Clinical Oral Investigations (2024) 28:544
https://doi.org/10.1007/s00784-024-05939-1

RESEARCH

Accuracy and consistency of ChatGPT-3.5 and −4 in providing differential diagnoses in oral and maxillofacial diseases: a comparative diagnostic performance analysis

Saygo Tomo[1] · Jérôme R. Lechien[2,3,4,5] · Hugo Sobrinho Bueno[6] · Daniela Filié Cantieri-Debortoli[6] · Luciana Estevam Simonato[6,7,8]

Received: 20 June 2024 / Accepted: 14 September 2024 / Published online: 24 September 2024

Abstract

Objective To investigate the performance of ChatGPT in the differential diagnosis of oral and maxillofacial diseases.

Methods Thirty-seven oral and maxillofacial lesions findings were presented to ChatGPT-3.5 and −4, 18 dental surgeons trained in oral medicine/pathology (OMP), 23 general dental surgeons (DDS), and 16 dental students (DS) for differential diagnosis. Additionally, a group of 15 general dentists was asked to describe 11 cases to ChatGPT versions. The ChatGPT-3.5, -4, and human primary and alternative diagnoses were rated by 2 independent investigators with a 4 Likert-Scale. The consistency of ChatGPT-3.5 and −4 was evaluated with regenerated inputs.

Results Moderate consistency of outputs was observed for ChatGPT-3.5 and −4 to provide primary ($\kappa = 0.532$ and $\kappa = 0.533$ respectively) and alternative ($\kappa = 0.337$ and $\kappa = 0.367$ respectively) hypotheses. The mean of correct diagnoses was 64.86% for ChatGPT-3.5, 80.18% for ChatGPT-4, 86.64% for OMP, 24.32% for DDS, and 16.67% for DS. The mean correct primary hypothesis rates were 45.95% for ChatGPT-3.5, 61.80% for ChatGPT-4, 82.28% for OMP, 22.72% for DDS, and 15.77% for DS. The mean correct diagnosis rate for ChatGPT-3.5 with standard descriptions was 64.86%, compared to 45.95% with participants' descriptions. For ChatGPT-4, the mean was 80.18% with standard descriptions and 61.80% with participant descriptions.

Conclusion ChatGPT-4 demonstrates an accuracy comparable to specialists to provide differential diagnosis for oral and maxillofacial diseases. Consistency of ChatGPT to provide diagnostic hypotheses for oral diseases cases is moderate, representing a weakness for clinical application. The quality of case documentation and descriptions impacts significantly on the performance of ChatGPT.

Clinical relevance General dentists, dental students and specialists in oral medicine and pathology may benefit from ChatGPT-4 as an auxiliary method to define differential diagnosis for oral and maxillofacial lesions, but its accuracy is dependent on precise case descriptions.

Keywords Artificial intelligence · Performance · ChatGPT-4 · ChatGPT-3.5 · Oral diseases · Diagnosis

✉ Saygo Tomo
saygotomo@hotmail.com

[1] Department of Pathology, School of Dentistry, University of São Paulo, Av. Professor Lineu Prestes 2227, São Paulo CEP 05508-000, Brazil

[2] Research Committee of Young-Otolaryngologists of the International Federations of Oto-rhino- laryngological Societies (YO-IFOS), Paris, France

[3] Department of Otorhinolaryngology and Head and Neck Surgery, CHU de Bruxelles, CHU Saint-Pierre, Brussels, Belgium

[4] Department of Otorhinolaryngology and Head and Neck Surgery, School of Medicine, Foch Hospital, UFR Simone Veil, Université Versailles Saint-Quentin-en-Yvelines (Paris Saclay University), Paris, France

[5] Department of Surgery, Faculty of Medicine, UMONS Research Institute for Health Sciences and Technology, University of Mons (UMons), Mons, Belgium

[6] Dental School, University Brasil, Fernandópolis, Brazil

[7] Medical School, University Brasil, Fernandópolis, Brazil

[8] Instituto Científico e Tecnológico, Programas de Bioengenharia e Ciências Ambientais, Universidade Brasil, Fernandópolis, Brazil

Introduction

The potential applications of Artificial Intelligence (AI) have gained prominence within healthcare and health sciences [1]. As AI technologies advance, their integration into medical specialties has become increasingly notable [1]. Among these innovations, the recently released AI-powered chatbot, Generative Pretrained Transformer (ChatGPT), has shown significant potential to address a wide range of medical inquiries [2]. Furthermore, this technology has the capability to enhance the quality of scientific reports and offer valuable insights into complex medical questions [2].

In the field of dentistry, the integration of AI tools for both academic and clinical purposes have similarly expanded [3, 4]. While many researchers and clinicians hold great expectations for the integration of tools like ChatGPT into oral healthcare, others may present some resistance, generating some debate [5–7]. The use of AI in oral medicine and pathology is a developing area but reports on the accuracy of ChatGPT in responding to targeted theoretical [8] and diagnostic [9] questions remain limited. A study by Vaira et al. [10] highlighted that ChatGPT-4 demonstrated high performance in addressing closed- and open-ended questions, as well as clinical scenarios related to oral surgery, oral oncology, and salivary gland pathology. However, this study did not include real clinical cases, what limits the practical applicability of these findings.

Despite limitations, the potential of ChatGPT to assist in clinical settings remains significant [11]. By providing differential diagnoses for oral and maxillofacial lesions, ChatGPT could greatly enhance clinical reasoning, decision-making, and patient guidance. The utility of AI-powered tools like ChatGPT has been demonstrated across various medical specialties, suggesting a promising future for their application in improving patient care and clinical outcomes [12–19].

The full extent of ChatGPT's potential to provide differential diagnoses for oral and maxillofacial lesions remains a topic of subjective evaluation. Thus, the aims of this study are to compare the performance of ChatGPT versions, specialists in Oral Medicine, General Dental Surgeons, and Dental Students to provide differential diagnoses for selected cases of oral diseases, and to assess the consistency of ChatGPT across repeated inputs of cases.

Table 1 Demographic and professional practice features of participants

	OMP ($n=18$)	DDS ($n=23$)	DDS-II ($n=16$)	DS ($n=18$)
Age				
Mean (range)	32.67 (25–53)	39.83 (24–60)	37.75 (24–58)	21.28 (19–60)
Sex				
Male	9 (50)	7 (30.4)	2 (12.5)	3 (16.7)
Female	9 (50)	16 (69.6)	14 (87.5)	15 (83.3)
Time of experience*				
2–5 years	8 (44.4)	2 (4.9)	5 (31.3)	-
6–10 years	5 (27.8)	7 (17.1)	3 (18.8)	-
> 10 years	5 (27.8)	14 (34.1)	8 (50)	-
Practice				
Public	10 (55.6)	13 (31.7)	7 (43.8)	-
Private	8 (44.4)	6 (14.6)	9 (56.3)	-
Public + private	0 (0)	4 (9.8)	0 (0)	-

OMP, specialists in Oral Medicine/Pathology; DDS, Dental Surgeons not specialists in Oral Medicine/Pathology; DDS-II, Dental Surgeons not specialists in Oral Medicine/Pathology that participated by describing cases; DS, Dental Students

Data presented as n (%)

* Refers to the time of experience in Oral Medicine/Pathology for OMP group, and time of general dental practice since graduation for DDS and DDS-II groups

Methods

Ethical considerations

Patient confidentiality in this study was ensured through a stringent anonymization process. The data collection forms did not include any personal or identifiable information about the participants. The study received approval from the *Universidade Brasil* Ethics Committee (73938823.7.0000.0075). The informed consent was obtained for all participants (professionals and dental students enrolled).

