UNIVERSIDADE FEDERAL DE MATO GROSSO DO SUL
CAMPUS CAMPO GRANDE
FACULDADE DE ODONTOLOGIA

ISABELA AMORIM DE OLIVEIRA

EFICÁCIA DO CHATGPT NO DIAGNÓSTICO DE LESÕES EM ODONTOLOGIA ATRAVÉS DE ANÁLISE DESCRITIVA DE ACHADOS MICROSCÓPICOS

CAMPO GRANDE
2025

ISABELA AMORIM DE OLIVEIRA

EFICÁCIA DO CHATGPT NO DIAGNÓSTICO DE LESÕES EM ODONTOLOGIA ATRAVÉS DE ANÁLISE DESCRITIVA DE ACHADOS MICROSCÓPICOS

Trabalho de Conclusão de Curso apresentado ao Curso de Odontologia da Universidade Federal de Mato Grosso do Sul (UFMS), como requisito para obtenção do título de Cirurgiã-Dentista

Orientador: Prof. Dr. Yuri Nejaim

CAMPO GRANDE

2025

ISABELA AMORIM DE OLIVEIRA

EFICÁCIA DO CHATGPT NO DIAGNÓSTICO DE LESÕES EM ODONTOLOGIA ATRAVÉS DE ANÁLISE DESCRITIVA DE ACHADOS MICROSCÓPICOS

Trabalho de Conclusão de Curso apresentado ao Curso de Odontologia da Universidade Federal de Mato Grosso do Sul (UFMS), como requisito para obtenção do título de Cirurgiã-Dentista

BANCA EXAMINADORA

_____

Prof. Dr. Yuri Nejaim
Faculdade de Odontologia da Universidade Federal de Mato Grosso de Sul / UFMS

_____

Prof.ª Dr.ª Gabriela Moura Chicrala Toyoshima
Faculdade de Odontologia da Universidade Federal de Mato Grosso de Sul / UFMS

_____

Prof. Dr. Gleyson Kleber do Amaral Silva
Faculdade de Odontologia da Universidade Federal de Mato Grosso de Sul / UFMS

Dedico este trabalho aos meus pais,

Magno e Renata.

AGRADECIMENTOS

Agradeço aos meus pais por me guiarem no caminho do ensino sempre com disposição e incentivo desmedido, e por estenderem os braços nos momentos difíceis.

À Universidade Federal de Mato Grosso do Sul, pela oportunidade de formação em uma instituição pública, gratuita e de excelência. Sinto orgulho em fazer parte de uma universidade que valoriza o conhecimento e o crescimento pessoal e coletivo.

À Faculdade de Odontologia da Universidade Federal de Mato Grosso do Sul que foi meu berço durante todos os anos da graduação, sendo muito bem administrada pelo nosso diretor, Prof. Dr. Fabio Nakao Arashiro, que guia com extremo zelo alunos e docentes, integrando o mundo UFMS aos nosso curso com a devida atenção. Agradeço pelo compromisso de todos os professores e funcionários na forma como conduzem o ambiente de trabalho moldando profissionais éticos e humanos para o mercado.

Ao meu orientador, Prof. Dr. Yuri Nejaim, pelo cuidado e empenho na condução deste trabalho, por sua competência e disposição em mudar o ambiente à sua volta para melhor, e por amparar a quem o recorra com carinho e interesse. Você é exemplo para mim e para muitos, irei sempre me lembrar de você com afeto e admiração.

Estendo meus agradecimentos aos professores Daniella Moraes, Gabriela Chicrala e Gleyson Amaral por contribuírem, de boa vontade, no desenvolvimento deste trabalho e desprender parte do seu tempo para me ajudar.

Aos meus amigos, que foram peça fundamental para que eu tivesse uma experiência descontraída e de boas memórias na graduação, minha atual e eterna saudade.

RESUMO

Oliveira IA. Eficácia do ChatGPT no diagnóstico de lesões em odontologia através da análise descritiva de achados microscópicos. Campo Grande, 2025. [Trabalho de Conclusão de Curso – Universidade Federal de Mato Grosso do Sul].

A inteligência artificial (IA) tem se mostrado uma ferramenta promissora na medicina e odontologia, especialmente em tarefas que exigem interpretação de informações complexas, como dados clínicos, laboratoriais e descrições histopatológicas. Entre as tecnologias emergentes, destacam-se os modelos de linguagem de grande escala (LLMs), como o ChatGPT-5, capazes de processar e gerar linguagem natural de forma contextualizada, simulando aspectos do raciocínio clínico humano. Este estudo investigou o desempenho diagnóstico do ChatGPT-5, na versão gratuita, na análise de informações microscópicas de laudos de patologia oral da Universidade Federal de Mato Grosso do Sul (UFMS), coletados entre 2023 e 2025. Foram selecionadas 135 lesões que atendiam aos critérios de inclusão, distribuídas em quatro grupos: lesões odontogênicas, lesões potencialmente malignas e carcinoma espinocelular (CEC), lesões traumáticas e reativas, e lesões por agentes biológicos. O modelo recebeu descrições detalhadas da microscopia, elaborou hipóteses diagnósticas principais e secundárias. Posteriormente, informações clínicas foram incluídas para avaliar o impacto no desempenho diagnóstico. Na primeira etapa, considerando apenas descrições microscópicas, o ChatGPT-5 apresentou 74,8% de acertos na hipótese diagnóstica principal. Com a inclusão de dados clínicos, a taxa global de acerto aumentou para 78,5%. O desempenho variou entre os grupos: as lesões potencialmente malignas e CEC alcançaram 97,7% de acerto, enquanto lesões odontogênicas apresentaram 60,0% após adição das informações clínicas. Lesões traumáticas e reativas e por agentes biológicos obtiveram 76,0% e 60,0% de acerto, respectivamente. Observou-se ainda que alguns acertos correspondiam a diagnósticos conceitualmente corretos, mas com nomenclatura diferente da utilizada nos laudos originais. Os resultados indicam que o ChatGPT-5 consegue interpretar descrições histopatológicas de forma consistente, especialmente quando informações clínicas são fornecidas, mas apresenta limitações em diferenciar lesões com padrões microscópicos semelhantes ou quando detalhes clínicos são determinantes. Além disso, erros semânticos e ocorrência de alucinações reforçam a necessidade de supervisão humana.

