v.6, n.1, 2026

REVISTA CIENTÍFICA RECISATEC
ISSN 2763-8405

DESEMPENHO DE MODELOS DE INTELIGÊNCIA ARTIFICIAL GENERATIVA EM QUESTÕES DE CONCURSO PÚBLICO DE ODONTOLOGIA: UM ESTUDO COMPARATIVO DA TAXA DE ACERTO EM SAÚDE BUCAL COLETIVA

PERFORMANCE OF GENERATIVE ARTIFICIAL INTELLIGENCE MODELS IN PUBLIC DENTISTRY EXAM QUESTIONS: A COMPARATIVE STUDY OF THE ACCURACY RATE IN COLLECTIVE ORAL HEALTH

RENDIMIENTO DE LOS MODELOS DE INTELIGENCIA ARTIFICIAL GENERATIVA EN LAS PREGUNTAS DE LOS EXÁMENES PÚBLICOS DE ODONTOLOGÍA: UN ESTUDIO COMPARATIVO DE LA TASA DE PRECISIÓN EN LA SALUD BUCAL COLECTIVA

Tânia Adas Saliba[1], Eder Akydawan de Paiva Gomes Fernandes[2], Cristhiane Martins Schmidt[3]

e61414
https://doi.org/10.70187/recisatec.v6i1.414

PUBLICADO: 05/2026

RESUMO

O avanço da inteligência artificial generativa tem despertado interesse em sua aplicação na área da saúde, incluindo a odontologia. No entanto, ainda são escassos os estudos que avaliam o desempenho dessas ferramentas em contextos específicos da odontologia coletiva, como a resolução de questões de concurso público. Diante disso, o presente estudo teve como objetivo avaliar e comparar as taxas de acerto de três modelos de inteligência artificial generativa: *ChatGPT, Gemini* e *DeepSeek*, em suas versões gratuitas, na resolução de 100 questões de concurso público na área de saúde bucal coletiva. As questões foram extraídas de bancos públicos de provas para cirurgiões-dentistas realizadas entre 2016 e 2026, abrangendo temas como epidemiologia oral, políticas públicas do SUS, vigilância em saúde, determinantes sociais e gestão de serviços. Cada questão foi aplicada individualmente aos três modelos utilizando prompt padronizado, sem histórico de conversas prévias, e a taxa de acerto foi calculada considerando cada questão correta como equivalente a 1 ponto percentual. Os resultados demonstraram que o ChatGPT obteve o melhor desempenho (75 acertos) seguido pelo Gemini (47 acertos) e pelo DeepSeek (23 acertos). As diferenças entre todos os pares foram estatisticamente significativas ($p < 0,001$), com o ChatGPT superando o Gemini em 28 pontos percentuais e o DeepSeek em 52 pontos percentuais. Conclui-se que, entre os modelos gratuitos testados, apenas o ChatGPT alcançaria a pontuação mínima para aprovação na maioria dos concursos públicos para cirurgiões-dentistas em saúde bucal coletiva, enquanto Gemini e DeepSeek, nas versões avaliadas, não se mostraram ferramentas confiáveis para esse fim.

PALAVRAS-CHAVE: Inteligência Artificial Generativa. Saúde Bucal. Saúde Pública.

ABSTRACT

The Advancement of generative artificial intelligence has sparked interest in its application in the health field, including dentistry. However, studies evaluating the performance of these tools in specific contexts of public dentistry, such as solving public service examination questions, are still scarce. Therefore, this study aimed to evaluate and compare the accuracy rates of three generative artificial intelligence models: ChatGPT, Gemini, and DeepSeek, in their free versions, in solving 100 public service examination questions in the area of public oral health. The questions were extracted from public exam banks for dentists conducted between 2016 and 2026, covering topics such as oral epidemiology, SUS (Brazilian Unified Health System) public policies, health surveillance, social determinants, and service management. Each question was applied individually to the three models

[1] Doutora em Odontologia Legal e Deontologia. Universidade Estadual Paulista "Júlio de Mesquita Filho". (UNESP). Araçatuba, São Paulo, Brasil.
[2] Doutorando em Saúde Coletiva em Odontologia. Universidade Estadual Paulista "Júlio de Mesquita Filho" - UNESP. Araçatuba, São Paulo, Brasil.
[3] Doutora em Biologia Buco-Dental. Universidade Estadual Paulista "Júlio de Mesquita Filho". (UNESP). Araçatuba, São Paulo, Brasil.