Study design and participants

This comparative diagnostic performance analysis involved a diverse cohort of participants with varying levels of training and experience in Oral Medicine and Oral Pathology. The first group comprised board certified Dental Surgeons specialists in Oral Medicine and/or Oral Pathology (OMP) with a minimum of two years of experience. The second group included board certified General Dental Surgeons (DDS) who are not specialists in Oral Medicine but had at least two years of general clinical experience. The final group consisted of dental students (DS) on their final year of dental school and who had passed their Oral Medicine and Oral Pathology courses without any failures.

The frequencies of demographic characteristics and professional experience of the participants are available in Table 1. The study included 18 specialists in Oral Medicine/Pathology (OMP), 23 General Dental Surgeons (DDS),

16 General Dental Surgeons who participated describing lesions to ChatGPT (DDS-II) as described below, and 18 Dental Students (DS).

Data collection

Two investigators (S.T. and L.E.S.) selected 37 archived patient cases, representing 11 categories of oral diseases. Cases' diagnoses and grouping are available in Appendix 1. Cases were selected based on the presence of classic characteristics for the given diagnosis, completeness of medical records, availability of clinical and radiographic images, and diagnosis confirmation through appropriate methods and complementary examinations (e.g., biopsy, radiography, computed tomography) (Fig. 1).

Participants from the OMP, DDS, and DS groups were presented with the 37 cases and asked to provide up to three differential diagnoses for each case, in order of probability. Each case description was prepared by S.T. and independently reviewed by L.E.S. to ensure the inclusion of all relevant information. Case details included age, gender, chief complaint, habits, relevant health issues and medication use, a full description of the lesion's clinical aspect, and imaging features when applicable. Data collection was conducted via online meetings, during which one investigator (S.T. or L.E.S.) presented the cases to the participants, who submitted their differential diagnoses through a Google® Form (Google®, Mountain View, CA, USA). The cases were presented using slide presentation software, with each slide containing the information for one case. Participants were given two minutes for each case (Fig. 1).

ChatGPT consultation

The descriptions of the 37 cases were individually entered into the ChatGPT-3.5 and −4 (OpenAI, CA, USA) API between January 2024 and April 2024. Each version of ChatGPT was prompted with the command: "*Please*, *give*

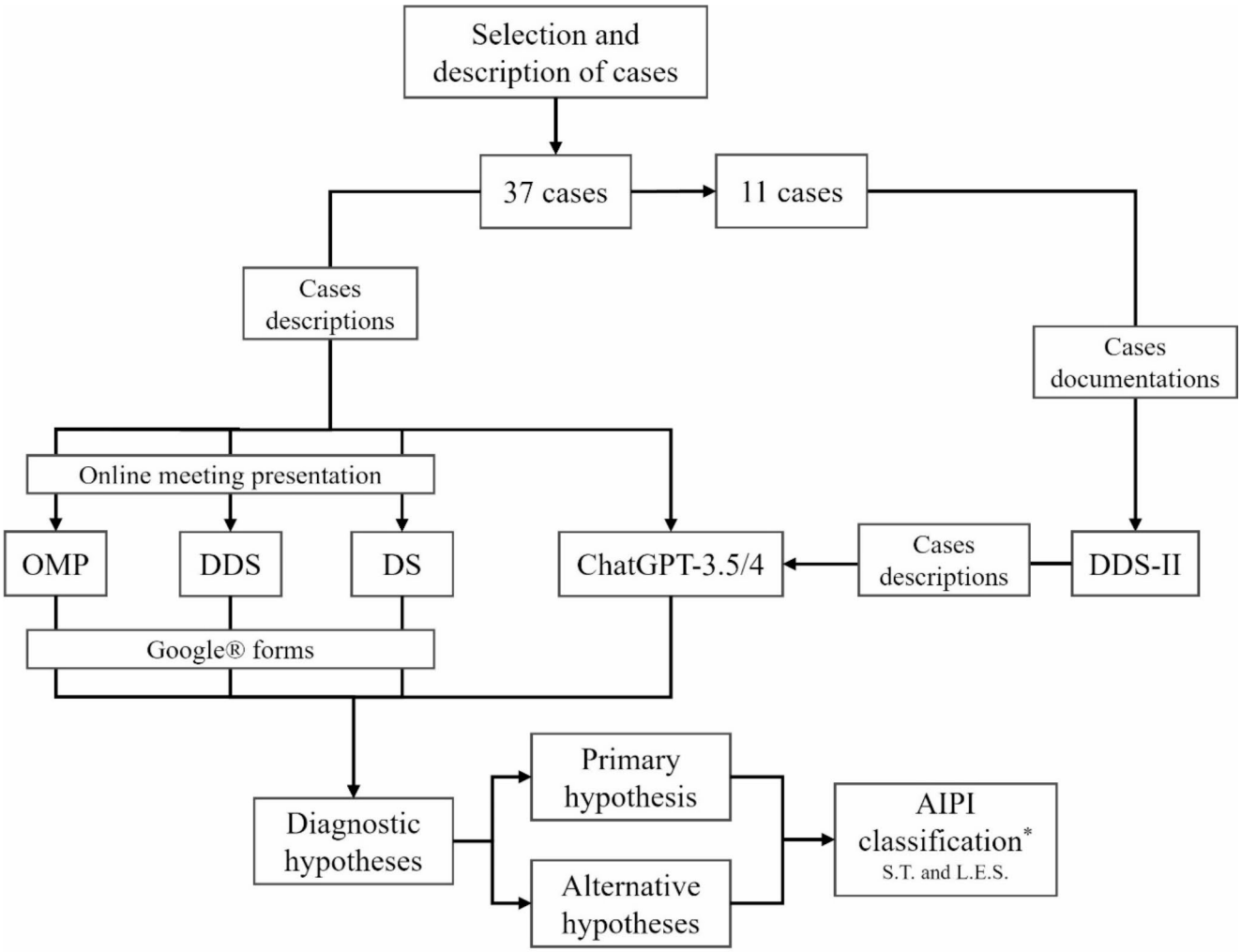

Fig. 1 Chart Flow of the study. OMP, specialists in oral medicine/pathology; DDS, general dental practioners; DS, dental students; AIPI, Artificial Intelligence Performance Instrument

me three differential diagnoses for the following case, in order of probability: [case description].” Each case was entered into each version of ChatGPT 15 times on different days (one per day) to evaluate the consistency of GPT versions (Fig. 1). The responses were collected on a database for further analysis.