Conclui-se que o ChatGPT-5 tem potencial como ferramenta complementar na análise de laudos histopatológicos em odontologia, oferecendo suporte ao raciocínio diagnóstico do patologista, sobretudo em lesões comuns e bem descritas. A integração futura com bases de dados específicas e informações radiográficas pode ampliar sua precisão e aplicabilidade clínica.

Palavras-chave: Inteligência Artificial Generativa, Patologia Bucal, Modelos de Linguagem de Grande Escala, Odontologia.

ABSTRACT

Oliveira IA. Performance of ChatGPT in the diagnosis of oral lesions through descriptive analysis of microscopic findings. Campo Grande, 2025. [Trabalho de Conclusão de Curso – Universidade Federal de Mato Grosso do Sul].

Artificial intelligence (AI) has emerged as a promising tool in medicine and dentistry, particularly for tasks that require interpreting complex information such as clinical data, laboratory results, and histopathological descriptions. Among emerging technologies, large language models (LLMs) such as ChatGPT-5 stand out for their ability to process and generate natural language in a contextualized manner, simulating aspects of human clinical reasoning. This study investigated the diagnostic performance of ChatGPT-5, in its free version, in analyzing microscopic information from oral pathology reports at the Federal University of Mato Grosso do Sul (UFMS), collected between 2023 and 2025. A total of 135 lesions meeting the inclusion criteria were selected and divided into four groups: odontogenic lesions, potentially malignant disorders and squamous cell carcinoma (SCC), traumatic and reactive lesions, and lesions caused by biological agents. The model received detailed microscopic descriptions and generated primary and secondary diagnostic hypotheses. Clinical information was later added to assess its impact on diagnostic performance. In the first stage, considering only microscopic descriptions, ChatGPT-5 achieved a 74.8% accuracy rate for the primary diagnostic hypothesis. After including clinical data, the overall accuracy increased to 78.5%. Performance varied across groups: potentially malignant disorders and SCC achieved 97.7% accuracy, while odontogenic lesions reached 60.0% after the addition of clinical information. Traumatic and reactive lesions and those caused by biological agents showed accuracy rates of 76.0% and 60.0%, respectively. Some responses were conceptually correct but differed in nomenclature from the original diagnoses. These findings indicate that ChatGPT-5 can consistently interpret histopathological descriptions, especially when supported by clinical data, although it still struggles to distinguish between lesions with similar microscopic patterns or when clinical details are critical. Moreover, semantic errors and occasional hallucinations highlight the need for human supervision.

In conclusion, ChatGPT-5 demonstrates strong potential as a complementary tool for analyzing histopathological reports in dentistry, supporting the diagnostic reasoning of

pathologists, particularly for common and well-characterized lesions. Future integration with specialized databases and radiographic information may further enhance its accuracy and clinical applicability.

Keywords: Generative Artificial Intelligence, Pathology, Oral, Large Language Models, Dentistry.

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A inteligência artificial (IA) tem emergido como uma das tecnologias mais transformadoras da medicina contemporânea, impulsionando avanços significativos em diagnóstico, prognóstico e apoio à decisão clínica. Seu princípio fundamental baseia-se na capacidade de máquinas simularem processos cognitivos humanos, como aprendizado, raciocínio e reconhecimento de padrões, por meio de algoritmos e redes neurais capazes de analisar grandes volumes de dados e gerar inferências automatizadas (Lobo, 2017). Na área da saúde, essa tecnologia tem se mostrado especialmente promissora em tarefas que exigem interpretação de informações complexas, como imagens médicas, dados laboratoriais e descrições clínicas detalhadas (Shen et al., 2024). De forma simplificada, os sistemas de IA aprendem a partir de exemplos, um processo conhecido como aprendizado supervisionado, e, progressivamente, ajustam seus parâmetros internos para reconhecer correlações e realizar previsões com maior precisão. Essa capacidade de aprendizado contínuo tem permitido a criação de modelos altamente sofisticados, capazes de replicar aspectos do raciocínio médico humano e de auxiliar profissionais de saúde em diferentes contextos diagnósticos (Amin & Baron, 2024).

