Journal of Pain Research

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Evaluating Large Language Models for Burning Mouth Syndrome Diagnosis

Takayuki Suga[1], Osamu Uehara[2], Yoshihiro Abiko[3], Akira Toyofuku[1]

[1]Department of Psychosomatic Dentistry, Graduate School of Medical and Dental Sciences, Institute of Science Tokyo, Tokyo, Japan; [2]Division of Disease Control and Molecular Epidemiology, Department of Oral Growth and Development, School of Dentistry, Health Sciences University of Hokkaido, Ishikari-Tobetsu, Hokkaido, Japan; [3]Division of Oral Medicine and Pathology, Department of Human Biology and Pathophysiology, School of Dentistry, Health Sciences University of Hokkaido, Ishikari-Tobetsu, Hokkaido, Japan

Correspondence: Takayuki Suga, Email suga.ompm@tmd.ac.jp

Introduction: Large language models have been proposed as diagnostic aids across various medical fields, including dentistry. Burning mouth syndrome, characterized by burning sensations in the oral cavity without identifiable cause, poses diagnostic challenges. This study explores the diagnostic accuracy of large language models in identifying burning mouth syndrome, hypothesizing potential limitations.

Materials and Methods: Clinical vignettes of 100 synthesized burning mouth syndrome cases were evaluated using three large language models (ChatGPT-4o, Gemini Advanced 1.5 Pro, and Claude 3.5 Sonnet). Each vignette included patient demographics, symptoms, and medical history. Large language models were prompted to provide a primary diagnosis, differential diagnoses, and their reasoning. Accuracy was determined by comparing their responses with expert evaluations.

Results: ChatGPT and Claude achieved an accuracy rate of 99%, while Gemini's accuracy was 89% ($p < 0.001$). Misdiagnoses included Persistent Idiopathic Facial Pain and combined diagnoses with inappropriate conditions. Differences were also observed in reasoning patterns and additional data requests across the large language models.

Discussion: Despite high overall accuracy, the models exhibited variations in reasoning approaches and occasional errors, underscoring the importance of clinician oversight. Limitations include the synthesized nature of vignettes, potential over-reliance on exclusionary criteria, and challenges in differentiating overlapping disorders.

Conclusion: Large language models demonstrate strong potential as supplementary diagnostic tools for burning mouth syndrome, especially in settings lacking specialist expertise. However, their reliability depends on thorough patient assessment and expert verification. Integrating large language models into routine diagnostics could enhance early detection and management, ultimately improving clinical decision-making for dentists and specialists alike.

Keywords: burning mouth syndrome, large language models, diagnostic accuracy, dentistry, artificial intelligence

---

Introduction

Burning Mouth Syndrome (BMS) is a disorder characterized by persistent burning sensations or pain in the oral cavity with no identifiable cause, particularly affecting middle-aged and older women.[1] Patients frequently report symptoms such as dry mouth sensation and taste disturbances, yet its exact etiology remains unclear.[1] Dysfunction in the pain processing system and possible brain involvement have been suggested, but definitive causative factors have not been established.[1] The diagnosis of BMS largely relies on ruling out other conditions, guided by exclusionary criteria like those outlined in the International Classification of Headache Disorders, 3rd edition (ICHD-3).[2] This process demands considerable clinical judgment and poses significant challenges, partly due to its overlapping symptomatology with other oral and maxillofacial pain disorders, as well as the absence of standardized, universally accepted biomarkers.

Current diagnostic methods for BMS underscore several critical gaps. First, the reliance on exclusionary criteria can lead to delayed or missed diagnoses when clinicians lack specialized training.[3] Second, patients' nuanced descriptions of pain or burning sensations often vary widely, further complicating accurate diagnosis.[4] Finally, many general dental

Received: 2 December 2024
Accepted: 26 February 2025
Published: 19 March 2025

© 2025 Suga et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v3.0) License (http://creativecommons.org/licenses/by-nc/3.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

practitioners may be unfamiliar with BMS or other medically unexplained oral symptoms (MUOS), resulting in diagnostic uncertainty and sometimes unnecessary treatments.[5] These gaps highlight the need for more efficient, accessible, and standardized diagnostic approaches that can assist clinicians in identifying BMS, while also addressing its strong association with psychiatric disorders and chronic pain.[3]

Generative Artificial Intelligence (AI), an advanced AI technology trained on vast amounts of data to generate human-like responses, has gained substantial attention in medical diagnostics. Among these AI systems, large language models (LLMs) are designed to produce fluent and coherent sentences, translations, and answers, functioning as highly reliable conversational agents—often referred to as "chatbots".[6] LLMs have been investigated for diagnostic support in multiple medical and dental fields, ranging from internal medicine and ophthalmology to oral pathology and radiology.[7–11] They analyze symptoms and patient complaints to propose possible diagnoses, guide treatment strategies, and even enhance patient–provider communication.[12,13] Furthermore, LLMs offer significant benefits in medical and dental education through simulations and case studies, aiding skill development and potentially improving healthcare delivery.[6,14] By leveraging LLMs alongside big data, new insights into treatments can emerge, promoting precision medicine and addressing healthcare disparities.[15] However, while research has explored the application of LLMs in diagnosing psychiatric and chronic pain conditions, no prior studies have investigated their role in diagnosing MUOS, including BMS.[16,17]

Against this backdrop, the present study aims to address a crucial gap in current diagnostic methods for BMS by evaluating the potential of LLMs to improve diagnostic accuracy. We posit that BMS, with its intricate and subjective symptom descriptions, may present significant challenges for LLM-based diagnostics, particularly when distinguishing between overlapping oral and maxillofacial pain syndromes. Building on broader efforts in AI-driven medical diagnostics, our research specifically focuses on whether LLMs can overcome existing limitations—such as the reliance on exclusionary criteria, subtle variations in patient-reported pain, and a lack of standardized clinical pathways—to provide reliable, supportive tools for clinicians. Therefore, our hypothesis is that diagnosing BMS will be challenging for LLMs, yet understanding the nature of these challenges could pave the way for more refined, AI-assisted diagnostic strategies. By systematically examining LLM diagnostic performance on BMS, we seek to inform both the dental community and AI researchers on how these models might be optimized or integrated into clinical practice for managing complex, medically unexplained oral conditions.

Materials and Methods

Vignettes

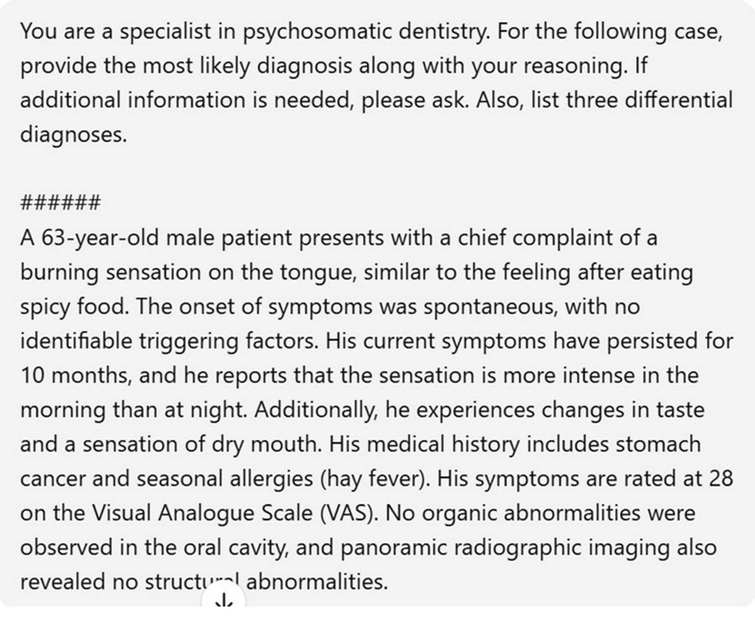
In this study, clinical vignettes were employed, following the approach of other diagnostic studies using similar LLMs.[17,18] To reflect the diverse symptoms and patient backgrounds associated with BMS, 100 cases were constructed from actual consecutive patient data, with details such as age, symptoms, and past medical history carefully modified to preserve patient privacy and maintain the key clinical features of BMS. Each vignette included information on the patient's age, gender, chief complaint, history of present illness, current symptoms, medical history (including psychiatric history), factors triggering symptom onset, and the corresponding visual analog scale (VAS; maximum 100) scores.