To evaluate the performance of ChatGPT in providing differential diagnoses for oral diseases when imputed with case descriptions made by general dental surgeons and compare the responses with the responses for the standard descriptions (made by S.T. and reviewed by L.E.S.), an additional group (DDS-II) described cases to ChatGPT. The general dental surgeons included in this group are not the same who participated on the previous phase. Eleven cases (one from each previously described group of oral lesions) were selected from the initial 37 cases for this phase. For this phase, 11 cases were selected because on a pilot study with 5 general dentists asked to describe the 37 cases, all of them left most cases undescribed because of the length of the activity. The cases selected for this phase are identified on Appendix 1. Participants had access to the medical records of patients and clinical and/or radiographic images, and they were invited to describe the case. Each description was entered into ChatGPT versions 3.5 and 4 using the same way (Fig. 1).

Performance assessment

The differential diagnoses provided by ChatGPT versions 3.5 and 4, OMP, DDS, and DS were evaluated and classified by two investigators (S.T. and L.E.S.) in a blind manner. Hypotheses were categorized using a 4 Likert-Scale (correct, plausible, non-plausible, or absent) following the classes of the diagnosis item of the Artificial Intelligence Performance Instrument (AIPI) [20]. A hypothesis was considered correct if it matched the case diagnosis, plausible if both investigators agreed that the case features justified the hypothesis, non-plausible if the case features did not justify the hypothesis, and absent if no primary or secondary hypotheses were provided. For each participant and each ChatGPT-3.5 and −4 input, the percentage of correct or plausible primary and alternative hypotheses was calculated. In cases of discordance between evaluators, both discussed together to reach an agreement.

Statistical analysis

Categorical variables, including the demographic and professional characteristics of participants, are presented through descriptive analysis. The rates of correct or plausible primary and alternative hypotheses were obtained as % for each participant. To address the non-independence of repeated ChatGPT outputs and the non-normality of the data when comparing the percentages of correct or plausible hypothesis across the groups, we employed a Generalized Linear Mixed Model (GLMM) with a *logit* transformation applied to percentage-based outcome variables. Repeated measures (ChatGPT-3.5 and −4 outputs) were modeled with random intercepts to account for within-group correlations, while independent measures were treated as fixed effects. The OMP group was set as the intercept. For the comparison of descriptions made by DDS-II with standard descriptions, the GLMM was applied with the percentages due to the normal distribution of data in this phase of the study. GLMM analyses were performed on Python (Python Software Foundation, version 3.12.5), and the ‘statsmodels’ package [21] and ‘matplotlib’ library to generate visual plots [22]. The indicators of quality of descriptions of lesions by DDS-II was analyzed by Kruskal-Wallis test followed by Dunn’s *post hoc* test on GraphPad Prism version 8.0®. Adjusted p-values for Dunn’s test are presented, considering multiple comparisons. The effect size (ES) of Kruskal-Wallis results were calculated using the *epsilon²* method. A power analysis of all Kruskal-Wallis tests was performed to assess the sensitivity to detect effects given the sample size, and all obtained values > 0.8. For analysis of the consistency of responses, the primary hypothesis and alternative hypothesis were coded according to the scores of the diagnostic item of AIPI and submitted to a Fleiss Kappa test using IBM SPSS Version 26®. The consistency was classified according to Koch’s interpretation. *P*-values < 0.05 were considered statistically significant.

Results

Diagnostic performance analysis

Descriptive analysis

As detailed on Table 2, OMPs achieved the highest percentages across most diagnostic measures, with scores consistently above the other groups. ChatGPT-4 followed closely and excelled OMP in the categories plausible primary hypotheses and plausible/correct alternative hypotheses. ChatGPT-3.5 showed moderate percentages across all measures, generally performing better than the DDS and DS groups. The DDS group had notably lower percentages, especially in plausible hypotheses and alternative diagnoses, while the DS group recorded the lowest percentages in all categories, with minimal variation between measures.

Table 2 Mean (SD) of correct diagnosis, correct primary hypothesis, plausible primary hypothesis and plausible/correct alternative hypothesis by OMP, ChatGPT-3.5, ChatGPT-4, DDS and DS

	Correct Diagnosis considering any hypothesis	Correct primary hypotheses	Plausible primary hypotheses	Plausible/correct alternative hypotheses
OMP	86.64 (6.25)	82.28 (8.04)	59.78 (20.61)	34.98 (16.57)
ChatGPT-3.5	64.86 (8.49)	45.95 (4.68)	52.11 (14.59)	67.57 (7.08)
ChatGPT-4	80.18 (6.02)	61.80 (5.29)	74.75 (13.65)	76.76 (4.98)
DDS	24.32 (13.31)	22.72 (12.33)	9.36 (6.63)	3.38 (4.55)
DS	16.67 (7.82)	15.77 (7.71)	7.34 (3.54)	1.20 (1.38)

Data is presented as %

OMP, Specialists in Oral Medicine/Pathology; DDS, Dental Surgeons; DS, Dental Students; SD, Standard Deviation

Table 3 Inter-rater reliability for ChatGPT-3.5 and −4 consults, OMP, DDS, and DS according to AIPI scores

	Primary hypothesis		Alternative hypotheses	
Group	κ-value	CI (95%)	κ-value	CI (95%)
ChatGPT-3.5	0.532	0.531–0.533	0.337	0.336–0.338
ChatGPT-4	0.533	0.532–0.533	0.368	0.367–0.369
OMP	0.217	0.217–0.218	0.182	0.181–0.183
DDS	0.161	0.160–0.161	0.071	0.070–0.071
DS	0.175	0.174–0.175	0.040	0.039–0.040

OMP, Specialists in Oral Medicine/Pathology; DDS, Dental Surgeons not specialists in Oral Medicine/Pathology; DS, Dental Students; CI, Confidence Interval

κ-Fleiss Kappa,

Stability of responses

For both primary and alternative hypotheses, ChatGPT-4 shows the highest κ-values ($\kappa = 0.533$ and $\kappa = 0.367$ respectively), closely followed by ChatGPT-3.5 ($\kappa = 0.532$ and $\kappa = 0.337$ respectively) (Table 3). For both versions, consistency was moderate for the primary hypothesis and fair-to-low for alternative hypotheses. For primary and alternative hypotheses, OMP, DDS and DS demonstrated a consistency considerably lower than both versions of ChatGPT (Table 3).