Entre as diversas aplicações da inteligência artificial, destacam-se os modelos de linguagem de grande escala (*Large Language Models*, LLMs), cujo principal exemplo hoje é o ChatGPT, desenvolvido pela OpenAI. Esses sistemas utilizam arquiteturas do tipo *transformer*, capazes de processar e gerar linguagem natural com elevado grau de coerência contextual. O aprendizado ocorre por meio da análise de grandes volumes de textos científicos e técnicos, nos quais o modelo identifica padrões linguísticos e semânticos, permitindo-lhe formular respostas probabilisticamente consistentes com o contexto apresentado (Brown et al., 2020; OpenAI, 2025). Essa estrutura possibilita que o ChatGPT simule processos de raciocínio semelhantes aos humanos, interpretando informações clínicas e científicas de forma dinâmica e contextualizada. Estudos recentes demonstram que LLMs têm se mostrado cada vez mais eficazes em tarefas médicas complexas, o que reforça seu potencial como ferramenta de apoio à decisão clínica (Thirunavukarasu et al., 2023; Gao et al., 2025).

Além disso, esses sistemas podem alcançar níveis de acurácia comparáveis aos de profissionais humanos sobretudo quando recebem dados estruturados e bem descritos (Hirosawa et al., 2024; Rutledge, 2024). Na radiologia, por exemplo, o ChatGPT-4V

(disponível na versão paga) apresentou desempenho semelhante ao de especialistas ao interpretar imagens e correlacioná-las com hipóteses clínicas (Suh et al., 2024), enquanto em especialidades médicas demonstrou capacidade de formular diagnósticos diferenciais consistentes com base em descrições sintomáticas complexas (Jeblick et al., 2024). Embora a aplicação direta do ChatGPT na odontologia ainda seja incipiente, a utilização da inteligência artificial na área já é uma realidade consolidada, abrangendo desde o diagnóstico por imagem até o planejamento cirúrgico e a análise de lesões orais (Kolarkodi & Alotaibi, 2023; Singh et al., 2024). Esses avanços sugerem que o emprego de sistemas baseados em linguagem natural, como o ChatGPT-5, pode representar um novo paradigma no apoio e no aprimoramento do raciocínio diagnóstico em odontologia.

Diante desse cenário, este trabalho teve como objetivo investigar o desempenho diagnóstico da versão gratuita do ChatGPT-5 na interpretação de informações microscópicas de laudos de patologia oral. A pesquisa foi realizada a partir da inserção de informações de microscopia presentes nos laudos elaborados por professores especialistas da Universidade Federal de Mato Grosso do Sul (UFMS), permitindo ao modelo gerar hipóteses diagnósticas e indicar o diagnóstico mais provável. Os resultados obtidos foram confrontados com os diagnósticos descritos nos laudos originais para analisar a precisão do modelo em ambos os níveis de avaliação, tanto nas hipóteses formuladas quanto na definição diagnóstica final. Ao explorar a capacidade do ChatGPT-5 em compreender e interpretar descrições clínicas textuais, o estudo visou testar a eficiência de contribuição do modelo na tomada de formulação de hipóteses do patologista, que em muitos momentos exige colaboração em equipe para chegar ao consenso diagnóstico. Além disso, a análise dos resultados pôde fornecer informações importantes sobre a viabilidade do uso de IA como coadjuvante na tomada de decisão em odontologia, fortalecendo a integração entre tecnologia e saúde bucal.

2 METODOLOGIA

Foram selecionadas informações microscópicas dos laudos de patologia oral da UFMS, correspondentes ao período de 2023 a 2025, sendo incluídos apenas aqueles que apresentavam um número mínimo de 5 e máximo de 10 laudos, da mesma lesão, nesses três anos, sendo desconsiderados diagnósticos sugestivos. Assim, dentre as 642 lesões registradas nesse período, 135 atenderam aos critérios estabelecidos. Os laudos referentes ao ano de 2025 foram considerados até o dia 30 de setembro, correspondendo aos documentos disponíveis no momento da coleta dos dados. As informações utilizadas foram obtidas através de um banco de dados sem identificação, acompanhado do respectivo Termo de Consentimento Livre e Esclarecido (TCLE).