REVISTA CIENTÍFICA RECISATEC
ISSN 2763-8405

using a standardized prompt, without a history of previous conversations, and the accuracy rate was calculated considering each correct answer as equivalent to 1 percentage point. The results demonstrated that ChatGPT achieved the best performance (75 correct answers), followed by Gemini (47 correct answers) and DeepSeek (23 correct answers). The differences between all pairs were statistically significant ($p < 0.001$), with ChatGPT outperforming Gemini by 28 percentage points and DeepSeek by 52 percentage points. It is concluded that, among the free models tested, only ChatGPT would achieve the minimum score for approval in most public examinations for dental surgeons in collective oral health, while Gemini and DeepSeek, in the versions evaluated, did not prove to be reliable tools for this purpose.

***KEYWORDS**: Generative Artificial Intelligence. Oral Health. Public Health.*

RESUMEN
El avance de la inteligencia artificial generativa ha despertado interés en su aplicación en el campo de la salud, incluyendo la odontología. Sin embargo, aún son escasos los estudios que evalúan el rendimiento de estas herramientas en contextos específicos de la odontología pública, como la resolución de preguntas de exámenes de servicio público. Por lo tanto, este estudio tuvo como objetivo evaluar y comparar las tasas de precisión de tres modelos de inteligencia artificial generativa: ChatGPT, Gemini y DeepSeek, en sus versiones gratuitas, para resolver 100 preguntas de exámenes de servicio público en el área de salud bucal pública. Las preguntas se extrajeron de bancos de exámenes públicos para odontólogos realizados entre 2016 y 2026, que abarcan temas como epidemiología oral, políticas públicas del SUS (Sistema Único de Salud de Brasil), vigilancia de la salud, determinantes sociales y gestión de servicios. Cada pregunta se aplicó individualmente a los tres modelos utilizando una consigna estandarizada, sin historial de conversaciones previas, y la tasa de precisión se calculó considerando cada respuesta correcta como equivalente a 1 punto porcentual. Los resultados demostraron que ChatGPT obtuvo el mejor rendimiento (75 respuestas correctas), seguido de Gemini (47 respuestas correctas) y DeepSeek (23 respuestas correctas). Las diferencias entre todos los pares fueron estadísticamente significativas ($p < 0,001$), con ChatGPT superando a Gemini por 28 puntos porcentuales y a DeepSeek por 52 puntos porcentuales. Se concluye que, entre los modelos gratuitos evaluados, solo ChatGPT alcanzaría la puntuación mínima para su aprobación en la mayoría de los exámenes públicos para cirujanos dentales en salud bucal colectiva, mientras que Gemini y DeepSeek, en las versiones evaluadas, no demostraron ser herramientas fiables para este fin.

***PALABRAS CLAVE**: Inteligencia Artificial Generativa. Salud Bucal. Salud Pública.*

INTRODUÇÃO

A inteligência artificial generativa tem avançado rapidamente no campo da saúde, incluindo aplicações em diagnóstico por imagem, planejamento terapêutico e educação continuada. Modelos de linguagem de grande escala, como os disponíveis comercial e academicamente, passaram a ser testados em contextos que vão desde o atendimento clínico simulado até a formulação de políticas públicas (Bhuyan *et al.,* 2025; Wang *et al.,* 2025). Na odontologia, contudo, a maior parte dos estudos ainda se concentra em aspectos técnicos e privados da prática profissional, deixando em segundo plano o potencial dessas ferramentas para apoiar a atuação no âmbito coletivo, onde questões como epidemiologia, gestão de serviços, vulnerabilidade social e equidade são centrais (Bamashmous, 2025).

