AVALIAÇÃO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL NA RESOLUÇÃO DE QUESTÕES DE MÚLTIPLA ESCOLHA EM UM EXAME PADRONIZADO DE ODONTOLOGIA (ENADE)

FELIPE DE OLIVEIRA CREMA[1]; TIAGO SCHLINDVEIN DE ARAUJO[2]; EDUARDO TROTA CHAVES[2]; VITOR HENRIQUE DIGMAYER ROMERO[2]; JAQUELINE BARBIERI MACHADO[2] WELLINGTON LUIZ DE OLIVEIRA DA ROSA[3]

[1]Universidade Federal de Pelotas –felipedeoliveiracrema@gmail.com
[2]Universidade Federal de Pelotas–tiagoschlar@gmail.com
[2]Universidade Federal de Pelotas–eduardo.trota@yahoo.com
[2]Universidade Federal de Pelotas–vitordigmayer@gmail.com
[2]Universidade Federal de Pelotas–jaquelineenalta@gmail.com
[3]Universidade Federal de Pelotas – darosawlo@gmail.com

1. INTRODUÇÃO

Os modelos de linguagem de grande escala (LLMs) têm ganhado espaço na educação em ciências da saúde. Baseados em redes neurais profundas, esses sistemas interpretam textos, geram conteúdo e realizam raciocínio lógico com elevado grau de complexidade (BROWN et al., 2020). Entre suas aplicações, destaca-se a resolução de questões de múltipla escolha (MCQs), formato amplamente utilizado em avaliações acadêmicas, exames de proficiência e certificações profissionais (GILSON et al., 2023). Embora apresentem bons resultados em exames padronizados, estudos indicam redução de desempenho em questões que exigem julgamento clínico, interpretação contextual ou raciocínio mais elaborado (LIÉVIN et al., 2023; LIU et al., 2024).

No Brasil, o Exame Nacional de Desempenho dos Estudantes (ENADE) é uma ferramenta utilizada para avaliar a qualidade dos cursos de graduação, incluindo Odontologia. Suas questões abrangem conhecimentos gerais e específicos e exploram a capacidade de raciocínio clínico, análise crítica e tomada de decisão dos estudantes, constituindo um instrumento adequado para examinar a atuação dos LLMs em diferentes níveis de complexidade cognitiva.

O objetivo deste estudo foi comparar o desempenho de cinco modelos de linguagem (Chat GPT-4o, Gemini 1.5 Flash, LLaMA 3, Perplexity AI e DeepSeek V3) na resolução de questões de múltipla escolha do ENADE Odontologia, considerando a acurácia, a consistência das respostas e a influência do tipo de conteúdo.

2. METODOLOGIA

O estudo foi registrado e disponibilizado publicamente no Open Science Framework (OSF), e pode ser acessado pelo link:

[https://osf.io/pa5bu/?view_only=c780ff6e73f1482eba31d1b5addd83c4].

O estudo analisou dados secundários de domínio público, não envolvendo participantes humanos, o que dispensou aprovação por comitê de ética. Foram

selecionadas 181 questões de múltipla escolha do ENADE Odontologia, aplicadas entre 2004 e 2023, sendo 33 de conhecimentos gerais e 148 de conhecimentos específicos. Apenas questões textuais foram incluídas, excluindo aquelas com imagens ou respostas discursivas.

Cada questão foi submetida manualmente a cinco modelos de linguagem: ChatGPT-4o, Gemini 1.5 Flash, LLaMA 3, Perplexity AI e DeepSeek V3. As interações ocorreram entre janeiro e março de 2025, utilizando versões públicas e gratuitas dos sistemas, com histórico de conversas limpo entre as rodadas. Cada item foi inserido cinco vezes em cada modelo, e a resposta final foi definida pela estratégia *best-of-five*. Nos casos sem maioria, realizaram-se três novas rodadas, adotando-se a estratégia *best-of-eight*.

Utilizou-se um único *prompt* padronizado, em inglês, com instruções claras sobre a escolha de uma alternativa (A–E) e apresentação de justificativa. As respostas foram comparadas aos gabaritos oficiais do INEP. A acurácia global e por componente foi avaliada pelo teste Q de Cochran, com comparações pareadas pelo teste de McNemar. A consistência interna foi medida pelo coeficiente de Fleiss (κ) e a variabilidade de respostas foi determinada pelo número médio de alternativas distintas entre rodadas corretas e incorretas.

3. RESULTADOS E DISCUSSÃO

Dentre os modelos avaliados, o DeepSeek apresentou a maior acurácia global (77,3%) e maior consistência interna (κ = 0,912). O Perplexity foi o segundo mais preciso (70,2%), mas com maior variabilidade nas respostas incorretas (83,3% de variação). ChatGPT, Gemini e LLaMA obtiveram acurácias semelhantes (entre 63,5% e 66,3%). Todos os modelos tiveram alto desempenho nas questões gerais (81,8% a 93,9%), mas maior dispersão nas questões específicas, com destaque novamente para o DeepSeek (75,0%) (Figura 1).

A análise da consistência interna mostrou que DeepSeek (κ = 0,912) e Gemini (κ = 0,890) apresentaram maior estabilidade nas respostas corretas, enquanto Perplexity exibiu elevada variabilidade em respostas incorretas (83,3% de mudança de alternativas). Esse padrão indica que modelos com alta acurácia e baixa variação entre tentativas tendem a oferecer maior confiabilidade para uso educacional, enquanto instabilidade em respostas incorretas pode limitar sua aplicabilidade (Figura 1).