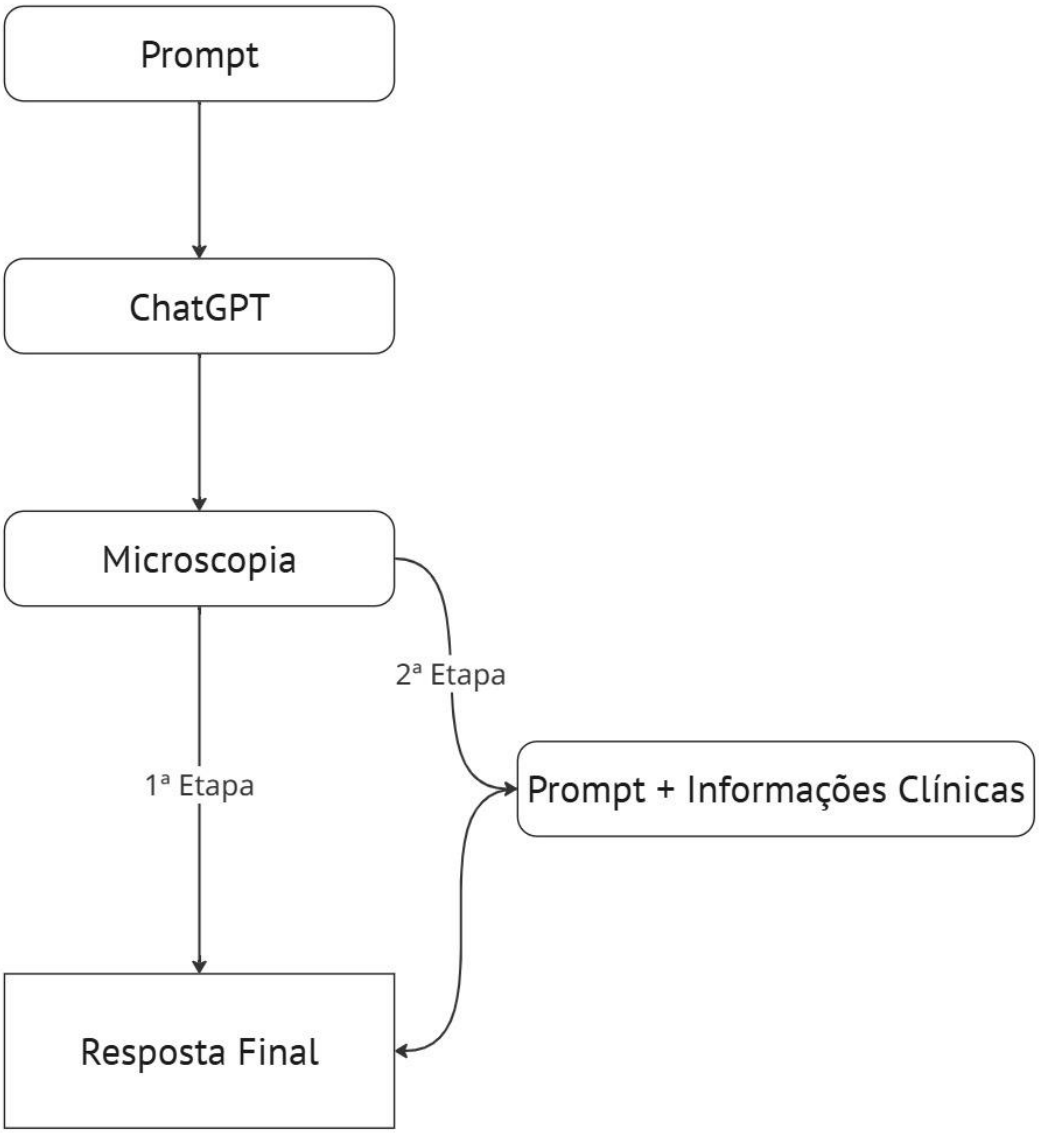
Para a extração das informações pelo ChatGPT, elaborou-se um *prompt* (instrução direcionada ao sistema de inteligência artificial para que produza uma resposta específica), o qual foi desenvolvido com o auxílio de um gerador de IA (Originality.AI, 2025), possibilitando a formulação de uma descrição mais detalhada do objetivo pretendido. Além disso, a elaboração baseou-se na arquitetura sugerida pelo próprio modelo. Assim, foi construído o seguinte texto: “A partir deste momento, você deve atuar como um especialista em Patologia Oral, com experiência em diagnóstico histopatológico de lesões bucais. Por meio das descrições microscópicas (tipo epitelial, camada basal, presença de citologia atípica, padrão de infiltração, tipo de estroma, inflamação, vasos, presença de queratinização, inclusão viral, etc.) que eu irei apresentar, elabore hipóteses diagnósticas entregando a resposta nos seguintes itens: ** > 1. **Hipótese diagnóstica principal** – a mais provável, com explicação detalhada (até 2-3 frases) sobre quais achados microscópicos sustentam essa hipótese. > 2. **Hipóteses diagnósticas secundárias** – duas outras possibilidades, cada uma acompanhada de justificativa breve (1 frase) indicando quais características poderiam levar a essa alternativa.

Após a inserção do *prompt*, as informações referentes à microscopia eram adicionadas separadamente. Com o objetivo de minimizar possíveis vieses de aprendizado do modelo, o comando foi aplicado de forma randomizada entre as diferentes lesões, iniciando-se um novo ciclo a cada resposta obtida.

Os dados coletados foram separados em uma planilha, onde foram assinaladas em cor verde para os acertos na hipótese principal e em vermelho para os acertos como hipóteses secundárias. Nos casos em que o modelo apresentou respostas em 2° e 3°

hipóteses ou incorretas, realizou-se uma nova rodada de inserções para essas lesões em específico. Para isso, o mesmo método aplicado inicialmente foi seguido, somado a um novo comando após dada a informação microscópica: “Agora, com essas informações clínicas, refaça suas hipóteses:”.

Figura 1 – Fluxograma da análise



Fonte: Autoral

Após cada nova inserção correspondente a essa etapa, foi determinado como 0 para as lesões que erraram totalmente as hipóteses, 1 para acertos na hipótese principal e 2 e 3 para segunda e terceira hipóteses, respectivamente.

