J Stomatol Oral Maxillofac Surg 127 (2026) 102972

Available online at ScienceDirect www.sciencedirect.com

Elsevier Masson France EM|consulte www.em-consulte.com

Original Article

Guideline-based evaluation of five generative artificial intelligence chatbot platforms for clinician-oriented questions on open temporomandibular joint surgery

Selin Gaş [a], Erdinç Sulukan [b], Büşra Korkmaz [a,*]

[a] Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Istanbul Gelisim University, Cihangir Mah., Şehit Jandarma Komando Er Hakan Öner Sok. No: 1, 34310 Avcılar, Istanbul, Turkey

[b] Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Istanbul Beykent University, Ayazağa Mah., Hadım Koruyolu Cad. No: 19, 34398 Sarıyer, Istanbul, Turkey

ARTICLE INFO

Keywords:
Artificial intelligence
Generative artificial intelligence
Large language models
Temporomandibular joint
Arthroplasty
Clinical decision-making

ABSTRACT

Objective: To compare the quality and readability of responses from five generative artificial intelligence chatbot platforms to clinician-oriented questions on open temporomandibular joint (TMJ) surgery against guideline-based reference answers.

Material and methods: Forty questions across eight domains were submitted on 16 July 2026 to ChatGPT (GPT-5.5), Claude (Opus 4.8), Gemini (3.1 Pro), Grok (4) and Perplexity (Pro), each via its paid tier at default settings (200 responses). Two blinded oral and maxillofacial surgeons applied the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool, the Global Quality Score (GQS) and a five-point overall quality rating using a priori key elements and written anchors. Readability was assessed with the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease Score (FRES); platforms were compared with the Friedman test and Bonferroni-corrected Wilcoxon post-hoc tests; inter-rater reliability used intraclass correlation coefficients (ICC).

Results: Inter-rater reliability was good to excellent (average-measure ICC 0.93–0.99). All outcomes except clarity differed among platforms ($P < .001$). Perplexity achieved the highest QAMAI total (27.5 ± 1.4; Kendall's W = 0.75), largely through retrieval-based source provision; excluding this domain, Claude, Perplexity and Gemini converged. Claude had the highest GQS (4.5 ± 0.6), Gemini the highest overall rating (4.3 ± 0.7); Grok scored lowest. Claude and Gemini were least readable (median FKGL 22.7 and 26.1; FRES −11.6 and −7.5). Length did not explain scores within platforms.

Conclusions: Performance varied substantially across platforms and dimensions; no platform optimized all outcomes. These findings describe informational quality, not clinical safety or decision-making, and support specialist verification before clinical use.

1. Introduction

Temporomandibular disorders (TMDs) are prevalent musculoskeletal conditions of the maxillofacial region and a major cause of chronic non-odontogenic orofacial pain [1]. Most patients can be managed with patient education, physiotherapy, occlusal appliances, pharmacotherapy, and, when indicated, minimally invasive procedures such as arthrocentesis or arthroscopy [2,3]. Open temporomandibular joint (TMJ) surgery is generally reserved for selected cases in which nonsurgical and minimally invasive approaches are ineffective or inappropriate, including ankylosis, refractory intra-articular disease, severe degenerative destruction, neoplasia, and end-stage pathology requiring reconstruction [3,4]. Given the complexity and potential irreversibility of these procedures, appropriate diagnosis, imaging, and patient selection are essential, and decision-making must balance joint-preserving and reconstructive options against procedure-specific risks such as facial nerve injury and complications of total joint replacement [5,6,5,6].

Large language models (LLMs) have rapidly emerged as tools for information retrieval, education, and potential clinical decision support across medicine and surgery [7,8]. In dentistry and oral and maxillofacial surgery, LLMs have been increasingly evaluated for answering clinical questions, supporting patient education, and facilitating knowledge synthesis [9,10]. However, fluent and authoritative-sounding responses do not necessarily indicate clinical validity. LLMs may generate inaccurate or misleading information, omit clinically important qualifications, or provide absent or unverifiable sources [11,12]. These limitations are

*Corresponding author.
E-mail addresses: sgas@gelisim.edu.tr (S. Gaş), erdincsulukan@beykent.edu.tr (E. Sulukan), bkorkmaz@gelisim.edu.tr (B. Korkmaz).

https://doi.org/10.1016/j.jormas.2026.102972
Received 21 August 2026; Accepted 30 August 2026
Available online 2 September 2026
2468-7855/© 2026 Elsevier Masson SAS. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

particularly relevant to open TMJ surgery, where inaccurate information on indications, imaging, procedure selection, or complications could adversely influence decision-making, so rigorous, procedure-specific evaluation is necessary.

A growing body of research has evaluated the accuracy, quality, and reliability of LLM-generated responses across dental disciplines, including orthodontics and oral and maxillofacial surgery, demonstrating that performance varies across models, prompts, and clinical topics [13,14]. TMD-specific studies have similarly reported promising but inconsistent performance; however, these investigations have focused primarily on general TMD knowledge or patient-oriented education rather than clinician-oriented questions concerning open TMJ surgery [15,16]. Consequently, it remains unclear whether LLM-generated information on surgical indications, imaging, procedure selection, and complications is consistent with contemporary guidance, and benchmarking against an explicit, contemporary reference standard is necessary to evaluate its potential utility in this context [12,17,12,17].

Contemporary LLMs differ in architecture, training, and update cycles, and comparative studies in dentistry indicate that their clinical performance should not be assumed to be equivalent [7,16,18]. Clinicians, however, do not interact with isolated model architectures but with commercial chatbot platforms that combine a model with platform-specific retrieval, citation and safety features. To our knowledge, no previous study has systematically compared multiple contemporary chatbot platforms using guideline-based, clinician-oriented questions specifically addressing open TMJ surgery. Therefore, this study evaluated and compared the quality and readability of responses generated by five chatbot platforms to clinician-oriented questions on open TMJ surgery, using guideline-based reference standard answers as the benchmark. We hypothesized that response quality and readability would differ significantly among the evaluated platforms.