The clinical vignettes were reviewed by a dentist with over 30 years of experience in treating MUOS in dentistry to ensure their plausibility as BMS cases. Given that, in our clinical experience, patients often report a sensation of "dry mouth" despite showing no abnormalities on Saxon tests or appearing to have moist oral cavities, the expressions "dry mouth" and "xerostomia" were deliberately avoided in favor of "dry mouth sensation".

Additionally, due to the challenges of synthesizing photographic evidence, the vignettes were designed to state explicitly that no organic abnormalities were observed during oral examinations, and panoramic radiographic imaging showed no structural issues. This was reflected in all prompts as the statement:

> No organic abnormalities were observed in the oral cavity, and panoramic radiographic imaging also revealed no structural abnormalities.

Details of all vignette and diagnostic results from the three LLMs are provided in the Supplementary Material 1. Figure 1 presents an example of a prompt containing a vignette scenario.



You are a specialist in psychosomatic dentistry. For the following case, provide the most likely diagnosis along with your reasoning. If additional information is needed, please ask. Also, list three differential diagnoses.

######
A 63-year-old male patient presents with a chief complaint of a burning sensation on the tongue, similar to the feeling after eating spicy food. The onset of symptoms was spontaneous, with no identifiable triggering factors. His current symptoms have persisted for 10 months, and he reports that the sensation is more intense in the morning than at night. Additionally, he experiences changes in taste and a sensation of dry mouth. His medical history includes stomach cancer and seasonal allergies (hay fever). His symptoms are rated at 28 on the Visual Analogue Scale (VAS). No organic abnormalities were observed in the oral cavity, and panoramic radiographic imaging also revealed no structural abnormalities.

Figure 1 An example of a prompt containing a vignette scenario.

Diagnosis by LLMs

In this study, we analyzed the diagnostic accuracy and reasoning process of three LLMs: ChatGPT-4o (OpenAI), Gemini Advanced 1.5 Pro (Google), and Claude 3.5 Sonnet (Anthropic). The prompts were submitted on November 15 and 16, 2024, to ensure that all LLMs were evaluated on stable, up-to-date versions and to minimize any confounding effects from incremental model updates, with the following content:

> You are a specialist in psychosomatic dentistry. For the following case, provide the most likely diagnosis along with your reasoning. If additional information is needed, please ask. Also, list three differential diagnoses.

Based on this prompt, we requested each LLMs to provide a diagnosis, the reasoning behind it, and three differential diagnoses.

The study was designed around a specific scenario, simulating typical yet diverse symptoms and patient backgrounds of BMS. Notably, we excluded the distinction between primary and secondary BMS from the evaluation of diagnostic accuracy. The key reasons are as follows:[19–23]

Diagnostic Ambiguity

Because primary BMS is idiopathic and primarily diagnosed by excluding other conditions, its diagnostic criteria can be inherently vague. This ambiguity often leads to inconsistent interpretations and potential misdiagnosis across different clinicians.

Potential Multifactorial Nature

BMS is considered to result from a complex interplay of neurological, psychological, and biological factors. Hence, a strict binary classification of "primary" versus "secondary" may not fully capture the diversity of clinical presentations or underlying etiologies.

Research Limitations and Clinical Utility

Current evidence is insufficient to establish whether distinguishing between primary and secondary BMS significantly benefits clinical practice. The extent to which this classification informs treatment decisions or patient management remains debatable, especially given that therapy often requires a multidisciplinary approach addressing both physical and psychological components.

For these reasons, we do not use this classification in clinical practice, nor did we incorporate it in LLM-based diagnostics.

The diagnostic evaluation criteria are as follows:

1. The diagnosis is considered correct only if the most likely diagnosis provided by the LLMs is BMS. The level of certainty in the diagnosis, such as "definitely" or "probably", is not evaluated.
2. A diagnosis of "Somatic Symptom Disorder" is deemed incorrect, as it reflects a psychiatric perspective and does not align with the viewpoint of specialists in oral medicine.
3. Gemini occasionally provided mixed diagnoses involving two conditions. If BMS was included and the other diagnosis was appropriate given the scenario, the answer was considered correct.
4. Failure to provide a response was considered incorrect.

Figure 2 illustrates an example of LLM responses to vignette scenarios.

Error Analysis

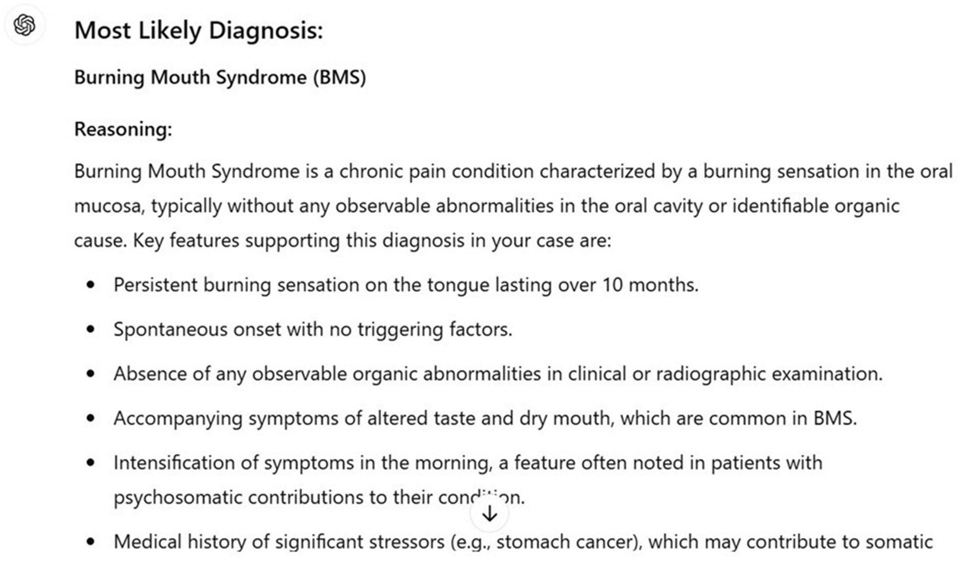
For responses identified as incorrect by the LLM, we classified them into the following categories:[24]

1. Logical Fallacy: The response exhibits sequential reasoning but does not adequately address the question.
2. Informational Fallacy: The response applies logical reasoning but neglects to incorporate a critical piece of information from the question prompt, leading to an incomplete or incorrect answer.
3. Explicit Fallacy: The response lacks logical coherence and fails to utilize the information provided in the question prompt to form an appropriate answer.

As illustrated in Figure 3, we thoroughly evaluated each LLM's output concerning the misdiagnosis case, examining both the logical consistency and the completeness of the prompts and training data. We then classified these outputs into the three categories described above.

Statistics

Statistical analyses were conducted using IBM SPSS Statistics version 29 (Armonk, NY). A chi-square test was conducted to evaluate differences in accuracy rates. When a chi-square test was conducted, we calculated Cramér's



Most Likely Diagnosis:

Burning Mouth Syndrome (BMS)

Reasoning:

Burning Mouth Syndrome is a chronic pain condition characterized by a burning sensation in the oral mucosa, typically without any observable abnormalities in the oral cavity or identifiable organic cause. Key features supporting this diagnosis in your case are:

- Persistent burning sensation on the tongue lasting over 10 months.
- Spontaneous onset with no triggering factors.
- Absence of any observable organic abnormalities in clinical or radiographic examination.
- Accompanying symptoms of altered taste and dry mouth, which are common in BMS.
- Intensification of symptoms in the morning, a feature often noted in patients with psychosomatic contributions to their condition.
- Medical history of significant stressors (e.g., stomach cancer), which may contribute to somatic

Figure 2 An example of Large language model (LLM) responses to vignette scenario.

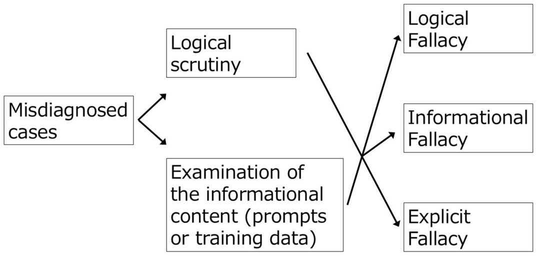


Figure 3 Error Analysis Flowchart: First, examine misdiagnosis cases with respect to logical consistency and sufficiency of information. Then, classify them into three categories: Logical Fallacy, Informational Fallacy, and Explicit Fallacy.

V to determine the effect sizes. In this study, we set the significance level at 0.05, taking into account (1) the sample size of 100, (2) the fact that previous similar studies have used $p < 0.05$ as their significance threshold, and (3) its customary use as a standard in medical research.[25] Patient data were expressed as mean ± standard deviation (SD).