A análise dos resultados foi feita tanto considerando os resultados gerais quanto para a seguinte divisão de grupos (Tabela 1):

Tabela 1 – Grupo de lesões

Grupo de Lesões	Lesões Incluídas
Lesões odontogênicas	Cisto dentígero; Cisto radicular; Queratocisto odontogênico; Capuz pericoronário
Lesões potencialmente malignas e carcinoma espinocelular (CEC)	Displasia epitelial leve; Displasia epitelial moderada; Displasia epitelial intensa; Carcinoma espinocelular; CEC bem diferenciado; CEC moderadamente diferenciado
Lesões traumáticas e reativas	Granuloma piogênico; Hiperplasia fibrosa; Hiperplasia fibrosa inflamatória; Hiperqueratose e acantose; Mucocele; Rânula
Lesões por agentes biológicos	Paracoccidioidomicose; Papiloma escamoso oral

Vale ressaltar que o uso do ChatGPT neste estudo foi realizado em sua versão gratuita, com o objetivo de avaliar seu desempenho de forma acessível e inclusiva, permitindo que seus resultados reflitam o potencial real da ferramenta quando utilizada por qualquer usuário, sem custo associado.

3 RESULTADOS

Na primeira etapa, considerando apenas as descrições microscópicas, o ChatGPT-5 apresentou 74,8% de acertos na formulação da hipótese diagnóstica principal, correspondendo a 101 acertos em um total de 135 lesões analisadas. Em 25,2% dos casos (34/135), as hipóteses sugeridas inicialmente não coincidiram com o diagnóstico final estabelecido pelo patologista.

Quando informações clínicas complementares foram adicionadas a esses 34 casos inicialmente incorretos, observou-se um aumento no desempenho diagnóstico. Dentre esses casos, 5 (14,7%) foram corrigidos na primeira hipótese, 1 (2,9%) apresentou acerto na segunda hipótese, e 5 (14,7%) acertaram na terceira hipótese. Ainda assim, 23 casos (67,6%) permaneceram incorretos, mesmo após a inclusão dos dados clínicos.

Dessa forma, ao considerar o conjunto das descrições microscópicas associadas às informações clínicas, o modelo atingiu um total de 106 acertos em 135 lesões, representando uma taxa global de acerto de 78,5%, enquanto 29 casos (21,4%) mantiveram diagnóstico incorreto (Tabela 2).

Tabela 2 – Análise geral

Etapa de avaliação	**Total de lesões avaliadas**	**Acertos (n)**	**Percentual de acerto**	**Erros (n)**	**Percentual de erro**
1ª etapa – Somente descrição microscópica	135	101	74,8%	34	25,2%
2ª etapa – Inclusão de informações clínicas nos 34 erros	34	5 (1ª Hipótese) 1 (2ª Hipótese) 5 (3ª Hipótese)	14,7% 2,9% 14,7%	23	67,6%
Total (microscopia + clínica)	135	106	78,5%	29	21,4%

Na análise por grupos de lesões (Tabela 3), observou-se que o desempenho diagnóstico do ChatGPT-5 variou de acordo com o tipo de patologia avaliada. Nas lesões

odontogênicas, o modelo apresentou 45,0% de acertos (9/20) ao considerar apenas as descrições microscópicas. Após a inclusão das informações clínicas complementares, foram obtidos 3 novos acertos, elevando a taxa de acerto para 60,0% (12/20).

Nas lesões potencialmente malignas e CEC, o desempenho foi significativamente superior, com 97,7% de acertos (44/45) já na primeira etapa. A adição de dados clínicos não alterou esse resultado, mantendo a taxa inicial**.**

Para as lesões traumáticas e reativas, o ChatGPT-5 atingiu 72,0% de acertos (36/50) na análise inicial. Após a consideração das características clínicas, ocorreram 2 novos acertos, resultando em uma taxa final de 76,0%.

Por fim, nas lesões por agentes biológicos, observou-se um desempenho intermediário, com 60,0% de acertos (12/20) e 40,0% de erros (8/20) na primeira etapa. A inclusão das informações clínicas não modificou o resultado, mantendo o mesmo percentual.

Tabela 3 – Análise por grupo

Grupo de Lesões	**Total de lesões avaliadas (n)**	**Acertos (n)**	**Percentual de acerto**	**Erros (n)**	**Percentual de erro**	**Acertos após inclusão de informações clínicas (n)**	**Percentual de acerto após informações clínicas**
Lesões odontogênicas	20	9	45,0%	11	55,0%	3	60,0%
Lesões potencialmente malignas e CEC	45	44	97,7%	1	2,2%	0	0%
Lesões traumáticas e reativas	50	36	72,0%	14	28,0%	2	76,0%
Lesões por agente biológico	20	12	60,0%	8	40,0%	0	0%

4 DISCUSSÃO

Os resultados obtidos neste estudo demonstram que o ChatGPT-5 gratuito apresentou desempenho expressivo na formulação de hipóteses diagnósticas baseadas em descrições microscópicas de lesões orais, alcançando uma taxa global de acerto de 74,8% na primeira etapa e de 78,5% após a inclusão de informações clínicas. Esses achados indicam que, mesmo sem o suporte de imagens histológicas, o modelo foi capaz de interpretar descrições textuais complexas e correlacioná-las a entidades patológicas específicas.