Overall analysis considering correct if any of the given hypotheses matched the diagnosis of the case

Significant differences were observed among the groups regarding the correct diagnoses when any of the provided hypotheses matched the diagnosis (Fig. 2A). The model's intercept (OMP) was significant ($z = 2.79$, $p = 0.0053$), reflecting a baseline correct diagnosis rate of $\beta = 2.0000$ (95% CI:0.5952 to 3.4046). ChatGPT-3.5 showed a non-significant decrease in correct diagnoses compared to the baseline (z=-1.344, $p = 0.1789$), with β=-1.3660 (95% CI: -3.3578 to 0.6257). ChatGPT-4 also did not significantly differ from the baseline (z=-0.538, $p = 0.5905$). DDS and DS demonstrated significant decreases, with β=-3.3238 (z=-3.298, $p = 0.0010$) and β=-3.7280 (z=-3.678, $p = 0.0002$), respectively. Full results table can be found on Appendix 3.

Analysis of the primary hypotheses – if the primary hypothesis matched the diagnosis of the case

Significant differences were observed among the groups regarding the primary hypothesis matching the diagnosis (Fig. 2B). The model's intercept was highly significant ($z = 9.102$, $p < 0.0001$), with a baseline correct diagnosis rate of $\beta = 1.6411$ (95% CI: 1.2878 to 1.9945). ChatGPT-3.5 showed a significant decrease compared to the baseline, with β=-1.8054 (z=-2.378, $p = 0.0174$). On the other hand, ChatGPT-4 did not significantly differ from the baseline (z=-1.520, $p = 0.1285$). Both DDS and DS showed highly significant decreases, with β=-3.0482 (z=-13.525, $p < 0.0001$) and β=-3.4546 (z = -13.549, $p < 0.0001$), respectively. Full results table can be found on Appendix 3.

Analysis of plausible primary hypotheses – if the primary hypotheses did not match the diagnosis of the case but is pertinent as differential diagnosis

For plausible primary hypotheses (Fig. 2C), the model's intercept (OMP) was not significant ($z = 1.583$, $p = 0.1134$), with $\beta = 1.0952$ (95% CI: -0.2607 to 2.4512). ChatGPT-3.5 exhibited a non-significant effect (z=-0.340, $p = 0.7340$), as well as ChatGPT-4 ($z = 0.036$, $p = 0.9710$). However, DDS and DS showed significant decreases, with β=-4.5287 (z=-5.237, $p < 0.0001$) and β=-3.7474 (z=-3.830, $p < 0.0001$), respectively. Full results table can be found on Appendix 3.

Analysis of correct or plausible alternative hypotheses – if any of the alternative hypotheses matched the diagnosis of the case or were pertinent as differential diagnosis

Significant differences were also observed for correct or plausible alternative hypotheses across groups (Fig. 2D). The intercept was not significant (z=-0.683, $p = 0.4949$). ChatGPT-3.5 showed a non-significant increase ($z = 0.313$,

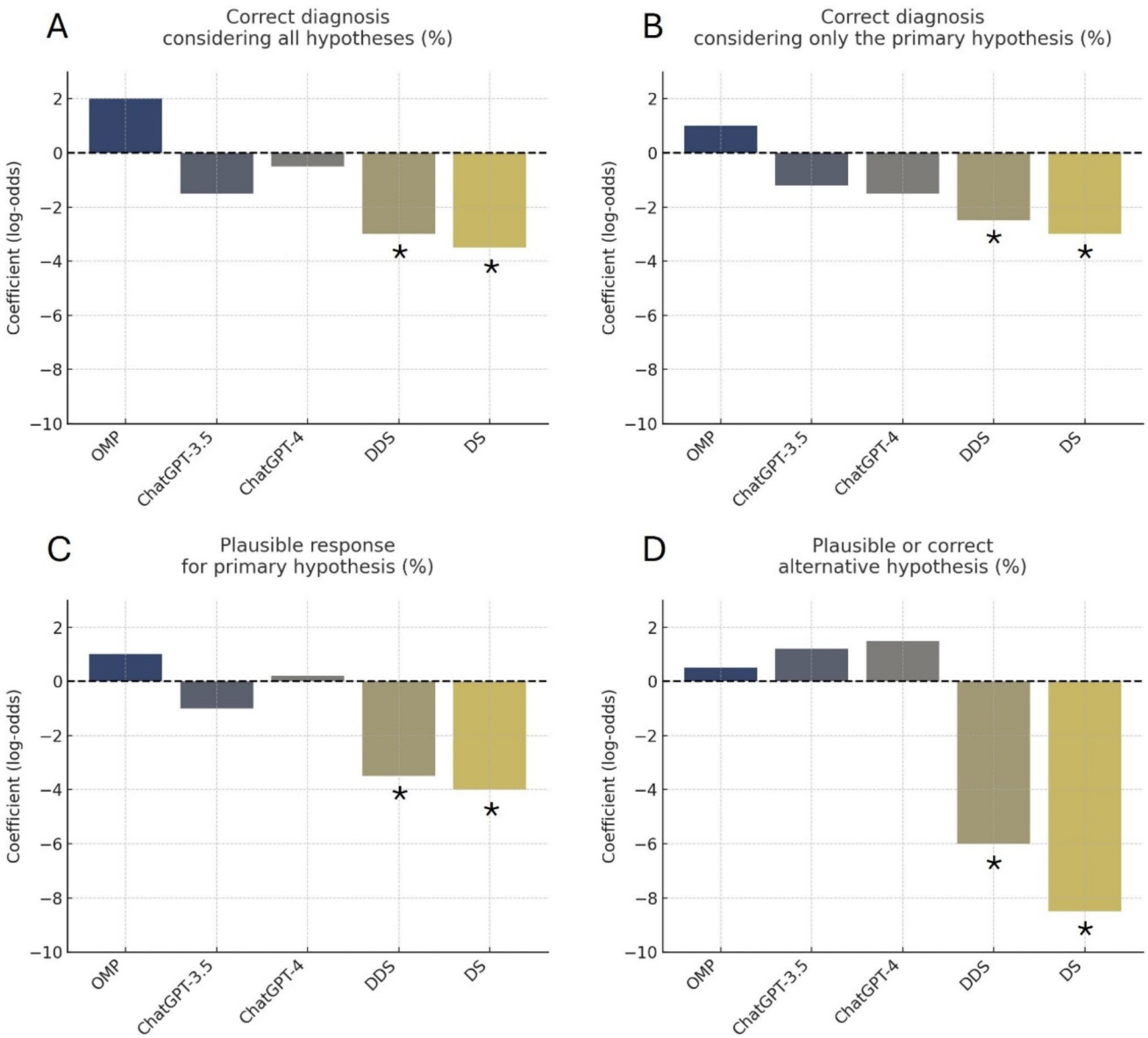

Fig. 2 Coefficient plot demonstrating the log-odds of study groups to provide (**A**) any correct diagnostic hypothesis, (**B**) correct primary diagnostic hypothesis, (**C**) primary plausible hypothesis, and (**D**) any plausible or correct alternative hypothesis. (* $p < 0.05$ compared to the OMP group)

$p = 0.7544$), ChatGPT-4 also showed a non-significant increase compared to the baseline ($z = 0.409$, $p = 0.6826$). Both DDS and DS exhibited highly significant decreases, with β=-7.2485 (z=-5.227, $p < 0.0001$) and β=-8.5107 (z=-5.425, $p < 0.0001$), respectively. Full results table can be found on Appendix 3.