A saúde bucal coletiva exige do profissional não apenas conhecimento biomédico, mas também domínio de políticas públicas, sistemas de informação, determinantes sociais, prevenção em

nível populacional e legislação do Sistema Único de Saúde (SUS). Esses temas são recorrentes em concursos públicos para odontólogos que atuam na atenção básica, vigilância em saúde, gestão e programas estratégicos como Saúde da Família (Narvai, 2006).

Embora estudos recentes tenham comparado o desempenho de diferentes Inteligências Artificiais (*Ias)* em provas médicas e jurídicas, ainda há pouca investigação focada na odontologia coletiva, especialmente em formato de concurso público (Bicalho; Oliveira; Guida, 2025). Sabe-se que modelos variam significativamente quanto à precisão, à consistência das respostas e à sensibilidade a nuances contextuais, como regionalização das políticas ou perfis epidemiológicos locais (Araújo *et al.,* 2025). No entanto, não está claro como as principais inteligências artificiais disponíveis se comportariam diante de itens objetivos extraídos ou inspirados em editais reais dessa área específica (Savegnago *et al.,* 2024).

Diante disso, o presente estudo teve como objetivo avaliar e comparar as taxas de acerto de diferentes modelos de inteligência artificial generativa na resolução de questões de concurso público voltadas para a saúde bucal coletiva, analisando seu desempenho relativo e possíveis padrões de erro.

MÉTODO

Trata-se de um estudo experimental, quantitativo, comparativo e transversal, conduzido para avaliar o desempenho de três modelos de inteligência artificial generativa na resolução de questões objetivas de concurso público na área de saúde bucal coletiva. O delineamento consistiu na aplicação padronizada de um mesmo conjunto de questões a cada sistema, com registro isolado das respostas e posterior cálculo da taxa de acerto por IA.

As questões foram extraídas de bancos públicos disponíveis gratuitamente na internet, especificamente de provas de concursos públicos para cirurgiões-dentistas realizadas entre 2016 e 2026. Foram selecionados exclusivamente itens referentes ao núcleo temático de saúde bucal coletiva, incluindo epidemiologia oral, políticas públicas, vigilância em saúde, determinantes sociais, promoção de saúde em nível populacional e gestão de serviços odontológicos no SUS.

Para possibilitar uma interpretação direta e intuitiva do desempenho de cada inteligência artificial, optou-se por calcular a taxa de acerto de forma que cada questão correta correspondesse exatamente a 1 ponto percentual na nota final. Essa escolha metodológica justifica-se pela amostra ter sido deliberadamente fixada em 100 questões, permitindo que o percentual de acertos coincida numericamente com o número absoluto de acertos, facilitando a compreensão dos resultados tanto para leitores da área odontológica quanto para profissionais da saúde sem formação aprofundada em estatística. Foram excluídas questões que exigiam interpretação de imagens (radiografias, gráficos complexos não descritos em texto), bem como itens com formulários incompletos ou cujo gabarito oficial não estava disponível em fontes confiáveis.

Foram testados três modelos de linguagem de grande escala, em suas versões gratuitas disponíveis durante o período da coleta (abril/2026):

1. ***ChatGPT*** *– modelo gratuito (ex.: ChatGPT 3.5 ou versão gratuita vigente, via plataforma web da OpenAI)*
2. ***DeepSeek*** *– versão gratuita disponível via aplicação web*
3. ***Gemini*** *– versão gratuita (ex.: Gemini 1.0 Pro ou Gemini Flash, via Google AI Studio ou interface web)*

Todos os sistemas foram acessados em janelas anônimas ou limpas, sem histórico de conversas prévias. Cada uma das 100 questões foi apresentada textualmente aos três sistemas de IA e o seguinte prompt padrão foi utilizado para todos os sistemas e todas as questões:

"Responda a seguinte questão de concurso público para cirurgião-dentista, na área de saúde bucal coletiva, assinalando apenas a alternativa correta (letra A, B, C, D ou E). Não forneça explicações, justificativas ou comentários adicionais. Questão: [texto completo da questão com alternativas]"

As respostas foram registradas literalmente pelo pesquisador em uma planilha eletrônica (Microsoft Excel), anotando-se a alternativa apresentada por cada IA. Em casos de respostas ambíguas (ex.: "Acredito que seja a letra B" ou múltiplas opções), a questão foi reaplicada uma única vez em nova sessão com o prompt reforçado; persistindo ambiguidade, a resposta foi considerada incorreta.

A taxa de acerto de cada IA foi calculada pela seguinte fórmula:

Taxa de acerto (%) = (Número de questões corretas / 100) × 100

Cada questão correta adicionou exatamente 1 ponto percentual à taxa final.

O estudo utilizou exclusivamente questões de domínio público (provas de concursos oficiais amplamente divulgadas), não envolvendo seres humanos, dados sensíveis ou conteúdo protegido por direitos autorais além do permitido para fins acadêmicos. Em razão dessas características o presente estudo dispensou a apreciação por um Comitê de Ética em Pesquisa (CEP), estando em conformidade com a LGPD.

RESULTADOS

Foram analisadas 100 questões de concurso público na área de saúde bucal coletiva, aplicadas a três modelos de inteligência artificial generativa (ChatGPT, DeepSeek e Gemini) em suas versões gratuitas. Cada questão correta correspondeu a 1 ponto percentual na taxa final de acerto. O desempenho geral dos sistemas é apresentado na **Tabela 1**.

REVISTA CIENTÍFICA RECISATEC
ISSN 2763-8405

Tabela 1. Desempenho geral dos modelos de IA nas 100 questões de saúde bucal coletiva

Gabarito	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
A																																																			
B																																																			
C																																																		
D																																																		
E																																																		
ChatGPT																																																		
A																																																		
B																																																		
C																																																		
D																																																		
E																																																		
DeepSeek																																																		
A																																																		
B																																																		
C																																																		
D																																																		
E																																																		
Gemini																																																		
A																																																		
B																																																		
C																																																		
D																																																		
E																																																		

Fonte: Os autores

Continuação da Tabela 1

Gabarito	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
A																																																		
B																																																		
C																																																		
D																																																		
E																																																		
ChatGPT																																																		
A																																																		
B																																																		
C																																																		
D																																																		
E																																																		
DeepSeek																																																		
A																																																		
B																																																		
C																																																		
D																																																		
E																																																		
Gemini																																																		
A																																																		
B																																																		
C																																																		
D																																																		
E																																																		

O ChatGPT obteve o melhor desempenho, com 75 acertos (75%), seguido pelo Gemini, com 47 acertos (47%), e pelo DeepSeek, com 23 acertos (23%). A diferença entre o primeiro e o último colocado foi de 52 pontos percentuais.

O ChatGPT manteve desempenho relativamente estável em todos os blocos, com discreta queda no último segmento. O Gemini apresentou bom rendimento inicial (34 acertos nas primeiras 35 questões), porém declínio acentuado a partir da questão 35. O DeepSeek demonstrou baixo desempenho homogêneo ao longo de toda a extensão do teste.