In accordance with the Declaration of Helsinki and following review by the Tokyo Medical and Dental University Hospital Ethical Committee (approval number: D2013-005-04), the patient who initially served as the basis for the clinical vignettes provided written, comprehensive research consent. Although the clinical vignettes used in this study were originally derived from actual patient information, details such as age, symptoms, and medical history were modified. Consequently, because no genuine patient data were employed, informed consent was deemed unnecessary, and, given the nature of the research, ethics committee review was not required.

Result

Patient Demographic From Vignette Scenarios

Table 1 presents the patient demographics from the vignette scenarios. The average age was 59.7 ± 13.5 years, and 10% of the patients were male. The VAS score was 57.9 ± 26.3. The age and VAS values were similar to those of actual patients with BMS. However, the proportion of male patients was slightly lower, as previous studies suggest a male ratio of 20–30%.[26]

Accuracy Rates of Three LLMs

Table 2 presents the number of correct and incorrect answers for three LLMs. ChatGPT and Claude achieved an accuracy rate of 99%, while Gemini's accuracy rate was 89%. A chi-square test revealed a P-value < 0.001. Cramer's V was approximately 0.23, which, for a 2×3 contingency table, is indicative of a medium level of association. Residual analysis indicated that Gemini's accuracy rate was significantly lower.

Incorrect Diagnoses

Table 3 summarizes the incorrect diagnoses made by the three LLMs. ChatGPT and Claude each misdiagnosed one case as Persistent Idiopathic Facial Pain (PIFP). Gemini made four diagnostic errors, combined BMS with other inappropriate diagnoses in six cases, and failed to explicitly provide a diagnosis in one case.

Table 1 Patient Demographic in Scenario Vignette

Age (yr)	59.7 ±13.5
Gender (male/ female)	10/90
Visual Analogue Scale of symptoms	57.9 ±26.3

Table 2 Number of Correct and Incorrect Answers of Three Large Language Models

	ChatGPT4o	Gemini Advanced 1.5 Pro	Claude 3.5 Sonnet
Correct	99	89	99
Incorrect	1	11	1

Table 3 Summary of Incorrect Diagnoses

ChatGPT4o		Gemini Advanced 1.5 Pro		Claude 3.5 Sonnet	
Persistent Idiopathic Facial Pain (PIFP)	1	Somatoform disorder	1	Persistent Idiopathic Facial Pain (PIFP)	1
		Drug-induced burning mouth sensation	1		
		BMS and neuropathic pain	1		
		BMS and Somatic Symptom Disorder	2		
		BMS and xerostomia	3		
		Post-Herpetic Neuralgia (PHN) with oral manifestations	1		
		Persistent Idiopathic Facial Pain (PIFP)	1		
		No single definitive answer	1		

Reasons for the Diagnosis of BMS

Table 4 summarizes the reasoning for diagnosing BMS as aggregated by three LLMs. Only reasoning with a count of four or more is included in the table All three LLMs identified "Absence of Organic Abnormalities", "Primary symptoms of BMS (persistent burning pain, especially on the tongue, feeling of dry mouth, etc)", and "Chronic Nature of

Table 4 Summary of Diagnosis Reasoning

ChatGPT4o		Gemini Advanced 1.5 Pro		Claude 3.5 Sonnet	
Absence of Organic Abnormalities	91	Primary symptoms of BMS	87	Absence of Organic Abnormalities	95
Primary symptoms of BMS	89	Absence of Organic Abnormalities	85	Primary symptoms of BMS	91
Chronic Nature of Symptoms	79	Chronic Nature of Symptoms	58	Chronic Nature of Symptoms	76
Psychological Factors	73	Diurnal pattern (worse in evening)	55	Middle-aged and older postmenopausal female	73
Diurnal pattern (worse in evening)	54	Sensation of dry mouth	42	Diurnal pattern (worse in evening)	68
Sensation of dry mouth	49	Psychological Factors	42	Sensation of dry mouth	59
Middle-aged and older postmenopausal female	31	Middle-aged and older postmenopausal female	35	Psychological Factors	46
Symptom relief during meals	27	Altered taste sensation	23	Symptom relief during meals	36
Altered taste sensation	23	Symptom relief during meals	19	High VAS score	35
High VAS score	10	High VAS score	9	Altered taste sensation	25
Post-dental treatment onset	8	Stress	7	Stress	8
Stress	8	Unilateral nature	4	Unilateral nature	7
Unilateral nature	6	Hormonal changes	4		

Abbreviations: BMS, Burning Mouth Syndrome; VAS, Visual Analogue Scale.

Symptoms" as major factors. However, additional tendencies were observed: ChatGPT emphasized "Psychological Factors", Gemini highlighted "Diurnal Pattern", and Claude placed importance on "Patient Demographics (Middle-aged and older postmenopausal female)".

Differential Diagnosis Statistics

Table 5 presents a summary of the diagnostic terms provided by three LLMs for the differential diagnosis of BMS. The differential diagnoses were compiled for all cases where the final diagnosis was confirmed as BMS. ChatGPT and Gemini

Table 5 Summary of Differential Diagnoses

ChatGPT4o		**Gemini Advanced 1.5 Pro**		**Claude 3.5 Sonnet**	
Neuropathy	56	Neuropathy	63	Geographic Tongue	68
Candidiasis	50	Nutritional Deficiency (Vitamin B12/ Folate iron Zinc etc.)	56	Candidiasis	50
Nutritional Deficiency (Vitamin B12/ Folate iron Zinc etc.)	39	Medication Side Effects	25	Nutritional Deficiency (Vitamin B12/ Folate iron Zinc etc.)	46
Sjögren's Syndrome	32	Allergy	19	Secondary BMS	25
Glossodynia	25	Gastro-Esophageal Reflux Disease (GERD)	13	Neuropathy	23
Oral Dysesthesia	17	Geographic Tongue	12	Xerostomia	19
Xerostomia	17	Candidiasis	12	Persistent Idiopathic Facial Pain	8
Medication Side Effects	17	Glossodynia	11	Glossodynia	8
Secondary Burning Mouth Syndrome	15	Psychological Disorders	10	Allergy	7
Oral Lichen Planus	5	Glossopharyngeal Neuralgia	9	Sjögren's Syndrome	7
Allergy	4	Sjögren's Syndrome	9	Complex Regional Pain Syndrome (CRPS)	4
Trigeminal Neuralgia	4	Xerostomia	7	Stomatitis	4
Stomatitis	4	Taste Disturbances	6	Somatization	4
Psychogenic Pain Disorder	2	Anxiety Disorder or Depression	4	Drug-Induced Dysgeusia	3
Persistent Idiopathic Facial Pain	2	Somatic Symptom Disorder	3	Oral Dysesthesia	3
Gastro-Esophageal Reflux Disease (GERD)	2	Persistent Idiopathic Facial Pain	1	Gastro-Esophageal Reflux Disease (GERD)	2
Referred Pain from Systemic or Local Conditions	1	Psychogenic Dysphagia	1	Medication Side Effects	2
Psychogenic Dysgeusia	1	Age-Related Changes	1	Hormonal-Induced Stomatodynia	1
Central Sensitization Disorders	1	Underlying Medical Conditions	1	Cervical Spondylosis with Glossopharyngeal Neuralgia	1
Glossopharyngeal Neuralgia	1	Referred Pain	1	Temporomandibular Disorder with referred pain	1
Postherpetic Neuralgia (PHN)	1	Burning Mouth Syndrome (BMS)	1	Early-stage Multiple Sclerosis	1

(*Continued*)

Table 5 (Continued).

ChatGPT4o		Gemini Advanced 1.5 Pro		Claude 3.5 Sonnet	
Psychological Factors	1	Systemic Disease	1	Post-stroke Central Pain Syndrome	1
		Temporomandibular Joint Disorder (TMD)	1	Trigeminal Neuralgia	1
				Post-traumatic Stress-related Oral Symptoms	1
				Contact Hypersensitivity Reaction	1
				Oral Allodynia	1
				Menopause-related atrophic changes	1
				Oral Lichen Planus	1
				Neuralgia-inducing Cavitational Osteonecrosis (NICO)	1
				Primary Burning Mouth Syndrome	1
				Drug-Induced Glossitis	1

identified neuropathy and nutritional deficiencies as the primary considerations, whereas Claude, in contrast to the other two, highlighted geographic tongue and secondary BMS as key differential diagnoses.