Subgroup analysis

Significant differences were observed for various lesion subgroups across the study groups (Fig. 3). ChatGPT-4 performed comparably to ChatGPT-3.5 for most subgroups, except for maxillary fibro-osseous lesions, where ChatGPT-3.5 showed a slight but significant decline (β=-4.9122, $p = 0.0319$). ChatGPT-3.5 showed no significant differences from the intercept for most subgroups but performed worse than the baseline for maxillary fibro-osseous lesions (β=-17.8067, $p < 0.001$). Both DDS and DS consistently showed significantly lower performance across all lesion subgroups ($p < 0.001$), indicating their overall reduced accuracy. Full results table can be found on Appendix 4.

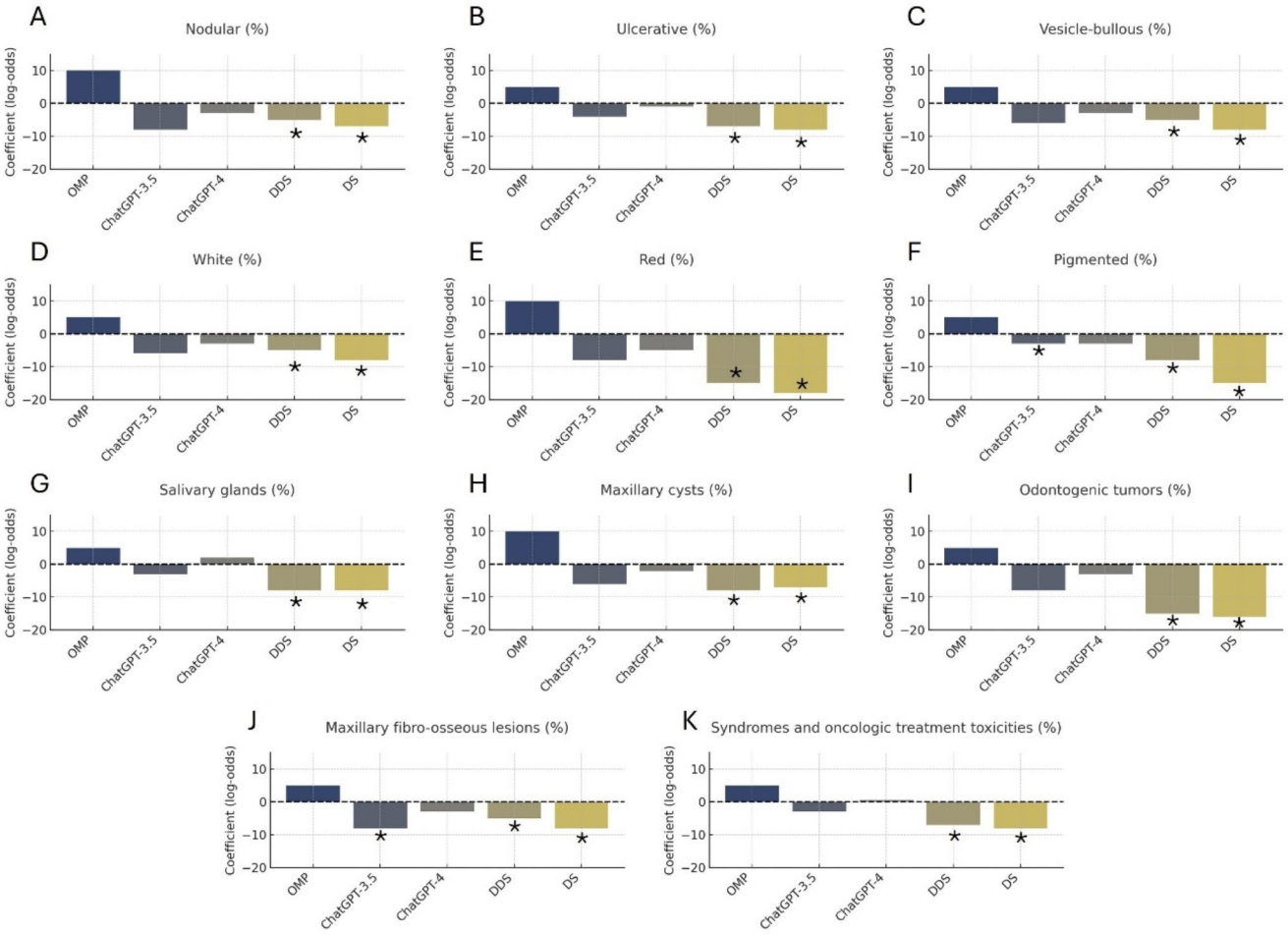

Fig. 3 Coefficient plot of the log-odds of OMP, ChatGPT-3.5, ChatGPT-4, DDS and DS to provide diagnostic hypothesis that match the correct diagnosis of (A) Nodular oral lesions, (B) Ulcerative oral lesions, (C) Vesicle-bullous oral lesions, (D) White oral lesions, (E) Red oral lesions, (F) Pigmented oral lesions, (G) Salivary glands diseases, (H) Maxillary cysts, (I) Odontogenic tumors, (J) Maxillary fibro-osseous lesions, and (K) Syndromes and oncologic treatment toxicities with oral involvement. (* $p < 0.05$ compared to the OMP group)

Impact of case descriptions by dentists

The performance of ChatGPT in providing diagnostic hypotheses for oral and maxillofacial lesions was significantly influenced by the quality of case descriptions provided by dentists. Significant differences were observed between groups considering any hypothesis that matched the correct diagnosis (Fig. 4A). For ChatGPT-3.5, the correct diagnosis rate was significantly reduced when using descriptions from non-specialist dentists (β=-30.341, $p = 0.047$). For ChatGPT-4, a similar significant decrease was observed (β=-46.212, $p = 0.027$). These findings suggest that both versions of ChatGPT perform better with standardized descriptions than with those provided by non-specialist dentists. Full results table can be found on Appendix 5.

Primary and alternative hypothesis analysis

When considering only the primary hypothesis, significant group differences were noted as well (Fig. 4B). For ChatGPT-3.5, there was a decrease in performance with non-specialist descriptions (β=-22.386), although it did not reach significance ($p = 0.243$). For ChatGPT-4, the decrease was more pronounced (β=-29.470) and nearly significant ($p = 0.051$). These results indicate that the primary hypothesis generation is less robust when relying on non-standard descriptions.