O aumento de acurácia após a adição das informações clínicas sugere que a integração entre dados clínicos e histopatológicos melhora o desempenho diagnóstico da IA, o que se alinha com a literatura que destaca a importância da contextualização clínica para a formulação diagnóstica precisa, tanto por humanos quanto por sistemas computacionais (Patel et al., 2019). Estudos prévios mostram que modelos de linguagem tendem a apresentar melhor desempenho quando recebem múltiplas fontes de informação, pois o raciocínio contextual reduz ambiguidades e aumenta a consistência das respostas (Anderson-Luk & Ip, 2024).

Ao analisar os grupos de lesões separadamente, observou-se variação significativa no desempenho diagnóstico do modelo. As lesões potencialmente malignas e CEC apresentaram o melhor resultado, com 97,7% de acertos, demonstrando que o ChatGPT-5 foi capaz de reconhecer padrões textuais característicos de displasias epiteliais e carcinomas, cuja descrição histológica tende a ser mais específica e padronizada. Esse resultado é consistente com estudos em medicina geral, nos quais os modelos GPT demonstraram alta precisão em identificar neoplasias e padrões celulares atípicos a partir de descrições clínicas e histológicas (Mavrych et al., 2025; Mazzucchelli et al., 2025). Além disso, o reconhecimento de termos associados à atipia celular, queratinização e invasão tecidual pode ter contribuído para a elevada taxa de acerto nesse grupo. Dos 20 casos de carcinoma de células escamosas incluídos, 13 foram corretamente classificados quanto ao grau de diferenciação, resultando em uma taxa de acerto de 65%.

Em contraste, o desempenho nas lesões odontogênicas foi inferior, 45,0% de acertos inicialmente e 60,0% após as informações clínicas. Esse resultado pode ser explicado pela sobreposição de características microscópicas entre diferentes cistos e tumores odontogênicos, que frequentemente compartilham padrões epiteliais

semelhantes e exigem correlação com achados radiográficos e clínicos para confirmação diagnóstica (Neville et al., 2016; Philipsen & Reichart, 2004). Assim, a ausência de dados radiológicos pode ter limitado a precisão do modelo, evidenciando uma limitação importante do uso exclusivo de IA textual na patologia oral.

Nas lesões traumáticas e reativas, a taxa de acertos foi intermediária, 72,0% inicialmente e 76,0% após informações clínicas, o que demonstra que o modelo conseguiu identificar padrões compatíveis com processos inflamatórios e reparadores. No entanto, ainda apresentou limitações em nomear corretamente algumas lesões específicas, como granuloma piogênico. Esse viés foi observado em 9 dos 10 laudos incluídos, indicando uma tendência do modelo em identificar o tipo de processo patológico, mas não o diagnóstico nominal preciso.

Por outro lado, o desempenho em lesões por agentes biológicos (60,0% de acertos) indica que o modelo possui certo reconhecimento de padrões infecciosos, mas sua precisão ainda é limitada. Isso pode estar relacionado à menor frequência dessas lesões nos bancos de dados de treinamento e à complexidade das manifestações histopatológicas causadas por microrganismos específicos, como fungos e vírus (Wang et al., 2023). Como exemplo podemos citar a blastomicose e histoplasmose onde houve alta preferência do modelo por indicá-las como hipótese principal para paracoccidioidomicose, o que pode ser explicado pela alta taxa dessas doenças nos Estado Unidos da América (EUA) em comparação com o Brasil (Kauffman, Pappas & Patterson, 2021; Shikanai-Yasuda et al., 2017) e, portanto, representar um possível aprendizado maior do modelo vindo dessa região.

Na análise detalhada dos acertos, observou-se que, embora o ChatGPT-5 tenha apresentado desempenho satisfatório na formulação das hipóteses diagnósticas, parte desses acertos não correspondeu exatamente ao nome específico da lesão, como já foi apontado. Dos 101 casos classificados como corretos, 35 (34,6%) referiam-se a diagnósticos que, apesar de conceitualmente compatíveis com a patologia descrita, não apresentavam o mesmo termo utilizado pelo patologista no laudo final. Esse achado evidencia uma limitação semântica do modelo, que demonstra compreensão contextual das características microscópicas, mas ainda apresenta dificuldade em empregar a nomenclatura técnica com precisão.

Além disso, na segunda etapa da pesquisa, em que foram incluídas informações clínicas adicionais, observou-se um comportamento distinto em relação à primeira fase. Das 48 novas inserções, que incluíam tanto casos inicialmente incorretos quanto aqueles

com erros na nomeação exata da lesão, verificou-se que, em 12 situações, o modelo modificou suas hipóteses e passou a acertar o diagnóstico antes mesmo da adição dos dados clínicos. Esse comportamento sugere que o ChatGPT-5 realiza um processo interno de reinterpretação a partir de novas tentativas, ajustando suas respostas sem necessariamente depender de informações complementares, possivelmente por reorganização probabilística dos padrões textuais utilizados em seu treinamento (Liu et al., 2024).

Durante o processo experimental, foi registrada também a ocorrência de uma alucinação (termo utilizado para descrever quando o modelo gera informações inexistentes ou incorretas), na qual o ChatGPT-5 mencionou uma patologia que não existe na literatura. Esse tipo de erro reforça a necessidade de supervisão humana constante, sobretudo em aplicações diagnósticas, uma vez que respostas semanticamente coerentes podem não ter respaldo científico.