Assembly of Requested Additional Information

Table 6 summarizes the information and tests requested by LLMs to enhance the reliability of a diagnosis of BMS. ChatGPT exhibited the highest number of requests, followed by Gemini, with fewer from Claude. The requested information varied across the three LLM systems: ChatGPT primarily requested psychiatric symptom assessment and

Table 6 Required Information to Enhance Diagnostic Reliability

ChatGPT4o		Gemini Advanced 1.5 Pro		Claude 3.5 Sonnet	
Psychiatric Symptom Assessment	82	Salivary Gland Function	77	Current medications	84
Current medications	79	Psychiatric Symptom Assessment	74	Blood tests	62
Salivary Gland Function	70	Blood Tests	59	Salivary Gland Function	34
Nutritional status	48	Detailed Medical History	52	Psychiatric Symptom Assessment	30
Blood tests	35	Current medications	25	Recent dental work	25
Detailed Medical History	31	Neurological assessment	14	Menopause or hormonal change	16
Habit (including oral habit)	27	Nutritional status	13	Daily fluctuation of symptoms	14
Taste testing	23	Allergies	8	Stress	12
Neurological and Systemic Workup	18	Taste testing	6	Remission and exacerbation factors	8
Lifestyle Factors	13	Lifestyle Factors	6	Taste testing	6

(*Continued*)

Table 6 (Continued).

ChatGPT4o		Gemini Advanced 1.5 Pro		Claude 3.5 Sonnet	
Allergies	8	Habit (including oral habit)	6	Location of the symptoms	3
Stress, Anxiety, or Depression	4	Recent dental work	6	Habit (including oral habit)	2
Remission and exacerbation factors	4	Stress	5	Lifestyle Factors	2
Daily fluctuation of symptoms	3	Location of the symptoms	3	Neurological assessment	1
Stress	1	Daily fluctuation of symptoms	2		
		Referral to a Mental Health Professional	2		
		Sleep quality	2		

Table 7 Error Analysis for Incorrect Diagnoses

	ChatGPT4o	Gemini Advanced 1.5 Pro	Claude 3.5 Sonnet
Logical Fallacy	N/A	3 (27.3%)	N/A
Informational Fallacy	1 (100%)	8 (72.7%)	1 (100%)
Explicit Fallacy	N/A	N/A	N/A

Abbreviation: N/A, not applicable.

information of current medications; Gemini focused on salivary gland function and psychiatric symptom assessment; Claude predominantly requested list of current medications and blood tests.

Error Analysis of Reasons for Incorrect Diagnoses

Table 7 presents the error analysis for incorrect diagnoses. No explicit fallacies were observed in any case. ChatGPT exhibited one instance (100%) of an informational fallacy. Gemini demonstrated three instances (27.3%) of logical fallacies and eight instances (72.7%) of informational fallacies. Claude displayed one instance (100%) of a informational fallacy.

Discussion

This study analyzed 100 case vignettes and LLM responses to evaluate their ability to diagnose BMS. All the LLMs demonstrated high diagnostic accuracy, with ChatGPT and Claude achieving a diagnostic rate of 99%. This finding indicates the possibility that LLMs can be effective in diagnosing BMS when symptoms, medical history, and patient background are thoroughly gathered in a clinical setting. However, as discussed later, concerns about reliability still persist, suggesting that LLMs are most effective when used as a supplementary tool.

Accuracy of Diagnosis

Before discussing the diagnostic performance, it is important to consider the timeframes in which these LLMs were trained. It is noteworthy that ChatGPT-4o, Gemini Advanced 1.5 Pro, and Claude 3.5 Sonnet have training cutoffs of October 2023, November 2023, and April 2024, respectively. Since no significant new findings on BMS were published between these cutoff dates and our study timeframe (November 15 and 16, 2024), we anticipate that the impact of these training cutoffs and training data on diagnostic performance is minimal.

ChatGPT and Claude: Both models achieved near-perfect accuracy with a performance rate of 99%, and their single incorrect diagnosis highlights the high consistency and robust capability of the algorithms in identifying BMS. This performance suggests that the models rely on finely tuned reasoning strategies closely aligned with the clinical features of BMS.

Gemini: The model achieved an accuracy of 89%, which is significantly lower compared to its peers. This performance, characterized by 11 incorrect diagnoses, suggests either a heightened sensitivity to borderline cases, resulting in an increased tendency for over-diagnosis of alternative conditions, or a failure to adequately capture the distinct features of BMS, leading to diagnostic drift.

BMS inherently exhibits overlapping features with other disorders and symptoms, including neuropathic components, xerostomia, and psychological factors. Consequently, we define cases in which these accompanying symptoms are markedly pronounced as "borderline cases", necessitating considerable emphasis on differential diagnosis. However, under the Gemini system, the exclusion of these concomitant symptoms appears insufficient, potentially increasing the risk of misdiagnosis in borderline cases.

Interpretation of Statistical Significance in a Clinical Context

While our results demonstrate a statistically significant difference in diagnostic accuracy (eg, 99% vs 89%), it is important not to draw conclusions based solely on the presence or absence of significance. Instead, these percentage differences should be examined in light of their clinical impact on patient care, such as the potential risk of misdiagnosis and subsequent treatment choices. For instance, even a relatively small gap in accuracy could lead to different clinical pathways, including additional diagnostic evaluations or consultations, which may affect the patient's overall experience and outcome. Therefore, statistical significance should be balanced against practical considerations in routine clinical practice, particularly for conditions like BMS where precise diagnosis is crucial for effective management.

Error Patterns and Misdiagnoses

ChatGPT and Claude: Both models primarily misdiagnosed PIFP, suggesting that they may over-prioritize chronic pain syndromes with symptomatology resembling BMS. Both models misdiagnosed the same clinical vignette as PIFP. The clinical vignette is shown below.

> A 55-year-old female patient presented with primary complaints of an oral sensation of pressure, numbness, and burning pain, which began following a dental procedure. Her symptoms have persisted for 34 months, with more intense discomfort on the left side. She has no notable medical history, and her symptom severity is rated at 78 on the Visual Analogue Scale (VAS). No organic abnormalities were observed in the oral cavity, and panoramic radiographic imaging also revealed no structural abnormalities.

Their error profiles reveal a shared limitation in differentiating BMS from PIFP, which often exhibit overlapping clinical characteristics, including chronic pain and an absence of visible pathology. Furthermore, incorrect diagnosis of PIFP is clinically regarded as more significant than incorrect diagnoses in Gemini. In this clinical vignette, it is considered that ChatGPT has committed an Informational Fallacy by failing to specify the exact location of the oral symptoms. As for Claude, it is also deemed to have committed an Informational Fallacy due to misinformation in its training data, by suggesting "Burning Mouth Syndrome (BMS)—Usually not related to dental procedures" as a differential diagnosis.

Gemini: Error analysis revealed that informational fallacies accounted for 72.7% of the identified issues. These fallacies stemmed from logical thought processes that were compromised by the overestimation or underestimation of the given information, as well as biases introduced by the inherent training data. The findings also highlighted a more diverse error profile, including somatoform disorders, which reflect difficulties in distinguishing functional syndromes; drug-induced burning mouth sensations, indicating an over-reliance on medication-related etiologies; and cases of BMS combined with neuropathic pain, revealing challenges in recognizing BMS as a standalone diagnosis. This diversity emphasizes the presence of less targeted diagnostic reasoning and a potentially weaker integration of the primary diagnostic criteria for BMS.

Below, we present a qualitative insight drawn from a misdiagnosis example—specifically, a case in which the LLM's training data was insufficient and thus serves as a typical example of an informational fallacy—to illustrate the performance of large language models (LLMs). The following content was input into the LLM:

> A 53-year-old female patient presents with a primary complaint of a burning sensation and dull pain in the palate. The symptoms began 13 months ago, following the emotional stress associated with caring for and the eventual passing of her mother. The pain is unilateral, affecting only the left side, and tends to be more intense in the evening compared to the morning. Interestingly, the symptoms diminish during meals. Her medical history includes a gastric hernia and postnasal drip. The severity of the symptoms, assessed by a Visual Analogue Scale, is 64. No organic abnormalities were observed in the oral cavity, and panoramic radiographic imaging also revealed no structural abnormalities.

In response to this clinical vignette, Gemini provided the following reasoning:

However, there are some atypical features for BMS:

- Unilateral pain: BMS usually presents bilaterally.
- Pain diminishes during meals: This is unusual for BMS, where eating can sometimes exacerbate symptoms due to increased awareness of oral sensations.