For plausible or correct alternative hypotheses (Fig. 4C), significant differences were observed primarily for ChatGPT-3.5, which saw a substantial decline with non-specialist descriptions (β=-53.475, $p = 0.003$). In contrast, ChatGPT-4 showed no significant change in performance (β=-22.133, $p = 0.457$). This suggests that while ChatGPT-3.5 is more

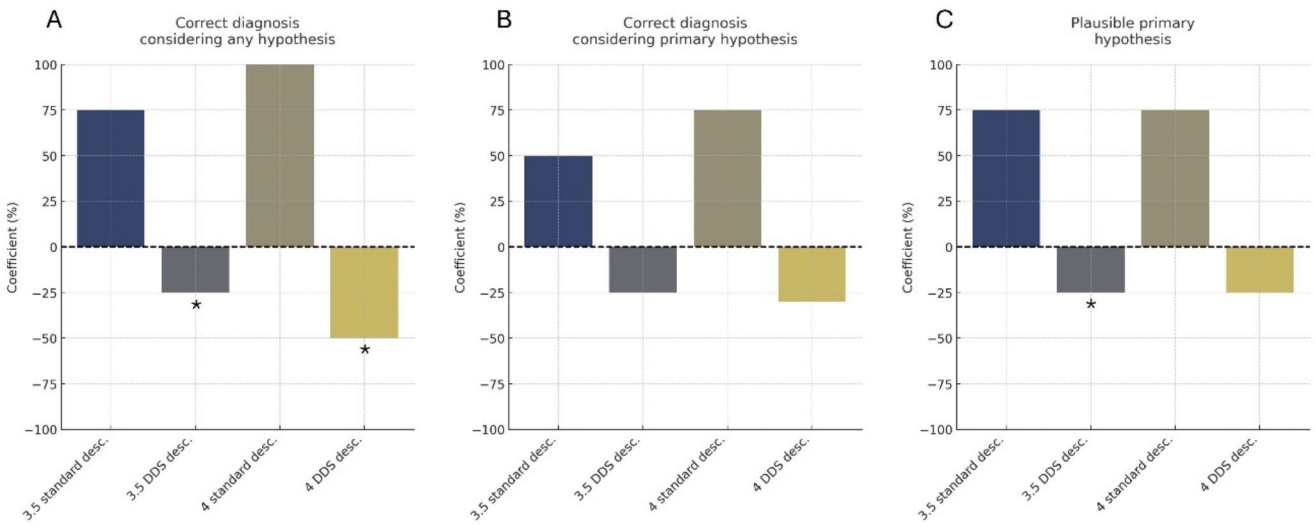

Fig. 4 Coefficient plot of percentages of (**A**) any correct hypothesis, (**B**) primary correct hypothesis, and (**C**) plausible correct hypothesis by ChatGPT-3.5 and ChatGPT-4 when tasked with standard cases' descriptions and cases' descriptions made by general dental practioners. (*$p < 0.05$ compared to the respective standard description)

Table 4 Description of lesions by Dental surgeons

Case information	Complete	Incomplete	Absent	H*	df*	ES	*p*-value*
Demographic	24.13	19.44	29.94	5.028	2	0.107	0.08
Main complaint	20.75	20.72	32.03	7.119	2	0.151	0.02
Medical history	22.19	20.16	31.16	5.837	2	0.124	0.05
Clinical description	14.91[a]	39.22[b]	19.38[a]	28.25	2	0.601	< 0.001
Radiologic description	17.56[a]	37.88[b]	18.06[a]	25.07	2	0.533	< 0.001

*Kruskal-wallis test. Data is presented as Mean Rank

df, Degree of Freedom; ES, Effect Size (*epsilon*^{2})

[a, b,c] Different letters indicate statistical significance on multiple comparison test (Dunn's corrected by Bonferroni's method)

Table 5 Completeness of clinical descriptions by Dental surgeons

Description	Demographic	Main Complaint	Medical history	Clinical description	Radiologic description	H*	df*	ES	*p*-value*
Complete	44.56	44.78	45.03	34.28	33.84	4.697	4	0.059	0.31
Incomplete	24.47[a]	29.22[a]	26.81[a]	61.06[b]	60.94[b]	42.82	4	0.542	< 0.001
Absent	46.47[a, b]	50.13[a]	48.44[a]	31.63[a, b]	25.84[b]	15.05	4	0.191	0.004

*Kruskal-wallis test. Data is presented as Mean Rank

df, Degree of Freedom; ES, Effect Size (*epsilon*^{2})

[a, b,c] Different letters indicate statistical significance on multiple comparison test (Dunn's corrected by Bonferroni's method)

susceptible to variability in case descriptions, ChatGPT-4 maintains more consistent performance. Full results table can be found on Appendix 5.

Quality of case descriptions

Analysis of case descriptions by dentists showed that most participants significantly provided incomplete clinical (H [df = 4, n = 48] = 28.25, $p < 0.001$) and radiologic (H [df = 4, n = 48] = 25.07, $p < 0.001$) descriptions (Table 4). Although significant differences were found among complete, incomplete, and absent descriptions for the main complaint using the Kruskal-Wallis test, Dunn's multiple comparisons did not indicate significant differences between these groups (Table 4). Detailed analysis showed that clinical and radiologic descriptions were more likely to be incomplete, and main complaint and medical history were more frequently absent than radiologic descriptions (Table 5).

Discussion

The results of the present study provide significant insights into the performance of ChatGPT in the differential diagnosis of oral and maxillofacial lesions in comparison with human specialists in this field, general dental practioners and dental students. The primary findings obtained demonstrated that the performance of ChatGPT-4 is, in general, comparable to that of OMPs to provide correct diagnosis for oral and maxillofacial lesions, indicating that this AI has the potential to reach the same level of diagnostic accuracy as specialists, what may have important implications for clinical practice.

The mean of overall correct diagnosis provided by ChatGPT-3.5 in this study (64.86%) is slightly lower than the rate of correct diagnosis obtained by ChatGPT-3.5 for ear, nose, and throat (ENT) pathologies (70.8%) in another study [16]. ChatGPT may have different levels of accuracy across healthcare specialties and groups of diseases. Lechien et al. [15] observed a rate of 62% correct primary diagnosis provided by ChatGPT-4 for otolaryngology – head and neck surgery cases, varying between 38% (laryngology cases) and 86% (head and neck cases), although no statistical difference was observed. Within oral and maxillofacial lesions in the present study, the mean of correct diagnosis for ChatGPT-4 was 80.18%, varying between 72.9% and 91.8%. A possible explanation for the higher precision observed in this study, is that most cases of oral diseases have an evident clinical lesion which can be observed and described in detail, while for some cases of otolaryngology and head and neck surgery only subjective information such as symptoms may be available for description. This study also demonstrated that ChatGPT-3.5 performed significantly worse than OMP for pigmented and bone lesions, while there was no statistically significant difference between ChatGPT-4 and OMPs for all subgroups, indicating that ChatGPT-3.5 has limitations for some groups of lesions.