2. Material and methods

2.1. Question development and categorization

The standardized question set was developed based on the American Association of Oral and Maxillofacial Surgeons (AAOMS) guidance document Management of Patients With Orofacial Pain and Temporomandibular Disorders [19]. Two oral and maxillofacial surgeons independently developed clinician-oriented questions on open temporomandibular joint (TMJ) surgery from the guideline's recommendations; the sets were then reviewed jointly and differences in selection, wording, and scope were resolved by consensus, yielding a final standardized set of 40 questions.

The questions covered eight domains: six topic-based clinical domains—indications and patient selection; diagnostic and imaging methods; surgical techniques and methods; complications and risks; long-term outcomes and follow-up; and ethics and informed consent—and two reasoning-based domains comprising case-based clinical scenarios and integrative clinical reasoning that required multistep decision-making. Sixteen representative questions, two from each domain, are presented in Table 1.

2.2. AI-generated answer collection and expert evaluation

Responses were collected on 16 July 2026 from five chatbot platforms, each accessed through its paid subscription tier: ChatGPT (GPT-5.5; ChatGPT Plus, OpenAI), Claude (Claude Opus 4.8; Claude Pro, Anthropic), Gemini (Gemini 3.1 Pro; Google AI Pro, Google DeepMind), Grok (Grok 4; SuperGrok, xAI) and Perplexity (Perplexity Pro, default model; Perplexity AI) (Table 2). All 40 questions were submitted to every platform on the same day to minimize the influence of platform updates during data collection, yielding 200 responses.

Each platform was used with its default settings. No system instructions, custom instructions, memory or personalization features, role assignment, response-length constraints, or output-format requests were applied. Each question was entered verbatim as the first message of a newly initiated, empty chat session; no follow-up, regeneration, or corrective prompts were used, and only the first response was retained. None of the 200 responses was refused, truncated, or accompanied by a safety warning. Under default settings, web search is available on all five platforms and is invoked automatically at the platform's discretion, whereas Perplexity is a retrieval-augmented system in which web search and citation are integral to every response. Decoding parameters such as temperature and top-p are not user-configurable in these consumer interfaces, so each platform was used with its undisclosed default sampling settings. The evaluation therefore characterizes the consumer-facing products as encountered by clinicians rather than the intrinsic capabilities of the underlying language models.

Two oral and maxillofacial surgeons with clinical experience in TMJ disorders and open joint surgery independently evaluated all responses. Before evaluation, the responses were coded, randomized, and presented without model identifiers to ensure blinding to the generating LLM. The evaluators completed their assessments independently and without discussion. Rather than reconciling scores through consensus, the two evaluators' ratings for each response were averaged to produce a single response-level value for each outcome, which was used in all subsequent analyses (40 responses per model; 200 in total).

The reported model names and versions correspond to the platform labels displayed on 16 July 2026; because commercial platforms are updated without full disclosure, the findings represent a time-specific evaluation.

2.3. Outcome measures

Response quality was assessed with three complementary instruments: the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool, the Global Quality Score (GQS), and a five-point overall quality rating. Each instrument was applied independently by the two blinded evaluators to all 200 responses, and the mean of the two evaluators' scores was used as the response-level value in the analyses. Averaging was chosen a priori because the unit of analysis was the response rather than the rater. Inter-rater reliability was assessed for each outcome measure (the six QAMAI items, the QAMAI total, GQS, and overall quality rating) with the intraclass correlation coefficient (ICC), estimated with a two-way mixed-effects model based on consistency. Both single-measure ICC(3,1) and average-measure ICC(3,2) values are reported with 95% confidence intervals; ICC(3,1) quantifies agreement between the individual raters and preserves information on rater variability, whereas ICC (3,2) represents the reliability of the averaged score used in the analyses. ICC values were interpreted as poor (<0.50), moderate (0.50−0.75), good (0.75−0.90), and excellent (>0.90) [20].

The QAMAI is a validated instrument for assessing the quality of AI-generated health information [21]; its six items (accuracy, clarity, relevance, completeness, source provision, and usefulness) are each rated from 1 to 5, giving a total of 6−30.

The GQS is a five-point global rating of the quality, flow, and usefulness of health information (1 = poor, 5 = excellent) [22].

The five-point overall quality rating captured each evaluator's holistic clinical appraisal, from 1 (very poor) to 5 (excellent).

A guideline-based reference standard answer was developed for each question by the two surgeons from the AAOMS guidance document [19] and agreed by consensus. The key elements expected for each question, their weighting, and the safety-critical and decision-dependent conditions were defined a priori from the AAOMS guidance and are provided in Supplementary Table S2. Written scoring anchors linking key-element coverage to each QAMAI item, the GQS, and the overall quality rating (for example, completeness: ≥90% of weighted key elements = 5, 70−89% = 4, 50−69% = 3, 25−49% = 2, <25% = 1) were provided to both evaluators before scoring (Supplementary Table S3), together with general principles governing synonym acceptance, partial credit,

Table 1
Representative clinician-oriented questions on open temporomandibular joint surgery, guideline-based reference standard answers, and responses generated by five chatbot platforms.