From a clinical standpoint, those familiar with BMS recognize that it does not necessarily manifest as bilateral pain in all cases.[27] In addition, it is well known that many patients report that their pain diminishes during meals. Based on other similar clinical vignettes, it appears that LLMs sometimes offer limited analysis concerning topics with multiple conflicting opinions—especially those that are clinically recognized but mentioned in only a few published studies.

When considering the clinical application of LLMs, this limitation should be addressed. For content or diagnoses that remain open to debate, a clinician's judgment is indispensable. Furthermore, whenever providing a diagnosis, it is prudent to make the LLM's reasoning explicit. If errors are identified, they should be corrected to facilitate ongoing learning and improvement of these models.

Deeper Analysis of Gemini's Informational Fallacies and Diagnostic Limitations

Gemini exhibited a notably high rate of informational fallacies (72.7%), suggesting a recurrent pattern of misinterpreting or overextending key clinical details. Unlike ChatGPT and Claude, both of which predominantly struggled with differentiating BMS from closely related conditions (eg, PIFP), Gemini's diverse range of diagnostic errors reflects less targeted diagnostic reasoning and weaker integration of BMS criteria. For instance, over-reliance on medication-related etiologies points to a tendency for Gemini to prioritize certain risk factors if they are prominently mentioned, even when contradictory signs or symptoms are present.

In clinical practice, these inaccuracies could lead to unnecessary diagnostic steps or, conversely, to overlooking hallmark features of BMS. This high error rate also indicates that Gemini may struggle when confronted with ambiguous or incomplete data, overestimating or underestimating the significance of specific details to fill perceived "gaps" in the clinical picture. Consequently, clinicians relying on Gemini—especially in settings with limited patient information—would need to confirm or refine the model's output through targeted questioning, additional diagnostic tests, or collaboration with specialists.

From a technical standpoint, further refinement of Gemini's training data or prompt-engineering strategies may help it more consistently align with established BMS criteria. Additionally, future model iterations could benefit from incorporating explicit rules or weighting systems that reduce the likelihood of logical overextensions. Such approaches may mitigate Gemini's tendency to over-diagnose alternative conditions and foster more reliable identification of key diagnostic thresholds for BMS.

Overall, recognizing these distinct error patterns is essential for clinicians and developers alike, as it underscores the importance of tailoring LLM usage to specific clinical contexts. While Gemini's broader error profile might be advantageous for atypical cases, the persistent risk of informational fallacies highlights the need for targeted oversight and the continuous recalibration of the model to maximize clinical safety and diagnostic accuracy.

Diagnostic Reasoning

The reasoning reflects what the LLMs prioritize in determining the diagnosis of BMS. It lists various factors, including characteristic symptoms like daily fluctuation. However, what truly matters is whether the criteria for BMS are met and whether any differential diagnoses can be ruled out.

ChatGPT: The key reasoning points for diagnosing BMS include several critical aspects. First, the absence of organic abnormalities is utilized in 91% of cases, underscoring the heavy reliance on exclusionary criteria. Second, the primary symptoms of BMS, such as pain, a burning sensation, an oral discomfort, are referenced in 89% of cases, demonstrating alignment with established clinical markers. Additionally, the chronic nature of symptoms is highlighted in 79% of cases, emphasizing the persistent course of the condition. The strengths of this diagnostic approach lie in its systematic and logical progression, transitioning from exclusionary reasoning to a symptomatic focus. However, a notable weakness lies in the overestimation of psychosocial and functional domains, which may lead to misdiagnosis and subsequently influence treatment strategies. ChatGPT often tends to associate cases with psychological factors, particularly when there is a history of mental illness, specific triggers, or multiple or severe past illnesses.

Gemini: The reasoning patterns related to BMS exhibit variability. Primary symptoms of BMS were referenced in 87% of cases, demonstrating usage comparable to peers. The absence of organic abnormalities was cited in 85% of cases, reflecting a reasonable application of exclusionary logic. However, the chronic nature of symptoms was highlighted in only 58% of cases, indicating a potential gap in emphasizing the persistent characteristics of BMS. This variability in reasoning offers strengths, such as the potential to identify edge cases or atypical presentations, but also reveals weaknesses, including a lack of consistency and an over-reliance on specific features, which may increase the risk of misdiagnosis.

Claude: In 95% of cases, the absence of organic abnormalities serves as the dominant reasoning for diagnosis. Primary symptoms of BMS are highlighted in 91% of cases, emphasizing their diagnostic relevance. The chronic nature of symptoms is considered in 76% of cases, which is slightly lower compared to ChatGPT. The approach demonstrates strengths through its structured methodology and high alignment with BMS diagnostic criteria. However, it shares similar limitations with ChatGPT, particularly in distinguishing BMS from closely related conditions.

Implications of Exclusionary Criteria and Psychosocial Emphasis

One important finding from our analysis is the LLMs' strong reliance on exclusionary criteria—particularly the absence of identifiable organic abnormalities—in diagnosing BMS. While exclusionary logic is aligned with many guidelines for diagnosing BMS, it may pose a risk of overlooking atypical presentations or comorbid conditions if clinicians rely solely on the model's output. Additionally, each LLM demonstrated a distinct emphasis on psychosocial or psychiatric factors, which can have practical implications for patient care. Overemphasizing psychosomatic components may lead to inappropriate referrals or under-treatment of potential biological contributors, whereas underestimating the psychological dimension could delay timely intervention for conditions like depression or anxiety. These patterns underscore the importance of maintaining a balanced, multidisciplinary approach—one that integrates thorough clinical evaluation and psychosocial assessment rather than relying solely on algorithmic exclusion. Recognizing and critically appraising each model's diagnostic rationale is essential for preventing misdiagnoses, ensuring appropriate referral pathways, and ultimately improving patient outcomes.

Information to Improve Diagnostic Reliability

Investigating the information requested as necessary is considered important for establishing a definitive diagnosis. The evaluation of psychiatric symptoms is crucial for treatment planning and collaboration with psychiatry. However, it should be noted that while BMS is associated with psychiatric disorders, the presence or absence of such disorders is not relevant to the diagnosis of BMS.

ChatGPT: The highlighted aspects of the assessment reveal critical areas of focus: Psychiatric Symptom Assessment (82%) underscores the sensitivity to psychosocial factors associated with BMS, Current Medications (79%) reflect an awareness of the potential for drug-induced oral conditions, and Salivary Gland Function (70%) suggests an emphasis on evaluating functional impairments. On the other hand, requests for blood tests and detailed medical history were comparatively less frequent. The approach exhibits strengths in its integration of a biopsychosocial perspective but could be further improved by incorporating a more detailed evaluation of systemic factors.

Gemini: The analysis highlights the emphasis on key diagnostic areas, with a prioritization of detailed medical history (52%), suggesting the importance of comprehensive data gathering. A significant focus is placed on salivary gland function (77%), aligning with its emphasis on functional assessments. Additionally, psychiatric symptom assessment (74%) shows considerable consistency with the approach utilized by ChatGPT. While the framework demonstrates strengths in prioritizing foundational diagnostic data, its limitations include a relatively lower emphasis on the role of medications, which may influence drug selection in future pharmacological therapies.

Claude: The suggested priorities indicate a strong emphasis on iatrogenic factors, as evidenced by the high consideration of current medications (84%). Blood tests, which highlight a preference for objective biomarkers, were prioritized by 62% of respondents. In contrast, the assessment of psychiatric symptoms, at 30%, received less emphasis compared to peer evaluations. This approach demonstrates strengths in prioritizing actionable clinical data but reveals a relative weakness in addressing psychiatric or psychosocial dimensions. Compared to other LLMs, Claude tended to require less supplementary information.

Differential Diagnoses

Overall, differential diagnoses have the potential to undermine the reliability of LLMs in diagnosing BMS. Neuropathy is considered a critical differential diagnosis by both ChatGPT and Gemini. Given that BMS may include neuropathic features such as allodynia, its inclusion as a differential diagnosis is naturally justified.[28] On the other hand, it is problematic to include easily distinguishable oral conditions such as geographic tongue, xerostomia, or oral candidiasis in the differential diagnosis. Nutritional deficiencies are frequently mentioned in reviews; however, they rarely considered in the clinical differential diagnosis. Moreover, their symptoms differ significantly from those of BMS; therefore, while they should be recognized as part of the differential diagnosis, they can often be excluded with relative ease.