ChatGPT-4 also demonstrated an accuracy comparable to OMPs to provide a correct primary diagnosis, what highlights its efficacy to define diagnostic hypotheses by probability. ChatGPT-3.5, ChatGPT-4 and OMPs were similarly likely to provide a plausible primary hypothesis, demonstrating that this AI tool is capable of 'thinking' broadly and provide valuable hypotheses. Moreover, we found that ChatGPT-4 presented a mean of 76.7% of plausible or correct alternative hypotheses, outperforming OMPs (34.9%). Although this difference was not statistically significant on GLMM analysis, this finding suggests that ChatGPT-4 carry the potential to serve as an auxiliary tool for specialists in cases where multiple differential diagnoses must be considered. An example is the perspective of integrating AI tools like ChatGPT with teledentistry for addressing complex cases [23]. Deeper investigations on this aspect must be conducted to better understand the extent of the applicability of ChatGPT for primary and specialized care.

AI tools like ChatGPT are heavily dependent on the quality of input data. A significant deterioration of the accuracy of both ChatGPT versions was observed when imputed with cases described by DDS-II, what can be mainly attributed to deficiencies in reporting patients' main complain and clinical and imaginologic descriptions of lesions. This analysis underscores the importance of complete documentation of cases and description to AI tools and the need for continued education and training of practioners to collect and provide accurate clinical data. This problem is ill-explored but chronic within Oral Medicine and Pathology since most referral letters from general practioners are of low-quality due to lack of essential information [24]. Oral Medicine specialists indicated that general dental practioners usually provide poor details of oral lesions clinical features and do not report patients' medical history on referrals [25]. The advent of AI tools in healthcare may shed light on this professionals' deficiency in dental and other health areas, since the implementation of AIs into clinical practice will depend on information collected and provided by professionals. With this, our results also indicate a possible challenge when future studies progress into investigating the application of Large Language Models (LLM) as diagnostic auxiliar in primary care settings.

ChatGPT-4 demonstrated better performance to diagnose or suggest plausible primary hypothesis for oral and maxillofacial lesions than DDSs and DSs and overcame OMPs in providing plausible alternative hypothesis. On the other hand, its precursor ChatGPT-3.5, exhibited significantly lower accuracy to provide correct primary hypothesis, and correct hypothesis for some subgroups of lesions. This finding is remarkable since it indicates the potential of ChatGPT-4 to support general practioners and students enhancing their diagnostic capabilities, what may have impact on clinical outcome for patients. Nevertheless, ChatGPT-3.5 has some limitations of which professionals and students should be aware since the free availability of this version may encourage many to opt for this version for clinical and educational assistance.

The inter-rater reliability of outputs across ChatGPT runs (15 for each version) only reached a moderate consistency for both versions for the primary hypotheses and low consistency for alternative hypotheses, indicating a limited consistency. Lack of consistency was also observed for ChatGPT-4 to propose additional examinations and therapeutic options for head and neck cancer [19]. As suggested by Lechien et al. [15], the knowledge and fine tuning of hyperparameters of LLMs may overcome this limitation. Nevertheless, consistency for both ChatGPT were substantially better than for

OMPs, DDSs, and DSs. The low consistency observed for OMP was not expected but, as a secondary result, exposes a potential problem occurring in the population studied. Thus, we recommend deeper investigations on the uniformity of clinical practice of specialists in oral medicine. Although there is space and need for improvement, ChatGPT may influence on the standardization of education and patient care, since an important discordance is evident among professionals.

The primary strength of this research is its originality because, to the best of our knowledge, this is the first study assessing the performance of two versions of ChatGPT in providing differential diagnoses for a wide variety of real cases of oral and maxillofacial lesions. This provides evidence not only to define its performance level, but to indicate that the inclusion of AI tolls into clinical practice may improve the diagnostic accuracy of specialists, general practioners and dental students. Two previous studies have investigated the accuracy of ChatGPT-3.5 to give diagnostic hypotheses for oral medicine cases [26, 27] but limited the investigation to the earlier version of the tool, which we demonstrated hereby to not be beneficial for clinical practice of general dentists and students. Furthermore, none of these studies compared the diagnostic performance of ChatGPT with the performance of specialists, what allowed us to demonstrate that ChatGPT-4 can reach the same accuracy as specialists and has the potential to overcome specialists in alternative hypotheses. Another strength is the inclusion of an analysis addressing the impact of cases descriptions made by general dentists on the performance of ChatGPT, which had not been assessed elsewhere.

The main limitation of this study is the low number of cases ($n=37$). Exapanding the number of cases, including a retrospective casuistic to better represent the routine of diagnostic process in single or multiple centers may help us to understand the applicability of LLMs on clinical routine. An additional limitation is that this study focused solely on the determination of differential diagnoses for the cases, what does not represent the full propedeutics work required to reach a final diagnosis. Future research must include the assessment of the performance of ChatGPT to recommend conducts for diagnosis of these lesions. Another limitation is that in order to maintain uniformity on the clinical information given to ChatGPT versions and human participants, the human groups of the first part of the study did not have access to clinical and radiologic images of the cases, what does not reflect a real-life situation, and may put ChatGPT versions on advantage since this AI tool is a language model. However, this research setup still gave us the advantage to observe and discuss how important high quality description of cases is important, since DDS-II group having access to clinical and radiologic images failed to obtain accurate hypothesis from ChatGPT versions due to low quality description of cases.

This study presents early yet significant results on the potential of ChatGPT as an auxiliary tool for the diagnosis of oral lesion. Some issues however remain to be explored in depth on future research. First, we recommend that future research avoid selected cases and include a retrospective bigger casuistic in order to expose the ChatGPT to routinely real clinical cases. With this, more consistent responses may be obtained regarding the potential of this tool as an auxiliary tool for daily clinical practice. Another issue to be better studied is how specialists may benefit from LLMs like ChatGPT, since we observed that although ChatGPT-4 have an accuracy comparable to OMP, the AI was more efficient to provide plausible alternative hypothesis, what could help specialists on challenging cases. These results must be confirmed and tested on future exploratory and prospective analyses. Furthermore, as AI tools gain space within healthcare, issues with the documentation and reporting of clinical data by professionals must be assessed with proposals to improve it, since AI tools are completely dependent on the quality and preciseness of imputed data.

Conclusion

ChatGPT-4 showed remarkable potential to provide pertinent diagnostic hypotheses for oral diseases, being comparable to specialists in Oral Medicine. Although ChatGPT-3.5 surpasses the diagnostic accuracy of DDSs and DSs, it remains inferior to both ChatGPT-4 and OMPs, especially considering specific groups of oral and maxillofacial diseases. The quality of input data significantly influenced ChatGPT's performance, underscoring the necessity for ongoing professional education and training to ensure comprehensive case documentation and precise descriptions with pertinent data. Both ChatGPT-3.5 and −4 have moderate consistency to provide differential diagnosis, what, currently, represents a disadvantage for clinical application.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00784-024-05939-1.

Acknowledgements S.T. acknowledges to the São Paulo State Research Foundation (FAPESP) for the postdoctoral fellowship (grant number 23/11402-4).