Domain	Question	ChatGPT-5.5 Plus (OpenAI)	Claude Opus 4.8 (Anthropic)	Gemini 3.1 Pro (Google DeepMind)	Grok 4 SuperGrok (xAI)	Perplexity Pro (Perplexity AI)	Reference standard answer
Indications and patient selection	**Q1 What are the definite indications for open TMJ arthroplasty within open joint surgery?**	Severe pain and loss of function resistant to conservative treatment, ankylosis, advanced degeneration, severe trauma sequelae, tumors, and surgically correctable structural joint pathologies are the main definite indications.	Bony/fibrous ankylosis, advanced degenerative joint disease (Wilkes IV−V) with refractory pain/dysfunction, condylar resorption or avascular necrosis, neoplasms (osteochondroma, chondromatosis, synovial pathologies), irreducible chronic dislocation, trauma sequelae (malunion/nonunion), and structural pathologies unresponsive to conservative plus minimally invasive treatment.	The definite indications for open TMJ arthroplasty are advanced internal derangements unresponsive to conservative treatment (Wilkes stages IV and V), fibrous or bony ankylosis, neoplasms or tumor-like lesions involving the joint, growth anomalies (e.g., condylar hyperplasia), and chronic recurrent luxations that cannot be reduced by conservative methods.	Advanced degeneration resistant to conservative treatment, ankylosis, multiple failed surgeries, tumor, and severe structural deformity.	Fibrous or osseous ankylosis, neoplasms (especially chondroma and chondrosarcoma), developmental disorders such as condylar hyperplasia, and chronic-recurrent dislocations are among the classic open arthrotomy indications.	Open surgery (arthroplasty) is indicated when intraarticular pathology (IPD: synovitis, chondromalacia, adhesions, disc morphology/position disorders) causes pain and loss of function unresponsive to conservative and minimally invasive treatment. Advanced osteoarthritis, ankylosis, condylar resorption, neoplasia/tumor-like lesions and irreducible chronic dislocation appear in the surgical algorithms. Imaging findings must be concordant with symptoms. (Ch. 5 & 7, Surgical Algorithms; Ch. 9 [19])
	Q3 What is the profile of a poor candidate for open joint surgery?	Patients without evident structural pathology, whose pain is predominantly due to central sensitization or psychosocial causes, who have unrealistic expectations, and who have severe uncontrolled systemic disease are poor candidates.	Patients with imaging-symptom mismatch, widespread pain syndrome/central sensitization (fibromyalgia), uncontrolled psychiatric illness, opioid dependence, uncontrolled parafunction-bruxism, unrealistic expectations, treatment non-compliance (physiotherapy non-adherence), active infection, uncontrolled systemic disease, and heavy smoking.	Patients with uncontrolled systemic disease, active local or systemic infection, unrealistic surgical expectations, severe psychosocial comorbidities (somatization, major depression, etc.), and chronic neuropathic or fibromyalgia-related widespread pain disorder are very poor candidates for open joint surgery.	Uncontrolled psychiatric illness, widespread chronic pain syndrome, unrealistic expectations, and non-compliance.	Patients with active infection, serious systemic disease, immunosuppression, severe metal allergy, insufficient bone support, severe compliance problems, or prominent neuropsychiatric disorder are considered poor candidates, particularly for TJR.	Pain predominance unexplained by structural pathology, widespread pain sensitivity/central sensitization, prominent psychosocial comorbidity (depression, anxiety, somatization, catastrophizing), unrealistic expectations, and non-compliance with postoperative physiotherapy are markers of a poor candidate. Cervical- or neuropathic-origin referred pain should also be ruled out before surgery. (Ch. 10, Failure to Make the Correct Diagnosis; Chronic Pain [19])
Diagnostic and imaging methods	**Q8 Why is MRI important in planning open joint surgery?**	MRI is the most important imaging modality for evaluating disc position, disc integrity, inflammation and other soft-tissue pathologies.	It is the only modality that shows disc position and reduction, disc deformation/perforation, effusion, synovitis, retrodiscal inflammation, condylar edema/avascular necrosis and soft-tissue tumors; it determines the choice between disc-preserving surgery and discectomy.	Because MRI is the non-invasive soft-tissue imaging modality that best demonstrates the anatomical position of the temporomandibular disc (anterior displacement, etc.), disc morphology, the condition of the retrodiscal tissues, joint effusion and bone marrow edema, it is critically important in defining surgical margins for soft-tissue pathology.	For demonstrating disc position, soft-tissue pathology, effusion and inflammation.	MRI answers whether the disc will be preserved or a discectomy performed by demonstrating disc position, retrodiscal tissues and intra-articular inflammation.	MRI is the standard modality for evaluating the articular disc, soft tissues and intracapsular effusion; it shows disc position and morphology and static/dynamic closed- and open-mouth images. Gadolinium-enhanced sequences distinguish inflamed synovium/arthropathy from osteomyelitis. It assesses disc displacement, effusion, synovial chondromatosis and osteoarthritis; bone detail is poor. (Ch. 4, Magnetic Resonance Imaging [19])
	Q10 When is CTA required in planning TMJ ankylosis that requires open joint surgery?	CTA is required if there is advanced ankylosis close to large vascular structures or suspicion of complex anatomical deformity.	When the ankylotic mass extends medially/to the skull base, in recurrent ankylosis, in suspected tumor/vascular malformation, and in Sawhney type III−IV cases where the relationship of the maxillary artery, internal carotid and retromandibular vein is uncertain.	CT Angiography (CTA) is required to minimize the risk of vascular injury in severe bony ankylosis cases where the ankylotic mass is very close to or abutting the skull base and major vascular structures such as the internal carotid artery, maxillary artery and internal jugular vein, and where anatomical boundaries are lost.	In ankylosis close to large vessels or in revision surgery.	CTA is required in advanced osseous ankylosis, in patients where vessels may be displaced by deformity and surgical bleeding risk is high.	CTA is ideal for identifying arterial and venous structures in bony TMJ ankylosis and is the initial evaluation for vascular pathology such as AV malformation/hemangioma. It is recommended when a vessel is suspected within the ankylotic mass, when anatomy is distorted by prior surgery, and when massive hemorrhage risk is anticipated. (Ch. 4, Computed Tomographic Angiography; Ch. 9, Ankylosis [19])

(*continued on next page*)

Table 1 (*Continued*)