Other major issues are the ambiguities in differentiating BMS, glossodynia, and oral dysesthesia, highlighting the potential inaccuracies in diagnostic approaches based on LLMs. To assess their differentiation capability, three LLMs were tasked with summarizing the distinctions among BMS, glossodynia, and oral dysesthesia in a tabular format. The results, which were largely similar across the models, are summarized in Table 8 using ChatGPT-generated content. Glossodynia is essentially an older term for BMS, with no fundamental differences in symptoms. In contrast, oral dysesthesia primarily refers to discomfort within the oral cavity, although it can include pain. These findings indicate that all three LLMs misinterpreted the distinctions between BMS, glossodynia and oral dysesthesia.

In oral medicine practice, critical differential diagnoses frequently include PIFP and oral cenesthopathy. PIFP is characterized by facial pain in the head and neck region, while oral cenesthopathy frequently involves oral discomfort, leading to symptom overlap.[29] The lack of a clear definition for oral cenesthopathy further complicates diagnosis.[30] Thus, it is essential to carefully assess the severity of the patient's pain and their oral discomfort symptoms. In such cases, the

Table 8 Summary of Differences

Term	Focus	Symptoms	Scope	Underlying Cause
Burning Mouth Syndrome	Chronic condition	Burning pain, dry mouth	Entire oral cavity	Idiopathic or systemic
Glossodynia	Symptom of the tongue	Pain (not necessarily burning)	Tongue only	Local or systemic
Oral Dysesthesia	Abnormal sensations	Burning, tingling, numbness	Any area of the oral cavity	Idiopathic or associated with causes

Oral Dysesthesia Rating Scale (OralDRS) is considered a useful tool for evaluating oral cenesthopathy symptoms.[31] However, diagnosing overlapping conditions, such as MUOS, remains challenging.

Complementary Relationship Between LLMs and Dentists

LLMs can analyze vast amounts of data quickly, enabling consistent diagnostic capabilities. However, adapting to patient backgrounds and contexts still relies on medical professionals' expertise. For diagnosing BMS, general dentists often have limited knowledge and capabilities, with diagnosis typically restricted to university hospitals or specialized institutions. In this context, LLMs show significant potential as a complementary tool to bridge this gap. Our study indicates that LLMs demonstrate diagnostic capabilities surpassing those of general dentists in identifying BMS, presenting a new paradigm for clinical support.

Nevertheless, LLMs have inherent limitations. Pain and discomfort described by patients are highly subjective and complex, making it challenging for LLMs to fully understand the associated contexts. Therefore, a "human-in-the-loop" approach, where a medical professional verifies the diagnostic results proposed by LLMs, is essential.[32] This approach not only mitigates the risk of misdiagnosis or overdiagnosis but also helps alleviate the workload of medical practitioners.

Physical diagnostic techniques, such as palpation, remain irreplaceable by LLMs. While LLMs have improved in areas like oral cavity examinations and panoramic X-ray evaluations, their reliability is still insufficient. These aspects continue to require the specialized expertise of dentists.

Real-world clinical contexts frequently involve incomplete or ambiguous patient data, such as missing laboratory results, vague symptom descriptions, or limited medical histories. In these scenarios, LLM-driven diagnoses may not achieve the same level of performance observed in carefully curated vignettes, underscoring the necessity of a "human-in-the-loop" approach. By gathering additional information through direct examination, specialized testing, and further history-taking, clinicians can fill these data gaps and refine or validate the AI-generated impressions. This synergy between LLMs and experienced professionals is thus particularly crucial when information is incomplete, ensuring that diagnoses remain accurate and clinically safe even under suboptimal data conditions.

Additionally, the unique risk of "hallucination", where LLMs produce erroneous information, must be considered.[33] This phenomenon can have critical consequences in the medical field, making it dangerous to accept LLMs' diagnostic results without scrutiny. A robust confirmation process by experienced dentists is indispensable to address the challenges posed by LLMs.

Integrating the complementary roles of LLMs and dentists offers the opportunity to establish a new clinical framework that not only improves diagnostic accuracy for BMS but also enhances treatment efficiency and the quality of patient care. Furthermore, leveraging LLMs can enable high-quality diagnostics in regions where access to specialists is limited, contributing to the reduction of healthcare disparities. Moving forward, efforts should focus on evolving the supportive role of LLMs to provide a beneficial clinical environment for both patients and dentists.

Actionable Recommendations for Clinicians Integrating LLMs Into Practice

To ensure the safe and effective adoption of LLMs in clinical settings, clinicians should undergo supplementary training that includes basic AI literacy, critical evaluation of AI-generated outputs, and awareness of common error patterns (eg, hallucinations or informational fallacies). Clear guidelines or protocols would help standardize LLM use, outlining when LLM-based advice may be considered sufficiently reliable and when further specialist consultation is warranted. Additionally, clinicians are encouraged to use LLM outputs as a complementary source—for instance, a second opinion in ambiguous cases—while consistently verifying AI-driven diagnoses through detailed patient assessments and confirmatory diagnostic tests. This structured approach can help mitigate the risk of over-reliance on AI and reinforce the human-in-the-loop framework, ultimately optimizing patient outcomes and minimizing misdiagnoses.

Ethical Considerations and the Risk of Over-Reliance on AI

Although our findings suggest that LLMs can provide substantial support for diagnosing BMS, it is crucial to address the ethical implications of implementing AI-driven tools in clinical settings. One primary concern is the potential for over-reliance on AI, wherein clinicians might accept the model's outputs uncritically, thereby overlooking atypical symptom presentations or failing to exercise independent clinical judgment. In such instances, systemic errors—such as algorithmic biases and "hallucination" phenomena—could go unchecked, resulting in suboptimal or even harmful patient outcomes. Moreover, the absence of regulatory frameworks specific to AI-assisted diagnostics raises questions about liability, informed consent, and patient autonomy. Clinicians must therefore remain vigilant, recognizing that AI tools are meant to augment rather than replace the nuanced decision-making abilities and empathetic understanding unique to human practitioners. A "human-in-the-loop" model, wherein a qualified professional verifies and contextualizes all AI-generated recommendations, can mitigate the ethical risks associated with over-reliance and ensure that patient welfare remains the central priority. By integrating transparent validation processes, adhering to strict data governance policies, and maintaining clear communication with patients about the capabilities and limitations of AI, healthcare teams can responsibly harness LLMs' diagnostic potential while safeguarding ethical standards in patient care.

Clinical Integration and Recommendations for General Dentists

From a practical standpoint, general dentists may find LLM-based diagnostic tools most beneficial when used as an initial screening or "second opinion" mechanism, rather than a definitive source of diagnostic authority. For instance, dentists could input patient data—such as symptom onset, location, and severity—into the LLM to generate potential diagnoses, including BMS and related conditions. However, given that LLMs occasionally misclassify BMS as Persistent Idiopathic Facial Pain (PIFP) due to overlapping symptoms and the absence of overt pathology, it is imperative that dentists critically evaluate these outputs. We recommend a structured "human-in-the-loop" approach, where dentists: (1) perform a thorough clinical examination and history-taking to identify or rule out organic abnormalities; (2) carefully compare the LLM's suggestions against known diagnostic criteria for BMS, PIFP, or other oral and maxillofacial pain disorders; (3) consider psychosocial factors that might exacerbate symptoms; and (4) confirm the final diagnosis through additional investigations or referrals (eg, neurology, psychiatry) when there is ambiguity. By following these steps, dentists can incorporate LLM-generated insights into their workflow while minimizing the risk of misdiagnosis, ensuring that clinical judgment remains paramount. This approach also acknowledges the inherent limitations of current LLMs in distinguishing among overlapping conditions, emphasizing that AI outputs should complement, rather than replace, expert decision-making.

Potential Biases and Strategies for Improving Model Reasoning

As described in the Materials and Methods section, the clinical vignettes were derived from consecutive real-world patient data and carefully modified in a manner that preserves the essential characteristics of BMS. Consequently, we believe these vignettes adequately reflect real-world scenarios with minimal bias in demographic or clinical representations. Nonetheless, certain atypical or rare presentations may not be fully captured, potentially limiting the models' applicability in highly complex cases. Looking ahead, further strategies for improving model reasoning in complex scenarios include the integration of more diverse patient data—such as imaging results or laboratory findings—enhanced prompt-engineering techniques that guide the model toward more nuanced clinical reasoning, and collaboration with multidisciplinary teams to address multifactorial presentations. By incorporating these approaches, LLMs may achieve greater robustness and reliability, particularly for patients presenting with overlapping or less typical symptoms.

