scientific reports**, Nature Publishing Group UK London, v. 7, n. 1, p. 16746, 2017.

Lima, L. *et al.* Artificial intelligence and its use in the educational process. **Navigating through the knowledge of education**, Seven Editora São José dos Pinhais, v. 1, 2024.

Morandín-Ahuerma, F. What is artificial intelligence? **International Journal of Research Publication and Reviews**, v. 3, n. 12, p. 1947–1951, 2022. Disponível em: ⟨https://doi.org/10.55248/gengpi.2022.31261⟩.

Nanda, A.; Mohapatra, B. B.; Mahapatra, A. P. K.; Mahapatra, A. P. K.; Mahapatra, A. P. K. Multiple comparison test by tukey's honestly significant difference (hsd): Do the confident level control type i error. **International Journal of Statistics and Applied Mathematics**, AkiNik Publications, v. 6, n. 1, p. 59–65, 2021.

Ram, B.; Verma, P. Artificial intelligence ai-based chatbot study of chatgpt, google ai bard and baidu ai. **World Journal of Advanced Engineering Technology and Sciences**, v. 8, n. 1, p. 258–261, 2023. Disponível em: ⟨https://doi.org/10.30574/wjaets.2023.8.1.0045⟩.

Santos, A. E. D. *et al.* Ensino de redes de computadores mediado por tecnologias educacionais: um mapeamento sistemático da literatura. **Revista Novas Tecnologias na Educação**, v. 18, n. 1, 2020.

Shen, S. Application of large language models in the field of education. **Theoretical and Natural Science**, v. 34, n. 1, p. 140–147, 2024. Disponível em: ⟨https://doi.org/10.54254/2753-8818/34/20241163⟩.

Yan, L. *et al.* Practical and ethical challenges of large language models in education: A systematic scoping review. **British Journal of Educational Technology**, Wiley Online Library, v. 55, n. 1, p. 90–112, 2024.

Zangana, H. M.; Mustafa, F. M.; Li, S. Large language models in cybersecurity. In: **Advances in Information Security, Privacy, and Ethics Book Series**. [s.n.], 2024. p. 277–300. Disponível em: ⟨https://doi.org/10.4018/979-8-3373-1102-9.ch009⟩.