Outro ponto de destaque foi a análise das respostas referentes à mucocele e rânula, lesões que compartilham a mesma natureza patológica, diferindo apenas pela localização anatômica. Mesmo após a inclusão de informações clínicas que indicavam claramente a região da lesão, o modelo manteve erros de classificação, acertando apenas 2 em um total de 5 tentativas. Essa dificuldade reforça que o ChatGPT-5 tende a priorizar descrições histológicas gerais, sem integrar adequadamente aspectos anatômicos específicos ao raciocínio diagnóstico. De modo semelhante, podemos mencionar os casos em que houve também indicações de diagnósticos clínicos para algumas lesões mesmo não fornecendo informações para tal, como no caso de algumas displasias.

Além disso, observou-se que o desempenho do modelo foi influenciado pela quantidade de informações fornecidas dentro de um mesmo chat. À medida que mais dados eram inseridos, o ChatGPT-5 passou a apresentar respostas mais lentas e aumento na taxa de erro, sugerindo possível sobrecarga contextual ou perda de coerência ao longo da interação. Curiosamente, ao iniciar uma nova conversa e reinserir o mesmo caso clínico, o modelo voltou a demonstrar maior precisão diagnóstica, indicando que o acúmulo de contexto pode interferir negativamente na consistência de suas respostas.

Não obstante, a variabilidade observada entre os grupos de lesões neste estudo, com maior desempenho em lesões potencialmente malignas e carcinoma espinocelular, e menor em lesões odontogênicas e reativas, sugere que o desempenho do modelo depende diretamente da frequência e qualidade das informações disponíveis durante o treinamento. Modelos de linguagem são mais eficientes na interpretação de descrições

amplamente documentadas em bases públicas (Kandpal et al., 2023), o que justifica o melhor desempenho em patologias de alta prevalência e terminologia bem consolidada.

A presença de erros semânticos e de alucinações, embora pontual, é amplamente reconhecida na literatura como um dos principais desafios para a aplicação clínica de LLMs (*Large Language Models*) em saúde (Pal, Umapathi & Sankarasubbu, 2023). Esses erros ocorrem devido à forma como o modelo prediz palavras com base em probabilidades linguísticas, e não em validação factual, o que pode gerar respostas plausíveis, porém cientificamente incorretas (OpenAI, 2025). Assim, a utilização do ChatGPT como ferramenta diagnóstica deve ser entendida como um apoio interpretativo, e não como um substituto do julgamento clínico do patologista.

Outrora, a dificuldade do modelo em diferenciar lesões semelhantes, reflete limitações na integração entre achados clínicos e histológicos, indicando que o raciocínio diagnóstico ainda se apoia majoritariamente em padrões textuais e não em correlações anatômicas complexas (Cheng, 2024). Esse ponto reforça a importância da contextualização clínica e da descrição detalhada no processo diagnóstico automatizado.

Outro aspecto relevante é a instabilidade de desempenho observada com o aumento do número de interações dentro de um mesmo chat. Estudos prévios também relatam que, à medida que o contexto se expande, o modelo tende a “diluir” informações e cometer mais erros interpretativos (Du et al., 2025). Esse fenômeno decorre da forma como o modelo armazena e prioriza dados no histórico da conversa, o que pode comprometer a precisão em análises extensas (Laban et al., 2025).

De modo geral, os resultados obtidos indicam que o ChatGPT-5, ainda que de forma gratuita, apresenta potencial significativo como ferramenta complementar para a análise de laudos histopatológicos em odontologia, sobretudo quando utilizado de forma controlada, com inserções objetivas e revisão humana. A tendência é que versões futuras, treinadas com bases específicas de patologia oral, possam apresentar ganhos substanciais em acurácia diagnóstica, minimizando as limitações aqui observadas.

5 CONCLUSÃO

Conclui-se que o ChatGPT-5 apresenta bom desempenho como ferramenta de apoio diagnóstico em patologia oral, com acertos de 74,8% baseados apenas nas descrições microscópicas e 78,5% após a inclusão de dados clínicos. Esses achados sugerem que o modelo consegue reproduzir, em parte, o processo de raciocínio diagnóstico do patologista humano, especialmente em lesões com características bem definidas, embora ainda apresente limitações relacionadas à clareza das informações. Ressalta-se ainda que, mesmo em sua versão gratuita, o modelo demonstrou capacidade de interpretação consistente e aplicabilidade prática, o que amplia seu potencial de uso em contextos acadêmicos e clínicos com recursos limitados, tornando a inteligência artificial mais acessível e democratizando seu papel. Assim, o ChatGPT-5 se mostra promissor como suporte cognitivo, auxiliando o profissional na formulação de hipóteses diagnósticas, desde que usado de forma complementar e sob supervisão humana.

6 REFERÊNCIAS

1. Amin T, Baron RJ. Large Language Models in Medical Practice: Opportunities, Risks, and Ethical Considerations. JAMA. 2024;331(4):305–306. PMID: 38746668

2. Anderson-Luk DW, Ip WCT, et al. Performance of GPT-4 and GPT-3.5 in generating accurate and comprehensive diagnoses across medical subspecialties. J Chin Med Assoc. 2024;87(5):[Epub ahead of print]. Available from: https://journals.lww.com

3. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–1901.