Broader Implications for AI in Clinical Decision-Making and Future Directions

In a broader context, our findings contribute to the growing body of evidence supporting the role of AI—particularly LLMs—in enhancing clinical decision-making. By demonstrating that LLMs can achieve a high degree of diagnostic accuracy for BMS, we provide insights into how such models may reduce diagnostic uncertainty in complex, subjective conditions. This work underscores the potential for LLMs to function as valuable adjuncts to clinician expertise, particularly in medical and dental fields where symptom presentations are nuanced and overlapping.

Looking beyond BMS, LLM-driven diagnostic support may offer substantial benefits in various medical domains that rely heavily on clinical judgment and patient-reported outcomes. For instance, the ability of LLMs to rapidly process patient histories and incorporate large bodies of medical literature could facilitate more timely and accurate diagnoses, particularly in under-resourced settings with limited access to specialists. However, real-world implementation requires careful consideration of ethical, legal, and educational factors to safeguard patient safety and maintain professional accountability.

Integration of LLMs Into Multi-Disciplinary Diagnostic Teams

Future research should explore the structured integration of LLMs into multi-disciplinary diagnostic teams composed of dentists, oral surgeons, psychiatrists, neurologists, and other relevant specialists. Such teams could leverage the computational strengths of LLMs—especially in systematically ruling out organic causes and synthesizing diverse patient data—while clinicians contribute the essential contextual, psychosocial, and experiential knowledge that AI models currently lack. By collaboratively evaluating high-risk or ambiguous cases, human–AI "co-diagnosis" could refine diagnostic accuracy, foster individualized treatment plans, and promote shared decision-making between patients and healthcare providers.

Broadening the Scope to Other Medically Unexplained Oral Symptoms

Beyond BMS, our findings carry implications for other MUOS, such as oral cenesthopathy, and persistent idiopathic facial pain. These conditions share the diagnostic challenges of subjective symptomatology and uncertain etiology, making them potential targets for LLM-assisted diagnostic support. By incorporating comprehensive patient histories, psychosocial factors, and multidisciplinary inputs (eg, from neurology, psychiatry, and oral medicine), future research could adapt our LLM-driven framework to systematically evaluate these complex disorders. Such an approach may facilitate earlier and more accurate diagnoses, reduce the burden of extensive referral processes, and ultimately enhance patient outcomes. Moreover, refining the training data to include a broader array of MUOS presentations could yield more generalizable AI tools, fostering cross-disciplinary collaboration and supporting the development of holistic, patient-centered care models.

Recommendations for Future Investigations

1. Prospective Multi-Center Trials: Conducting large-scale, real-world clinical evaluations in multiple institutions would help clarify the utility and limitations of LLMs in diverse populations and practice settings.
2. Context-Aware Model Enhancements: Improving prompt engineering and incorporating domain-specific data (eg, imaging, laboratory findings, multidisciplinary case notes) could help LLMs navigate subtle diagnostic distinctions.
3. Interdisciplinary Collaboration: Engaging AI engineers, clinicians from various specialties, ethicists, and policy-makers will be vital to develop guidelines that address liability, data governance, and training requirements for integrated AI systems.
4. Educational Resources: Developing standardized training modules for clinicians and students could help them understand how best to interpret and utilize LLM outputs in practice, including recognizing potential sources of bias or "hallucinations".

By embedding LLMs into a framework that respects both computational capabilities and the indispensable role of human expertise, we can move toward more equitable and efficient clinical decision-making processes. Ultimately, this integrative approach can reduce diagnostic delays, optimize patient outcomes, and ensure robust ethical safeguards in the rapidly evolving intersection of AI and healthcare.

Limitation

This study has several limitations. The use of only 100 vignettes, while consistent with similar LLM diagnostic studies, may limit the statistical power to detect smaller differences in performance between models and restrict the scope of LLM evaluation. These scenarios may not comprehensively capture the full range of symptoms and patient backgrounds relevant to BMS, and the relative paucity of male-oriented vignettes could compromise the validation of LLMs' ability to diagnose BMS in male patients. Despite these constraints, the chi-square analysis revealed significant performance disparities ($p < 0.001$), underscoring the robustness of the detected differences. Future research should thus employ larger and more diverse datasets, both to confirm the generalizability of these findings and to enhance statistical precision, particularly in evaluating LLM performance across a broader spectrum of clinical presentations.

Another limitation of this study is that we did not distinguish between primary and secondary BMS in our diagnostic criteria. Consequently, the generalizability of our findings to clinical settings or research protocols that rely on this classification remains unclear. Future investigations should incorporate explicit primary vs secondary BMS distinctions, as recommended by certain clinical guidelines, to determine whether such a classification significantly impacts diagnostic accuracy or therapeutic decision-making.

Furthermore, our study was designed to assess the diagnostic accuracy of each LLM based solely on its initial output. We did not examine the potential for iterative prompting, additional inquiries, or subsequent interactions that might correct misdiagnoses. Consequently, the study may not fully reflect the models' capacity to refine or improve their diagnostic suggestions in a real-world clinical workflow, where such iterative processes are common. Future investigations should incorporate repeated exchanges and follow-up prompts to determine whether these interactions enhance diagnostic accuracy and reduce error rates in actual practice.

The cross-sectional design poses challenges in tracking how LLMs performance evolves over time, especially considering that these systems undergo regular updates and improvements. Adopting a longitudinal approach in future studies could provide deeper insights into how LLMs performance progresses and its role in disseminating medical knowledge over time.

This study did not incorporate image generation capabilities, and thus assumed no organic abnormalities in intraoral or panoramic radiographs. However, diagnosing the absence of organic abnormalities in clinical practice requires sufficient expertise in oral medicine, oral surgery, and dental radiology. Moreover, as BMS could potentially mask underlying malignancies, clinical skills such as palpation are indispensable.[34] If any suspicion arises, it is crucial to refer the patient to an appropriate specialist.

Conclusion

We found that LLMs can diagnose BMS with high accuracy, contrary to our initial expectation that the condition's complexity and subjective nature would pose significant challenges. Although these LLMs performed well overall, we observed a few logical gaps and incorrect diagnoses, highlighting the need for expert review. At present, LLMs appear most suitable as supplementary tools to assist clinicians rather than as standalone diagnostic systems.

Moving forward, it is important to refine the way LLMs are trained and to explore how they can be integrated effectively into clinical practice. For example, we can diversify and expand the patient data used for training and involve specialists in oral medicine to reduce errors. We should also define clear protocols for "human–AI collaboration", ensuring that clinicians confirm AI-generated diagnoses, especially for complex cases.

Beyond BMS, LLMs may have broader applications in diagnosing other MUOS and in various clinical fields such as neurology or psychiatry, where diagnosis often depends on subtle, subjective information. However, ethical issues—like the risk of over-reliance on AI or potential misinformation ("hallucination")—must be considered. Practical safeguards, transparent data use, and accountability frameworks are essential to maintain patient safety and trust.

Future research should focus on:

Prospective Clinical Trials: Testing LLMs in real-world settings to track their performance over time.

Treatment and Education: Exploring their role in treatment planning, patient education, and professional training.

Model Updates: Regularly auditing and refining LLMs so they remain aligned with the latest evidence-based guidelines.

Although LLMs show promise in improving BMS diagnostics, human expertise remains crucial. Clinicians should validate AI outputs, manage complex cases that go beyond the models' capabilities, and offer personalized care. By combining the strengths of LLMs with expert clinical judgment, we can potentially enhance diagnostic accuracy, streamline workflows, and extend quality care to regions with limited specialist availability, all while minimizing ethical and safety risks.

Abbreviations

AI, Generative Artificial Intelligence; LLMs, large language models; MUOS, medically unexplained oral symptoms; BMS; burning mouth syndrome; ICHD-3, the International Classification of Headache Disorders, 3rd edition; VAS, visual analog scale; PIFP, Persistent Idiopathic Facial Pain.

Data Sharing Statement

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Ethics Approval and Informed Consent

In accordance with the Declaration of Helsinki and following review by the Tokyo Medical and Dental University Hospital Ethical Committee (approval number: D2013-005-04), the patient who initially served as the basis for the clinical vignettes provided written, comprehensive research consent. Although the clinical vignettes used in this study were originally derived from actual patient information, details such as age, symptoms, and medical history were modified. Consequently, because no genuine patient data were employed, informed consent was deemed unnecessary, and, given the nature of the research, ethics committee review was not required.

Consent for Publication

The clinical vignettes used in this study were originally derived from actual patient information, details such as age, symptoms, and medical history were modified. Consequently, because no genuine patient data were employed, informed consent was deemed unnecessary.