As análises das respostas ao ENADE 2019 , confirmaram que os modelos têm melhor desempenho em itens fáceis. Apenas DeepSeek e Gemini acertaram a questão mais difícil do exame (Q25, acerto de 14% entre estudantes),evidenciando sensibilidade à dificuldade. Modelos como Perplexity e ChatGPT mostraram grande instabilidade em respostas incorretas, alterando-as em mais de 60% das tentativas (Tabela 1).

Os dados sugerem que a escolha do modelo influencia significativamente a qualidade das respostas, sobretudo em perguntas clínicas. DeepSeek e Gemini demonstraram melhor equilíbrio entre precisão e consistência. Por outro lado, o

uso educacional de modelos menos consistentes pode ser válido como ferramenta de revisão, desde que acompanhado por validação profissional.

No componente de conhecimentos gerais, todos os modelos obtiveram altos índices de acerto, variando de 81,8% a 93,9%, sem diferenças significativas. Já no componente específico, o desempenho foi mais heterogêneo: DeepSeek manteve a liderança com 75,0%, seguido por Perplexity (66,9%), enquanto Gemini, ChatGPT e LLaMA apresentaram resultados próximos de 60%. Esses valores comparativos, com indicação de diferenças estatísticas, encontram-se detalhados na Tabela 1.

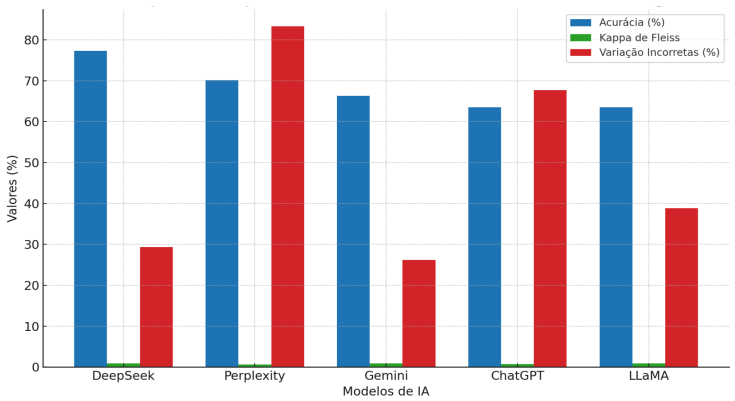

Figura 1: A figura ilustra visualmente os resultados dos modelos em termos de acurácia, consistência (kappa de Fleiss) e variabilidade nas respostas incorretas. Fonte**:** Próprio autor.

Tabela 1. Comparação da acurácia e desempenho de cada LLM por componente do ENADE. Fonte: Próprio autor.

Modelo	Total de Respostas	Respostas Corretas	Respostas Incorretas	Acurácia % (IC 95%)	Alternativa Mais Frequente (%)
QUESTÕES DE CONHECIMENTOS GERAIS (n = 33)					
ChatGPT	33	27	6	81.8% (65.6 – 91.4)^{A}	A (39.4%)
Gemini	33	31	2	93.9% (80.4 – 98.3)^{A}	A (36.4%)
LLaMA	33	28	5	84.8% (69.1 – 93.3)^{A}	A (36.4%)
Perplexity	33	28	5	84.8% (69.1 – 93.3)^{A}	A (33.3%)
DeepSeek	33	29	4	87.9% (72.7 – 95.2)^{A}	A (33.3%)
QUESTÕES DE CONHECIMENTOS ESPECÍFICOS (n = 148)					
ChatGPT	148	88	60	59.5% (51.4 – 67.0)^{AB}	D (25.0%)

Gemini	148	89	59	60.1% (52.1 – 67.7)[AB]	C (24.3%)
LLaMA	148	87	61	58.8% (50.7 – 66.4)[A]	E (22.3%)
Perplexity	148	99	49	66.9% (59.0 – 74.0)[B]	D (23.0%)
DeepSeek	148	111	37	75.0% (67.5 – 81.3)[C]	C (23.0%)

4. CONCLUSÕES

Todos os modelos de linguagem analisados foram capazes de resolver questões do ENADE Odontologia com desempenho moderado a alto. DeepSeek e Perplexity se destacaram tanto na acurácia quanto na estabilidade das respostas, especialmente em itens clínicos e de maior complexidade. Os resultados indicam potencial para uso educacional, mas reforçam a necessidade de validação profissional antes de qualquer aplicação em contextos que exijam tomada de decisão baseada em evidências.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BROWN, T. B. et al. **Language models are few-shot learners.** arXiv preprint arXiv:2005.14165, 2020. Disponível em: https://arxiv.org/abs/2005.14165. Acesso em: 03 de ago. 2025.

GILSON, A. et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? **JMIR Med Educ.**, 2023.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Exame Nacional de Desempenho dos Estudantes (ENADE).** Brasília, 2023. Disponível em: https://www.gov.br/inep. Acesso em: 03 de ago. 2025

LIÉVIN, V. et al. **Can large language models reason about medical questions?** arXiv preprint arXiv:2207.08143, 2023. Disponível em: https://arxiv.org/abs/2207.08143. Acesso em: 03 de ago. 2025.

LIU, M. et al. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. **Journal of Medical Internet Research**, 2024.