Domain	Question	ChatGPT-5.5 Plus (OpenAI)	Claude Opus 4.8 (Anthropic)	Gemini 3.1 Pro (Google DeepMind)	Grok 4 SuperGrok (xAI)	Perplexity Pro (Perplexity AI)	Reference standard answer
Surgical techniques and methods	**Q13 When should a disc-sparing approach be used in open joint surgery?**	It should be applied if the disc is in a preservable condition and can be functionally repositioned.	If the disc is morphologically intact, not perforated or severely deformed; in Wilkes II–III stage, in a young patient, and when the condylar surface is largely preserved, repositioning/plication is preferred.	Disc-preserving approaches (e.g., discopexy) should be used in early-to-mid internal derangement (Wilkes stages II and III) where disc morphology is preserved, the disc is not severely deformed or perforated, and it has enough flexibility to be surgically sutured into its anatomical position in the fossa.	If disc reduction is possible and tissue quality is good.	Disc repositioning, discectomy/meniscectomy, gap arthroplasty, interpositional arthroplasty, autogenous graft reconstruction and total joint replacement are the main techniques.	When the superior disc surface is intact and there is no perforation or condylar exposure/cartilage destruction, a disc-sparing incision and repositioning/plication/imbrication (discopexy) is preferred. If the disc is perforated, deformed, irreparable or causing pain/dysfunction, meniscectomy is performed. (Ch. 9, Arthroplasty [19])
	Q15 What is the most critical anatomical risk in patients requiring meniscectomy during open joint surgery?	Facial nerve and especially injury to the medially located maxillary artery are among the most important anatomical risks.	Vascular structures posteromedial to the retrodiscal tissues — especially the maxillary artery and retromandibular vein/pterygoid plexus; and, in medially advancing dissection, the skull base and facial nerve branches.	The most critical anatomical risk during and after meniscectomy is severe hemorrhage from injury to the internal maxillary artery in the deep dissectional area medial to the disc, as well as perforation of the skull base (roof of the glenoid fossa) immediately superior to the joint capsule and injury to the temporal branch of the facial nerve.	Facial nerve (temporal and zygomatic branches) injury.	The most critical risk is injury to the facial nerve branches; the external auditory canal and skull base are also in close proximity.	The most critical risk is perforation of the medial capsule during deep dissection at the medial disc attachment; this leads to serious hemorrhage from injury to the middle meningeal artery, internal maxillary artery or pterygoid venous plexus. The medial disc attachment is the most difficult to remove and requires controlled dissection. (Ch. 9, Meniscectomy [19])
Complications and risks	**Q19 How can facial nerve injury be prevented in open joint surgery?**	Risk can be reduced with dissection consistent with anatomical planes, careful tissue manipulation, and nerve monitoring when needed.	Subfascial dissection via a preauricular/endaural incision (Al-Kayat–Bramley), staying in the periosteal plane over the zygomatic arch, blunt dissection, avoiding excessive retractor traction, using bipolar cautery, and nerve monitoring when needed.	Facial nerve injury can be prevented by performing the surgical incision (e.g., preauricular) and dissection beneath the deep layer of the temporoparietal fascia while adhering to anatomical landmarks such as the Al-Kayat and Bramley modification, avoiding aggressive retraction, and using intraoperative facial nerve monitoring.	By isolating the nerve and using careful retraction.	A correct dissection plane, appropriate incision, avoiding excessive retraction, adherence to anatomical layers, and nerve monitoring when needed reduce the risk.	The temporal (and zygomatic) branch of the facial nerve is most at risk; most injuries result from excessive retraction rather than direct transection. Risk is reduced by subperiosteal dissection in the correct fascial plane, gentle retraction with the smallest retractors, recognition of anatomical landmarks (Al-Kayat), and a nerve stimulator when needed. Inability to raise the eyebrow occurs in 5% of cases, usually resolving within 3 months, and is permanent in <1%. (Ch. 10, Facial Nerve Injury; Ch. 9 [19])
	Q21 How should middle cranial fossa perforation and CSF leak be managed in open joint surgery?	The CSF leak should be repaired, neurosurgery consultation obtained when needed, and the patient closely monitored.	Intraoperative diagnosis; temporal fascia/muscle/fat graft for small defects, fibrin sealant, dural suture if needed; neurosurgery consultation, lumbar drainage if the leak persists, head elevation, avoidance of straining, and close monitoring for meningitis.	When glenoid fossa perforation and CSF leak develop, the leak area should be immediately plugged with a temporalis muscle graft, bone wax or fat graft to create a watertight barrier, postoperative measures to lower intracranial pressure should be taken, prophylactic meningitis antibiotic therapy started, and neurosurgery consultation requested.	With emergent repair, tissue graft and neurosurgery consultation.	The dural defect should be closed with primary repair or appropriate grafts, neurosurgical support obtained if needed, and infection risk reduced.	Small perforations without CSF leak may require no treatment. For large perforations or CSF leak: repair with autogenous cortical graft (or custom TJR), dural repair (temporal fascia, fibrin glue), and ICP reduction with lumbar drainage when needed. Neurosurgery consultation and postoperative non-contrast CT should be considered. In ankylosis surgery, cuts are recommended to be planned at least 5 mm away from the MCF. (Ch. 10, Perforation of the EAC/MCF [19])
Long-term outcomes and follow-up	**Q24 Why is physical therapy important after open joint surgery?**	Physical therapy preserves mouth opening, reduces the risk of fibrosis and re-ankylosis, and supports functional recovery.	It prevents fibrosis and re-ankylosis, preserves mouth opening and translation, restores muscle strength and coordination, and reduces edema and pain. It should be started early and continued for months; it is the most decisive postoperative factor for surgical success.	Postoperative physical therapy is vital to prevent soft-tissue scarring around the joint, to preserve and increase joint range of motion (ROM), to resolve spasm in the masticatory muscles and re-establish functional coordination, and to prevent the joint from progressing to early ankylosis or re-ankylosis.	Mandatory to preserve maximum mouth opening and prevent scar tissue.	Physical therapy preserves range of motion, reduces the risk of adhesion and re-ankylosis, and ensures maintenance of the function gained through surgery.	Aggressive and early joint mobilization is mandatory for the success of any intraarticular joint surgery. Range-of-motion exercises begin on postoperative day 4; the goal is to increase ROM and mobilize surgical edema. Early motion reduces the risk of adhesion, fibrosis and re-ankylosis. (Ch. 9, Post-Operative Considerations [19])