4. Cheng J. Applications of Large Language Models in Pathology. Bioengineering. 2024;11(4):342.

5. Du Y, Tian M, Ronanki S, Rongali S, Bodapati S, Galstyan A, et al. Context length alone hurts LLM performance despite perfect retrieval. arXiv preprint arXiv:2510.05381. 2025.

6. Gao S, et al. Artificial intelligence in dentistry: a narrative review of diagnostic and therapeutic applications. Med Sci Monit. 2025;40195079.

7. Hirosawa T, et al. Evaluating ChatGPT-4's Accuracy in Identifying Final Diagnoses Within Differential Diagnoses Compared With Those of Physicians. JMIR Form Res. 2024;38924784.

8. Jeblick K, et al. ChatGPT in medical diagnostics: performance and limitations in complex clinical cases. NPJ Digit Med. 2024;40312678.

9. Kandpal N, Deng H, Roberts A, Wallace E, Raffel C. Large Language Models Struggle to Learn Long Tail Knowledge. Proceedings of the 40th International Conference on Machine Learning. 2023;202:15696 15707.

10. Kauffman CA, Pappas PG, Patterson TF. Blastomycosis and histoplasmosis: clinical and epidemiologic features. Clin Chest Med. 2021;42(3):331-344.

11. Kolarkodi SH, Alotaibi KZ. Artificial Intelligence in Diagnosis of Oral Diseases: A Systematic Review. J Contemp Dent Pract. 2023;37189014.

12. Laban P, Hayashi H, Zhou Y, Neville J. LLMs get lost in multi-turn conversation. Proc Assoc Comput Linguistics. 2025; [online ahead of print].

13. Liu D, Nassereldine A, Yang Z, Xu C, Hu Y, Li K, et al. Large Language Models have Intrinsic Self Correction Ability. arXiv preprint arXiv:2406.15673. 2024.

14. Lobo LC. Inteligência artificial e medicina. Rev Bras Educ Med. 2017;41(4):1–8. doi:10.1590/1981-52712015v41n4e20170016

15. Mavrych V, Yousef EM, Yaqinuddin A, Bolgova O. Large language models in medical education: a comparative cross-platform evaluation in answering histological questions. Med Educ Online. 2025;30(1):2534065. doi:10.1080/10872981.2025.2534065. PMID:40651009

16. Mazzucchelli M, Salzano S, Caltabiano R, Magro G, Certo F, Barbagallo G, Broggi G, et al. Diagnostic performance of ChatGPT-4.0 in histopathological analysis of gliomas: a single institution experience. Neuropathology. 2025;45(4):e70023. doi:10.1111/neup.70023. PMID:40726356

17. Neville BW, Damm DD, Allen CM, Chi AC. Oral & Maxillofacial Pathology. 4th ed. Elsevier; 2016.

18. OpenAI. Introducing GPT-5 [Internet]. 2025 Aug 7 [cited 2025 Oct 16]. Available from: https://openai.com/index/introducing-gpt-5

19. OpenAI. Why language models hallucinate [Internet]. OpenAI Research Publication; 2025 Sep 5 [cited 2025 Oct 16]. Available from: https://openai.com/index/why-language-models-hallucinate

20. Originality.AI. AI Prompt Generator [Internet]. 2025 [cited 2025 Oct. Available from: https://originality-ai.translate.goog/blog/ai-prompt-generator?_x_tr_sl=en&_x_tr_tl=pt&_x_tr_hl=pt&_x_tr_pto=tc

21. Pal A, Umapathi LK, Sankarasubbu M. Med HALT: Medical Domain Hallucination Test for Large Language Models. In: Proc. 27th Conf. on Computational Natural Language Learning (CoNLL). Singapore; 2023:314 334.

22. Patel BN, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. NPJ Digit Med. 2019;2:111.

23. Philipsen HP, Reichart PA. Odontogenic tumours and allied lesions. Quintessence Publishing; 2004.

24. Rutledge GW. Diagnostic accuracy of GPT-4 on common clinical scenarios and challenging cases. Learn Health Syst. 2024;39036534.

25. Shen Y, Heacock L, Elias J, et al. Large Language Models in Medicine: The Potentials and Pitfalls. J Am Med Inform Assoc. 2024 Mar;31(3):561–568. PMID: 38285984

26. Shikanai-Yasuda MA, Mendes RP, Colombo AL, et al. Brazilian guidelines for the clinical management of paracoccidioidomycosis. Rev Soc Bras Med Trop. 2017;50(5):715-740.

27. Singh N, et al. Applications of Artificial Intelligence in Dental Diagnostics and Decision-Making: A Scoping Review. Int J Environ Res Public Health. 2024; PMC10982745.

28. Suh PS, et al. Comparing Diagnostic Accuracy of Radiologists versus GPT-4V and Gemini Pro Vision Using Image Inputs from Diagnosis Please Cases. Radiology. 2024;38980179.

29. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nat Med. 2023;29(8):1930–1940. PMID: 37548042.

30. Wang L, Zhang Y, Liu C, et al. Artificial intelligence in the diagnosis of infectious diseases: a comprehensive review. Frontiers in Microbiology. 2023;14:1180672.