A comparação estatística entre os modelos foi realizada por meio do teste qui-quadrado de homogeneidade, utilizando tabelas de contingência 2×2 (acerto/erro) para cada par de IAs. Os resultados são apresentados na **Tabela 2.**

Tabela 2 – Comparação pareada entre os modelos de IA – Teste qui-quadrado

Comparação	Acertos (IA1 / IA2)	Erros (IA1 / IA2)	χ^2	gl	p-valor
ChatGPT (75) vs. Gemini (47	75 / 47	25 / 53	16,47	1	< 0,001
ChatGPT (75) vs. DeepSeek (23)	75 / 23	25 / 77	54,45	1	< 0,001
Gemini (47) vs. DeepSeek (23)	47 / 23	53 / 77	12,71	1	< 0,001

Todas as comparações pareadas revelaram diferenças estatisticamente significativas ($p < 0,001$). O ChatGPT superou o Gemini em 28 pontos percentuais e o DeepSeek em 52 pontos percentuais. O Gemini, por sua vez, superou o DeepSeek em 24 pontos percentuais.

Em relação aos padrões de erro, observou-se que, das 25 questões erradas pelo ChatGPT, o Gemini também errou 20 (80%), sugerindo dificuldades compartilhadas em determinados conteúdos da saúde bucal coletiva. O DeepSeek apresentou o maior número absoluto de erros com suas respostas divergindo completamente do gabarito, incluindo opções sem proximidade temática com a resposta correta.

DISCUSSÃO

O presente estudo demonstrou que o modelo ChatGPT (75% de acertos) superou significativamente o Gemini (47%) e o DeepSeek (23%) na resolução de questões de concurso público em saúde bucal coletiva. A diferença observada foi estatisticamente robusta ($p < 0,001$ para ambas as comparações), sugerindo que, entre os modelos gratuitos testados, o ChatGPT apresenta desempenho substancialmente superior. Em um contexto real de certame público, onde a nota de corte frequentemente se situa entre 50% e 60% para cargos de cirurgião-dentista na atenção básica, apenas o ChatGPT alcançaria aprovação, enquanto Gemini e DeepSeek ficariam aquém do mínimo exigido pela maioria das bancas examinadoras (Martins *et al.,* 2026).

Os achados estão parcialmente alinhados com estudos prévios que compararam inteligências artificiais generativas em provas de saúde. Pesquisas com exames médicos relataram

que versões pagas ou avançadas do ChatGPT atingem taxas entre 60% e 75%, enquanto modelos gratuitos ou menores apresentam desempenho inferior, frequentemente abaixo de 50%. No entanto, a superioridade expressiva do ChatGPT sobre o Gemini e o DeepSeek no presente estudo contrasta com relatos da literatura técnica, que apontam o Gemini como competidor direto em tarefas de raciocínio clínico. Essa divergência pode ser possivelmente atribuída a três fatores: o conteúdo específico de saúde bucal coletiva, que exige conhecimento de políticas públicas brasileiras (SUS, PNAB, e-SUS) e epidemiologia social — temas nos quais o ChatGPT pode ter sido mais amplamente treinado em fontes de língua portuguesa; a versão gratuita do Gemini utilizada, possivelmente com capacidades reduzidas em relação à sua versão paga; e diferenças na arquitetura de cada modelo quanto à recuperação de informações contextuais e normativas (Figueiredo *et al.,* 2026).

Um achado particularmente relevante foi o padrão de erro compartilhado entre ChatGPT e Gemini. Das 27 questões erradas pelo ChatGPT, o Gemini também errou 20 (80%), indicando dificuldades comuns em determinados subdomínios da saúde bucal coletiva. Uma análise sugere que esses erros se concentraram em questões que exigiam conhecimento de portarias ministeriais específicas podendo levar a IA a “alucinar”. (Alansari; Luqman, 2026).

Em particular, análises informalizadas e testes práticos com a versão gratuita do DeepSeek sugerem que a plataforma ainda pode entregar respostas imprecisas, especialmente em tópicos normativos e de cálculo epidemiológico, o que torna seu uso desaconselhável como principal ferramenta para preparação direta de provas de concursos públicos odontológicos no Brasil, exigindo sempre validação por fontes oficiais (Park *et al.,* 2024; Yalamanchili *et al.,* 2024).