(*continued on next page*)

Table 1 (*Continued*)

Domain	Question	ChatGPT-5.5 Plus (OpenAI)	Claude Opus 4.8 (Anthropic)	Gemini 3.1 Pro (Google DeepMind)	Grok 4 SuperGrok (xAI)	Perplexity Pro (Perplexity AI)	Reference standard answer
	Q26 **How can heterotopic bone and re-ankylosis be reduced after TJR in open joint surgery?**	The risk can be reduced with early motion exercises, appropriate surgical technique, and regular rehabilitation.	Meticulous removal of bone debris and blood, wide resection/gap creation, autogenous fat graft interposition, early and aggressive physiotherapy, additional coronoidectomy, hemostasis control, and NSAIDs in selected cases; radiotherapy is not routine in the TMJ.	To reduce the risk of heterotopic ossification and re-ankylosis, the area around the prosthesis should be filled with a free fat graft (abdominal or buccal) during surgery, aggressive physical therapy started in the early postoperative period (first 24–48 h), and prophylactic non-steroidal anti-inflammatory drugs (e.g., indomethacin) or low-dose localized radiotherapy preferred.	With NSAIDs, radiation and aggressive physical therapy.	Leaving an adequate surgical gap, appropriate interposition, early exercise, and NSAIDs or adjuvant measures in selected protocols can reduce the risk of heterotopic bone and re-ankylosis.	The most common and least problematic approach is to create a critical gap arthroplasty, use an alloplastic TJR and fill the dead space with a fat graft. The fat graft limits differentiation of undifferentiated mesenchymal/pluripotent cells, reducing fibrosis and heterotopic calcification. Evidence for NSAIDs (indomethacin) and low-dose radiotherapy is weak and not routine in the TMJ. (Ch. 10, Heterotopic Bone and Ankylosis [19])
Ethics and informed consent	**Q27** **What should be included in the informed consent form for open joint surgery?**	It should include the diagnosis, treatment options, expected benefits, possible complications, need for rehabilitation, and the possibility of additional surgery.	Diagnosis and alternatives, the planned procedure and its aim, realistic expected gain (pain may not fully resolve, opening may increase only to a limited extent), risks (facial nerve injury, sensory loss, bleeding, infection, EAC/dura injury, malocclusion, re-ankylosis, prosthesis lifespan and revision possibility), the requirement for physiotherapy and diet, the need for long-term follow-up, and the expected course if surgery is not performed.	The consent form should clearly state the details of the surgical procedure to be performed, alternative treatment options, facial nerve injury (risk of permanent paralysis), bleeding, hearing loss or EAC perforation, infection, prosthesis or graft failure, development of malocclusion, the risk of persistent chronic pain, and the need for long-term intensive physical therapy.	Risks, alternatives, success rates and long-term expectations.	Surgery-specific risks such as diagnosis, alternative treatments, expected benefits, nerve injury, infection, bleeding, CSF leak, persistent pain, re-ankylosis, malocclusion and the need for rehabilitation should be included.	Consent should include a description of the procedure, expected benefits, alternative treatments and major complications, together with the possibility of revision/additional surgery. Surgery-specific risks: facial nerve injury, EAC/tympanic membrane injury and hearing loss, MCF perforation/CSF leak, massive hemorrhage, infection and persistent pain. (Ch. 10, Complications [19])
	Q28 **Why should psychosocial comorbidities be specifically emphasized in the consent for open joint surgery?**	It must be stated that psychosocial comorbidities can affect pain perception and surgical satisfaction.	Because pain may persist even if the surgery is technically successful; the expectation that "surgery will end my pain" is unrealistic. Including this explicitly in the consent is necessary for expectation management, legal protection and ethical transparency.	Patients with psychosocial comorbidity may continue to feel pain despite anatomical success due to central sensitization; the consent should clarify that this is not a surgical failure but a disorder of the nervous system's pain processing, and expectations should be set on a realistic basis to reduce legal and clinical risks.	Because of the potential to adversely affect surgical success and patient satisfaction.	Because these conditions can markedly affect postoperative pain perception, expectations, compliance and satisfaction.	Despite an accurate diagnosis and technically successful surgery, a subgroup of patients may have persistent pain through central sensitization/neuropathic mechanisms. This should be clearly stated in the consent; the expectation that "surgery will completely eliminate pain" must be placed on a realistic basis. It is important for expectation management and to avoid patient abandonment. (Ch. 10, Chronic Pain; Biopsychosocial Appendix [19])
Case-based clinical scenarios	**Q30** **What is the strongest next option in a patient with persistent pain and severe dietary restriction after two failed open joint surgeries?**	If advanced structural damage is confirmed, total joint replacement is the strongest next surgical option.	If the pain source is confirmed to be structural, alloplastic total joint replacement, preferably a patient-specific prosthesis; this should be accompanied by early aggressive physiotherapy and fat graft interposition. Myofascial/central pain and psychosocial factors should be excluded beforehand.	In this patient who has undergone two failed open joint surgeries, joint anatomy and biomechanics will be severely disrupted, so the strongest and most predictable next option is complete resection of the damaged joint structures and replacement with a custom-made Total Joint Prosthesis (TJR).	Total joint replacement (TJR).	If end-stage damage or severe loss of function is confirmed clinically and radiologically, total TMJ replacement is the strongest next option.	In a multiply-operated joint with advanced structural damage, alloplastic total joint replacement (preferably a patient-specific prosthesis) is the most stable and predictable option; it is performed with a critical gap and fat graft interposition. However, if the pain is of myofascial/central/neuropathic origin, these must be excluded first, otherwise the outcome is poor. (Ch. 9, Total Joint Replacement; Ch. 10, Chronic Pain [19])
	Q31 **What steps are mandatory in the preoperative planning of a patient with post-traumatic bony ankylosis requiring open joint surgery?**	Detailed clinical examination, CT/CBCT, CTA if needed, occlusion analysis and rehabilitation planning are mandatory.	Airway assessment and fiberoptic intubation plan, 3D mapping of the ankylotic mass with thin-slice CBCT/CT (Sawhney classification), CTA when needed, evaluation of the coronoid and contralateral joint, analysis of occlusion-ramus height and dentofacial deformity, virtual surgical planning/cutting guides, blood preparation, graft/prosthesis selection, and pre-planning the physiotherapy program.	Preoperatively, it is mandatory to clarify the relationship of the ankylotic mass with the skull base and adjacent vascular structures using three-dimensional CT/CBCT, to obtain CTA if needed, to define osteotomy lines with virtual surgical planning (VSP), to produce a stereolithographic model, and to have an aggressive postoperative physical therapy protocol ready.	Detailed CBCT/MRI, CTA (if needed), psychosocial evaluation and prosthesis planning.	It is mandatory to define the ankylotic mass with CBCT/CT, the vascular anatomy with CTA if needed, airway planning, possible coronoidectomy need, and functional goals.	Mandatory steps: mapping the ankylotic mass and resection boundaries with axial/coronal/sagittal CT; CTA when a vessel is suspected; virtual surgical planning (VSP), stereolithographic model, cutting guides and navigation; and a safe airway plan coordinated with anesthesia (often fiberoptic nasoendotracheal intubation, with tracheostomy preparation if needed). (Ch. 9, Ankylosis / Virtual Surgical Planning [19])