Author contributions Saygo Tomo, Design, data acquisition, data analysis and interpretation, drafting, editing final version, final approval. Jérôme R. Lechien, Data analysis and interpretation, drafting, final approval. Hugo S. Bueno, Design, data acquisition, final approval. Daniela F. Cantieri-Debortoli, Design, data acquisition, final approval. Luciana E. Simonato, Design, project administration, data acquisition, data analysis and interpretation, final approval.

Funding None to declare.

Data availability Data of the present study is available upon reasonable request to the corresponding author.

Declarations

Competing interests The authors declare no competing interests.

References

1. Haug CJ, Drazen JM (2023) Artificial intelligence and machine learning in clinical medicine. N Engl J Med 388:1201–1208. https://doi.org/10.1056/NEJMra2302038
2. Lee P, Bubeck S, Petro J (2023) Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 388:1233–1239. https://doi.org/10.1056/NEJMra2302039
3. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK (2023) ChatGPT in dentistry: a comprehensive review. Cureus 15:e38317. https://doi.org/10.7759/cureus.38317
4. Vimalraj S, Sekaran S (2023) ChatGPT: empowering dentistry with future possibilities. Oral Oncol 144:106496. https://doi.org/10.1016/j.oraloncology.2023.106496
5. Cunha JL (2024) Comment on 'ChatGPT-4 as auxiliary tool in the temporomandibular disorders diagnostic: an opinion'. Oral Surg 7:296–297. https://doi.org/10.1111/ors.12876
6. Cunha JL, Alves PM, Nonaka CF (2023) ChatGPT in oral medicine: striking a balance between technological advancement and clinical expertise. Oral Dis Early view. https://doi.org/10.1111/odi.14823
7. Yang Y, Ngai EW, Wang L (2024) Resistance to artificial intelligence in health care: literature review, conceptual framework, and research agenda. Inf Man 61:103961. https://doi.org/10.1016/j.im.2024.103961
8. Diniz-Freitas M, Fernández-Sanromán J, Alonso-Del-Hoyo JR, Blanco-Carrión A, López-Cedrún JL (2023) Application of ChatGPT in oral and maxillofacial surgery. J Oral Maxillofac Surg 81:1277–1285. https://doi.org/10.1016/j.joms.2023.04.021
9. Alawi F (2023) Differential diagnosis in the future. Oral Surg Oral Med Oral Pathol Oral Radiol 136:119–121. https://doi.org/10.1016/j.oooo.2023.05.003
10. Vaira LA, Lechien JR, Abbate V et al (2023) Accuracy of ChatGPT-Generated information on Head and Neck and Oromaxillofacial surgery: a Multicenter Collaborative Analysis. Otolaryngol Head Neck Surg. https://doi.org/10.1002/ohn.489
11. de Souza LL, Lopes MA, Santos-Silva AR, Vargas PA (2024) The potential of ChatGPT in oral medicine: a new era of patient care? Oral Surg Oral Med Oral Pathol Oral Radiol 137:1–8. https://doi.org/10.1016/j.oooo.2023.09.010
12. Lang A, Smith R, Johnson T (2024) ChatGPT's role in modern healthcare education. Med Educ 58:45–51. https://doi.org/10.1111/medu.14782
13. Bilika T, Rodriguez A, Svensson P (2023) Artificial intelligence in dental research: a narrative review. J Dent Res 102:1442–1450. https://doi.org/10.1177/00220345231159928
14. Li DJ, Kao YC, Tsai SJ, Bai YM, Yeh TC, Chu CS, Hsu CW, Cheng SW, Hsu TW, Liang CS, Su KP (2024) Comparing the performance of ChatGPT GPT-4, Bard, and Llama-2 in the Taiwan Psychiatric Licensing Examination and in differential diagnosis with multi-center psychiatrists. Psych Clin Neurosc. Article in. https://doi.org/10.1111/pcn.13656
15. Lechien JR, Naunheim MR, Maniaci A, Radulesco T, Saibene AM, Chiesa-Estomba CM, Vaira LA (2024) Performance and consistency of ChatGPT-4 Versus otolaryngologists: a clinical Case Series. Otolaryngol Head Neck Surg Article in press. https://doi.org/10.1002/ohn.759
16. Makhoul M, Melkane AE, El Khoury P, El Hadi C, Matar N (2024) A cross-sectional comparative study: ChatGPT 3.5 versus diverse levels of medical experts in the diagnosis of ENT diseases. Eur Arch Otorhinolaryngol 281:1234–1243. https://doi.org/10.1007/s00405-024-08509-z
17. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T (2023) Diagnostic accuracy of differential-diagnosis lists generated by generative pre-trained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 20:3378. https://doi.org/10.3390/ijerph20043378
18. Chiesa-Estomba CM, Lechien JR, Vaira LA et al (2024) Exploring the potential of Chat-GPT as a supportive tool for sialendoscopy clinical decision making and patient information support. Eur Arch Otorhinolaryngol 281:2081–2086. https://doi.org/10.1007/s00405-023-08104-8
19. Lechien JR, Chiesa-Estomba CM, Baudouin R, Hans S (2024) Accuracy of ChatGPT in head and neck oncological board decisions: preliminary findings. Eur Arch Otorhinolaryngol 281:2105–2114. https://doi.org/10.1007/s00405-023-08326-w
20. Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA (2024) Validity and reliability of an instrument evaluating the performance of an intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI). Eur Arch Otorhinolaryngol 281:2063–2079. https://doi.org/10.1007/s00405-023-08219-y
21. Seabold S, Perktold J (2010) Statsmodels: Econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference 92–96. https://doi.org/10.25080/Majora-92bf1922-011
22. Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9(3):90–95. https://doi.org/10.1109/MCSE.2007.55
23. Santana LADM, Floresta LG, Alves EVM et al (2024) Advancing oral cancer diagnosis in Brazil: integrating artificial intelligence with teledentistry for enhanced patient outcomes. Oral Oncol 151:106741. https://doi.org/10.1016/j.oraloncology.2024.106741
24. Rodrigues CRD, Fernandes PM, Santos-Silva AR, Vargas PA, Lopes MA (2021) Evaluation of the quality of referral letters: experience of a Brazilian oral medicine service. Braz Oral Res 35. https://doi.org/10.1590/1807-3107bor-2021.vol35.0037
25. Guan G, Lau J, Yew V et al (2020) Referrals by general dental practitioners and medical practitioners to oral medicine specialists in New Zealand: a study to develop protocol guidelines. Oral Surg Oral Med Oral Pathol Oral Radiol 130:43–51. https://doi.org/10.1016/j.oooo.2020.03.050
26. Albagieh H, Alzeer ZO, Alasmari ON et al (2024) Comparing Artificial Intelligence and Senior Residents in Oral Lesion Diagnosis: A Comparative Study. *Cureus* 2024 16: e51584. https://doi.org/10.7759/cureus.51584
27. Uranbey Ö, Özbey F, Kaygısız Ö, Ayrancı F (2024) Assessing ChatGPT's Diagnostic Accuracy and therapeutic strategies in oral pathologies: a cross-sectional study. Cureus 16:e58607. https://doi.org/10.7759/cureus.58607