O estudo apresenta limitações que devem ser consideradas. Primeiramente, utilizou-se apenas a versão gratuita de cada modelo, não sendo possível afirmar se versões pagas ou atualizações posteriores alterariam o ranking de desempenho. Em segundo lugar, as questões foram extraídas de bancos públicos com diferentes bancas examinadoras, podendo haver variabilidade no nível de dificuldade e na clareza dos enunciados. Terceiro, o estudo não controlou o efeito da ordenação das questões (apresentadas sempre na mesma sequência para todas as inteligências artificiais), embora o uso de novas sessões a cada questão tenha minimizado o viés de memória contextual. Por fim, a generalização dos resultados para outros domínios da odontologia (como cirurgia, periodontia ou ortodontia) não é automática, uma vez que o estudo focou exclusivamente na saúde bucal coletiva.

CONSIDERAÇÕES FINAIS

Os resultados demonstraram diferenças expressivas no desempenho dos três modelos de inteligência artificial avaliados. Apenas o ChatGPT alcançaria o percentual mínimo para aprovação na maioria dos certames para cirurgiões-dentistas, indicando que, nas versões gratuitas testadas, Gemini e DeepSeek no momento, não são ferramentas confiáveis para esse fim específico.

AGRADECIMENTOS

CAPES- Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

REFERÊNCIAS

ALANSARI, Aisha; LUQMAN, Hamzah. Large language models hallucination: A comprehensive survey. **Computer Science Review**, v. 61, p. 100970, 1 ago. 2026.

ARAÚJO, Samara Lavínnya Serrano de Souza et al. Impactos do ChatGPT no ensino da Odontologia: Uma revisão de escopo. **Arquivos em Odontologia**, v. 61, p. 213–228, 20 dez. 2025.

BAMASHMOUS, Mohamed. The Role of Artificial Intelligence in Transforming Dental Public Health: Current Applications, Ethical Considerations, and Future Directions. **The Open Dentistry Journal,** 11 feb. 2025.

BHUYAN, Soumitra S. et al. Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency. **Journal of Medical Systems**, v. 49, n. 1, p. 10, 2025.

BICALHO, Gabriela Magalhães; OLIVEIRA, Arthur Henrique de; GUIDA, José Paulo de Siqueira. Desempenho da inteligência artificial em questões de processo seletivo de residência médica. **Femina**, v. 52, n. 6, p. 370–373, 14 maio 2025.

FIGUEIREDO, Maria Clara Pimenta de et al. Performance of the Artificial Intelligence large language models ChatGPT 3.5, Gemini (Google Bard), ChatGPT 4.0, and Gemini 2.5 flash in surgical subspecialty questions of Brazilian medical residency exams. **Einstein**, (São Paulo), v. 24, 2026.

MARTINS, Diogo Gonçalves dos Santos et al. Análise comparativa de desempenho entre ChatGPT, Scholar GPT e DeepSeek em provas teóricas do Conselho Brasileiro de Oftalmologia 2022. **Rev. bras.oftalmol.**, v. 85, 11 fev. 2026.

NARVAI, Paulo Capel. Saúde bucal coletiva: caminhos da odontologia sanitária à bucalidade. **Revista de Saúde Pública**, v. 40, p. 141–147, 2006.

PARK, Ye-Jean et al. Assessing the research landscape and clinical utility of large language models: a scoping review. **BMC Medical Informatics and Decision Making**, v. 24, n. 1, p. 72, 12 mar. 2024.

SAVEGNAGO, Gleica Dal' Ongaro et al. Inteligência artificial na odontologia: uma revisão narrativa de literatura. **RFO UPF**, 2024.

WANG, Shanshan et al. Generative Artificial Intelligence in Medical Imaging: Foundations, Progress, and Clinical Translation. **Research**, v. 8, p. 1029, 2025.

YALAMANCHILI, Amulya et al. Quality of Large Language Model Responses to Radiation Oncology Patient Care Questions. **JAMA Network Open**, v. 7, n. 4, p. e244630, 2 abr. 2024.