(*continued on next page*)

Table 1 (*Continued*)

Domain	Question	ChatGPT-5.5 Plus (OpenAI)	Claude Opus 4.8 (Anthropic)	Gemini 3.1 Pro (Google DeepMind)	Grok 4 SuperGrok (xAI)	Perplexity Pro (Perplexity AI)	Reference standard answer
Integrative clinical reasoning	**Q37 Why must cervical-origin or neuropathic pain be ruled out in a patient scheduled for open joint surgery?**	Because these pain sources can mimic TMJ disease and lead to unnecessary surgery.	These pains mimic TMJ pain; if the source is not in the joint, surgery is both useless and may add new injury/neuropathy and lead to dissatisfaction. Exclusion is essential for correct indication, realistic expectations and avoiding unnecessary surgery.	Because cervical spine pathologies and neuropathic pain produce pain referred directly to the TMJ region; if these primary sources are not eliminated before surgery, the patient will undergo unnecessary invasive open joint surgery and the referred pain character will not change at all postoperatively.	They must be ruled out because they can lead to surgical failure.	Because these pains can mimic TMJ-origin pain and result in a wrong surgical indication, causing persistent postoperative pain.	Cervical pain is one of the most common sources of pain referred to the face and jaws; conditions such as migraine, trigeminal neuralgia and cervicogenic headache mimic TMJ pain. Failure to recognize referred pain leads to unnecessary TMJ surgery with poor outcomes. All cervical pain should be managed before surgery (Spurling test, referral to physical therapy/physiatry) and the patient reassessed. (Ch. 10, Failure to Make the Correct Diagnosis [19])
	Q39 What are the possible causes of new-onset malocclusion after total joint replacement in open joint surgery, and how should the initial assessment be performed?	Prosthesis position, screw stability, condylar seating and occlusion should be evaluated clinically and radiologically.	Causes: intraoperative prosthesis malposition, inadequate intermaxillary fixation or its early release, screw loosening/component migration, edema-hematoma, muscle spasm, contralateral joint pathology, heterotopic bone, or pre-existing occlusal discrepancy. Initial assessment: clinical occlusion examination and comparison with preoperative records, CBCT/CT showing prosthesis position, fixation integrity check; elastic guidance and physiotherapy in mild cases, revision if there is a structural error.	Malocclusion may develop from loosening or fracture of prosthesis screws, displacement of prosthesis components, acute degeneration/condylar resorption in the contralateral joint, or postoperative muscle spasm; the initial assessment should check occlusal contacts and confirm prosthesis integrity with panoramic or CBCT imaging.	Due to prosthesis position, condylar height or muscle imbalance; clinical and radiological evaluation.	Component malposition, screw loosening, incorrect height restoration or contralateral joint/muscle adaptation problems may be the cause; clinical occlusion analysis and imaging should be performed at the first stage.	Possible causes: intraoperative component malposition, change in seating of the condyle in the fossa, early release/inadequacy of MMF, screw loosening/component migration, postoperative edema-muscle imbalance, and contralateral joint pathology. Initial assessment: detailed clinical occlusion examination and comparison with preoperative records, CBCT/CT showing component position; guiding elastics in mild cases, revision for structural error. (Ch. 9, Total Joint Replacement; Ch. 10 [19])

Two representative questions are shown for each of the eight clinical domains; question numbers correspond to the complete set of 40 questions and all corresponding responses provided as Supplementary Material. Reference standard answers were developed from the AAOMS guidance document [19]; parenthetical codes denote the supporting chapter and section heading. AI, artificial intelligence.

Table 2
Chatbot platforms evaluated (all queries submitted on 16 July 2026).

Platform	Model version	Subscription tier	Web search / retrieval	Settings
ChatGPT (OpenAI)	GPT-5.5	ChatGPT Plus	Default; invoked automatically by the platform	Default; no custom instructions
Claude (Anthropic)	Claude Opus 4.8	Claude Pro	Default; invoked automatically by the platform	Default; no custom instructions
Gemini (Google DeepMind)	Gemini 3.1 Pro	Google AI Pro	Default; invoked automatically by the platform	Default; no custom instructions
Grok (xAI)	Grok 4	SuperGrok	Default; invoked automatically by the platform	Default; no custom instructions
Perplexity (Perplexity AI)	Perplexity Pro (default model)	Perplexity Pro	Always on (retrieval-augmented by design; citations appended to every response)	Default; no custom instructions

Session protocol for all platforms: new empty chat per question, question entered verbatim, first response retained. Decoding parameters (temperature, top-p) are not user-configurable and were left at platform defaults.

incorrect statements, decision questions, and list padding (Supplementary Table S4). Representative questions, reference answers, and platform responses are presented in Table 1.