Author Contributions

All authors contributed substantially to the work reported—whether through conceptualization, study design, execution, data acquisition, or analysis and interpretation, or through multiple of these tasks—and participated in drafting, revising, or critically reviewing the manuscript. They provided final approval for the version submitted, agreed on the choice of journal for submission, and accept responsibility for all aspects of the work.

Funding

This research was funded by KAKENHI from the Japanese Society for the Promotion of Science (JSPS), grant number 22K10141 to AT.

Disclosure

The authors declare that there are no competing interests.

References

1. Lin X, Jin R, Huang W, Ye Y, Jin J, Zhu W. Trends of burning mouth syndrome: a bibliometric study. *Front Neurol*. 2024;15:1443817. doi:10.3389/fneur.2024.1443817
2. Headache Classification Committee of the International Headache Society (IHS) The International Classification of Headache Disorders, 3rd edition. *Cephalalgia*. 2018;38(1):1–211. doi:10.1177/0333102417738202
3. Bogetto F, Maina G, Ferro G, Carbone M, Gandolfo S. Psychiatric comorbidity in patients with burning mouth syndrome. *Psychosomatic Med*. 1998;60(3):378–385. doi:10.1097/00006842-199805000-00028
4. Toyofuku A, Matsuoka H, Abiko Y. Reappraising the psychosomatic approach in the study of "chronic orofacial pain": looking for the essential nature of these intractable conditions. *Front Pain Res*. 2024;5:1349847. doi:10.3389/fpain.2024.1349847
5. Toyofuku A. Psychosomatic problems in dentistry. *Biopsychosoc Med*. 2016;10:14. doi:10.1186/s13030-016-0068-2
6. Luo S, Ivison H, Han SC, Poon J. Local interpretations for explainable natural language processing: a survey. *ACM Comput Surv*. 2024;56(9):1–36. doi:10.1145/3649450
7. Arruzza ES, Evangelista CM, Chau M. The performance of ChatGPT-4.0o in medical imaging evaluation: a preliminary investigation. *J Educ Eval Health Prof*. 2024;21:29. doi:10.3352/jeehp.2024.21.29
8. Chen JS, Reddy AJ, Al-Sharif E, et al. Analysis of ChatGPT responses to ophthalmic cases: can ChatGPT think like an ophthalmologist? *Ophthalmol Sci*. 2025;5(1):100600. doi:10.1016/j.xops.2024.100600

9. Fraser H, Crossland D, Bacher I, Ranney M, Madsen T, Hilliard R. Comparison of diagnostic and triage accuracy of Ada Health and WebMD Symptom Checkers, ChatGPT, and physicians for patients in an emergency department: clinical data analysis study. *JMIR mHealth uHealth*. 2023;11:e49995. doi:10.2196/49995
10. Huang H, Zheng O, Wang D, et al. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. *Int J Oral Sci*. 2023;15(1):29. doi:10.1038/s41368-023-00239-y
11. Wang D, Zhang S. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artif Intell Rev*. 2024;57(11). doi:10.1007/s10462-024-10921-0
12. Bellanda VCF, Santos MLD, Ferraz DA, Jorge R, Melo GB. Applications of ChatGPT in the diagnosis, management, education, and research of retinal diseases: a scoping review. *Int J Retina Vitreous*. 2024;10(1):79. doi:10.1186/s40942-024-00595-9
13. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Commun Med*. 2023;3(1):141. doi:10.1038/s43856-023-00370-1
14. Uehara O, Morikawa T, Harada F, et al. Performance of ChatGPT-3.5 and ChatGPT-4o in the Japanese National Dental Examination. *J Dent Educ*. 2024. doi:10.1002/jdd.13766
15. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230–243. doi:10.1136/svn-2017-000101
16. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. *Asian J Psychiatr*. 2023;89:103770. doi:10.1016/j.ajp.2023.103770
17. Labinsky H, Nagler LK, Krusche M, et al. Vignette-based comparative analysis of ChatGPT and specialist treatment decisions for rheumatic patients: results of the Rheum2Guide study. *Rheumatol Int*. 2024;44(10):2043–2053. doi:10.1007/s00296-024-05675-5
18. Hirosawa T, Kawamura R, Harada Y, et al. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. *JMIR Med Inform*. 2023;11:e48808. doi:10.2196/48808
19. Kolkka-Palomaa M, Jaaskelainen SK, Laine MA, Teerijoki-Oksa T, Sandell M, Forssell H. Pathophysiology of primary burning mouth syndrome with special focus on taste dysfunction: a review. *Oral Dis*. 2015;21(8):937–948. doi:10.1111/odi.12345
20. Gu Y, Baldwin S, Canning C. Hypovitaminosis D, objective oral dryness, and fungal hyphae as three precipitating factors for a subset of secondary burning mouth syndrome. *Heliyon*. 2023;9(9):e19954. doi:10.1016/j.heliyon.2023.e19954
21. Kishore J, Shaikh F, Zubairi AM, et al. Evaluation of serum neuron specific enolase levels among patients with primary and secondary burning mouth syndrome. *Cephalalgia*. 2022;42(2):119–127. doi:10.1177/03331024211046613
22. Adamo D, Spagnuolo G. Burning mouth syndrome: an overview and future perspectives. *Int J Environ Res Public Health*. 2022;20(1). doi:10.3390/ijerph20010682
23. Shim Y. Treatment for burning mouth syndrome: a clinical review. *J Oral Med Pain*. 2023;48(1):11–15. doi:10.14476/jomp.2023.48.1.11
24. Gupta R, Herzog I, Park JB, et al. Performance of ChatGPT on the plastic surgery inservice training examination. *Aesthet Surg J*. 2023;43(12):NP1078–NP1082. doi:10.1093/asj/sjad128
25. Mendonca de Moura JD, Fontana CE, Reis da Silva Lima VH, de Souza Alves I, Andre de Melo Santos P, de Almeida Rodrigues P. Comparative accuracy of artificial intelligence chatbots in pulpal and periradicular diagnosis: a cross-sectional study. *Comput Biol Med*. 2024;183:109332. doi:10.1016/j.compbiomed.2024.109332
26. Calabria E, Canfora F, Leuci S, et al. Gender differences in pain perception among burning mouth syndrome patients: a cross-sectional study of 242 men and 242 women. *Sci Rep*. 2024;14(1):3340. doi:10.1038/s41598-024-53074-4
27. Kim MJ, Kim J, Kho HS. Comparison of clinical characteristics between burning mouth syndrome patients with bilateral and unilateral symptoms. *Int J Oral Maxillofac Surg*. 2020;49(1):38–43. doi:10.1016/j.ijom.2019.06.013
28. Kouri M, Adamo D, Vardas E, et al. Small fiber neuropathy in burning mouth syndrome: a systematic review. *Int J mol Sci*. 2024;25(21):11442. doi:10.3390/ijms252111442
29. Gerwin R. Chronic facial pain: trigeminal neuralgia, persistent idiopathic facial pain, and myofascial pain syndrome-an evidence-based narrative review and etiological hypothesis. *Int J Environ Res Public Health*. 2020;17(19):7012. doi:10.3390/ijerph17197012
30. Umezaki Y, Miura A, Watanabe M, et al. Oral cenesthopathy. *Biopsychosoc Med*. 2016;10:20. doi:10.1186/s13030-016-0071-7
31. Uezato A, Toyofuku A, Umezaki Y, et al. Oral dysesthesia rating scale: a tool for assessing psychosomatic symptoms in oral regions. *BMC Psychiatry*. 2014;14:1696. doi:10.1186/s12888-014-0359-8
32. Xiao H, Wang P. Llm a*: human in the loop large language models enabled a* search for robotics. *arXiv preprint arXiv:231201797*. 2023.;
33. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55(12):1–38. doi:10.1145/3571730
34. Suga T, Tu TTH, Takenoshita M, et al. Case report: hidden oral squamous cell carcinoma in oral somatic symptom disorder. *Front Psychiatry*. 2021;12:651871. doi:10.3389/fpsyt.2021.651871

Journal of Pain Research

Dovepress
Taylor & Francis Group

Publish your work in this journal

The Journal of Pain Research is an international, peer reviewed, open access, online journal that welcomes laboratory and clinical findings in the fields of pain research and the prevention and management of pain. Original research, reviews, symposium reports, hypothesis formation and commentaries are all considered for publication. The manuscript management system is completely online and includes a very quick and fair peer-review system, which is all easy to use. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/journal-of-pain-research-journal