2.4. Outcome assessment and statistical analysis

Readability was quantified with the Flesch Reading Ease Score (FRES) and Flesch-Kincaid Grade Level (FKGL), computed with the textstat library (version 0.7.13, Python 3.12) from the raw response text after removal of URLs and citation strings. Higher FRES values indicate greater reading ease, whereas higher FKGL values indicate a higher grade level. The FRES formula has no lower bound and yields negative values for highly complex text; no floor, truncation, or other post-processing was applied. FRES >60 and FKGL <8 were used as general benchmarks, and readability was interpreted descriptively and independently from clinical quality.

Because the same 40 questions were submitted to all five platforms, the data have a repeated-measures structure in which each question constitutes a block with five related observations. Between-platform comparisons were therefore performed using the Friedman test (k = 5 related samples, n = 40 blocks). The score analyzed for each question-platform pair was the mean of the two evaluators' ratings (QAMAI items, QAMAI total, GQS, and overall quality rating); readability indices were single values per response. Kendall's W was calculated as the effect size for the Friedman test (0.1 small, 0.3 moderate, 0.5 large). When the Friedman test was significant, pairwise post-hoc comparisons were performed using the Wilcoxon signed-rank test with Bonferroni correction for the 10 pairwise comparisons within each outcome (adjusted $\alpha = 0.005$; reported P values are Bonferroni-adjusted), and the effect size of each pairwise comparison was expressed as $r = Z/\sqrt{n}$ (0.1 small, 0.3 medium, 0.5 large). To examine whether response length influenced scoring, the word count of each response (after removal of URLs) was correlated with each outcome using Spearman's ρ, both across all 200 responses and within each platform. Descriptive statistics are reported as mean ± standard deviation and median (interquartile range). All tests were two-sided, with statistical significance set at $P < .05$. Analyses were performed in Python 3.12 (SciPy, statsmodels, pingouin 0.6).

3. Results

A total of 200 responses were generated, with 40 responses from each of the five platforms. Inter-rater reliability was good to excellent for all outcome measures (Table 5). Across the 200 pooled responses, single-measure ICC(3,1) ranged from 0.863 (95% CI 0.82−0.89) for completeness to 0.970 (95% CI 0.96−0.98) for the QAMAI total score, and average-measure ICC(3,2) ranged from 0.927 (95% CI 0.90−0.94) to 0.985 (95% CI 0.98−0.99). Platform-specific ICCs (Supplementary Table S1) were ≥0.70 for single measures and ≥0.83 for average measures; for the source-provision item of Perplexity, the two evaluators assigned identical scores to every response (ICC = 1.00; confidence interval not estimable).

The Friedman test showed significant between-platform differences for every outcome except clarity (Table 3). The largest effect was observed for the QAMAI total score ($\chi^2(4) = 119.4$, $P < .001$, Kendall's W = 0.75): Perplexity obtained the highest scores (27.52 ± 1.43; median 27.5, IQR 26.5−29.0), followed by Claude (25.45 ± 1.49), Gemini (25.14 ± 1.17), ChatGPT (23.04 ± 2.38), and Grok (21.60 ± 1.95). Bonferroni-adjusted Wilcoxon signed-rank tests showed that all pairwise differences in total score were significant ($P \leq .007$, r = 0.53−0.87) except Claude versus Gemini ($P = 1.00$) (Table 4).

Perplexity's advantage was driven mainly by source provision ($\chi^2(4) = 107.4$, $P < .001$, W = 0.67), in which it scored higher than every other platform (4.90 ± 0.30 versus 2.41−2.83; all $P < .001$, r = 0.89−0.91), whereas the other four platforms did not differ from one another except Gemini versus Grok ($P = .011$). When the source-provision item was excluded, the median five-item QAMAI score was 23.0 for Claude, 22.8 for Perplexity, 22.0 for Gemini, 20.3 for ChatGPT, and 19.0 for Grok, indicating that Perplexity's overall lead was attributable mainly to source provision, whereas the lower scores of ChatGPT and Grok persisted. For accuracy ($\chi^2(4) = 101.0$, $P < .001$, W = 0.63), Claude (4.41 ± 0.55), Gemini (4.42 ± 0.57), and Perplexity (4.54 ± 0.50) did not differ from each other but all scored higher than ChatGPT (3.75 ± 0.60; $P \leq .003$) and Grok (2.92 ± 0.62; $P < .001$); ChatGPT also scored higher than Grok ($P < .001$, r = 0.75). Grok obtained the lowest scores for relevance, usefulness, GQS, and overall quality (all comparisons against Grok P < .001). Claude achieved the highest usefulness (4.54 ± 0.51; $\chi^2(4) = 72.5$, $P < .001$, W = 0.45) and GQS (4.51 ± 0.58; $\chi^2(4) = 73.9$, $P < .001$, W = 0.46), differing significantly from ChatGPT, Gemini, and Grok ($P \leq .038$) but not from Perplexity (4.31 ± 0.53). For the five-point overall quality rating ($\chi^2(4) = 99.1$, $P < .001$, W = 0.62), Gemini (4.29 ± 0.71) and Claude (4.17 ± 0.66) scored higher than ChatGPT (3.64 ± 0.51; $P \leq .001$) and Grok (2.40 ± 0.51; $P < .001$), and Gemini also scored higher than Perplexity (3.76 ± 0.64; $P = .022$). Completeness differed mainly because ChatGPT scored lower than the other four platforms ($P \leq .023$). Clarity did not differ between platforms ($\chi^2(4) = 5.04$, $P = .283$, W = 0.03).

Readability differed significantly between platforms (FKGL: $\chi^2(4) = 92.3$, $P < .001$, W = 0.58; FRES: $\chi^2(4) = 41.0$, $P < .001$, W = 0.26; Table 3). Claude (median FKGL 22.7, FRES −11.6) and Gemini (FKGL 26.1, FRES −7.5) produced the most complex text and were significantly harder to read than ChatGPT (FKGL 16.4, FRES 10.0), Grok (14.1, 10.1), and Perplexity (17.9, 10.2) (all $P \leq .037$); Claude and Gemini did not differ from each other, nor did ChatGPT, Grok, and Perplexity on FRES. All platforms exceeded the benchmarks for lay readers.

Response length differed markedly between platforms (mean 9.4 words for Grok, 17.3 for ChatGPT, 22.8 for Perplexity, 37.1 for Claude, and 47.4 for Gemini). Across all 200 responses, word count correlated moderately with the QAMAI total (Spearman $\rho = 0.50$), but this was largely explained by platform identity: within ChatGPT, Claude, and Gemini the correlation was weak and non-significant ($\rho = 0.17$, 0.10, and 0.15), whereas within Grok ($\rho = 0.49$) and Perplexity ($\rho = 0.63$) shorter responses omitted guideline key elements and were scored lower, primarily on accuracy and usefulness. Completeness correlated only weakly with word count (pooled $\rho = 0.14$), and Gemini, the longest-answering platform, did not obtain the highest scores (Supplementary Table S5).

Table 3
Quality and readability measures according to chatbot platform (Friedman test).

Measure	Platform	Mean ± SD	Median (IQR)	Mean rank	χ^2 (df = 4)	*P* value	Kendall's W
QAMAI total	ChatGPT-5.5 Plus	23.04 ± 2.38	23.0 (21.4–25.0)	2.04	119.36	<.001	0.746
	Claude Opus 4.8	25.45 ± 1.49	25.5 (24.5–26.5)	3.65			
	Gemini 3.1 Pro	25.14 ± 1.17	25.0 (24.8–26.0)	3.21			
	Grok 4 (SuperGrok)	21.60 ± 1.95	21.8 (20.4–23.0)	1.34			
	Perplexity Pro	27.52 ± 1.43	27.5 (26.5–29.0)	4.76			
QAMAI accuracy	ChatGPT-5.5 Plus	3.75 ± 0.60	4.0 (3.0–4.0)	2.42	101.02	<.001	0.631
	Claude Opus 4.8	4.41 ± 0.55	4.0 (4.0–5.0)	3.71			
	Gemini 3.1 Pro	4.42 ± 0.57	4.5 (4.0–5.0)	3.74			
	Grok 4 (SuperGrok)	2.92 ± 0.62	3.0 (2.9–3.0)	1.23			
	Perplexity Pro	4.54 ± 0.50	5.0 (4.0–5.0)	3.90			
QAMAI clarity	ChatGPT-5.5 Plus	4.21 ± 0.63	4.0 (4.0–5.0)	2.71	5.04	.283	0.031
	Claude Opus 4.8	4.47 ± 0.52	4.5 (4.0–5.0)	3.27			
	Gemini 3.1 Pro	4.36 ± 0.48	4.0 (4.0–5.0)	2.95			
	Grok 4 (SuperGrok)	4.29 ± 0.62	4.0 (4.0–5.0)	2.86			
	Perplexity Pro	4.46 ± 0.50	4.0 (4.0–5.0)	3.20			
QAMAI relevance	ChatGPT-5.5 Plus	4.49 ± 0.52	4.5 (4.0–5.0)	3.08	40.73	<.001	0.255
	Claude Opus 4.8	4.65 ± 0.47	5.0 (4.0–5.0)	3.45			
	Gemini 3.1 Pro	4.58 ± 0.49	5.0 (4.0–5.0)	3.24			
	Grok 4 (SuperGrok)	3.91 ± 0.60	4.0 (3.9–4.0)	1.85			
	Perplexity Pro	4.62 ± 0.48	5.0 (4.0–5.0)	3.39			
QAMAI completeness	ChatGPT-5.5 Plus	4.29 ± 0.71	4.0 (4.0–5.0)	2.12	32.85	<.001	0.205
	Claude Opus 4.8	4.67 ± 0.46	5.0 (4.0–5.0)	2.96			
	Gemini 3.1 Pro	4.84 ± 0.40	5.0 (5.0–5.0)	3.41			
	Grok 4 (SuperGrok)	4.78 ± 0.39	5.0 (4.5–5.0)	3.21			
	Perplexity Pro	4.80 ± 0.39	5.0 (5.0–5.0)	3.29			
QAMAI source provision	ChatGPT-5.5 Plus	2.62 ± 0.55	3.0 (2.0–3.0)	2.44	107.40	<.001	0.671
	Claude Opus 4.8	2.70 ± 0.54	3.0 (2.0–3.0)	2.56			
	Gemini 3.1 Pro	2.83 ± 0.40	3.0 (3.0–3.0)	2.90			
	Grok 4 (SuperGrok)	2.41 ± 0.54	2.0 (2.0–3.0)	2.10			
	Perplexity Pro	4.90 ± 0.30	5.0 (5.0–5.0)	5.00			
QAMAI usefulness	ChatGPT-5.5 Plus	3.67 ± 0.56	4.0 (3.0–4.0)	2.40	72.47	<.001	0.453
	Claude Opus 4.8	4.54 ± 0.51	5.0 (4.0–5.0)	4.16			
	Gemini 3.1 Pro	4.11 ± 0.63	4.0 (4.0–4.2)	3.24			
	Grok 4 (SuperGrok)	3.29 ± 0.59	3.0 (3.0–4.0)	1.71			
	Perplexity Pro	4.20 ± 0.58	4.0 (4.0–5.0)	3.49			
GQS	ChatGPT-5.5 Plus	4.21 ± 0.58	4.0 (4.0–5.0)	3.29	73.93	<.001	0.462
	Claude Opus 4.8	4.51 ± 0.58	5.0 (4.0–5.0)	3.88			
	Gemini 3.1 Pro	4.03 ± 0.66	4.0 (4.0–4.1)	2.91			
	Grok 4 (SuperGrok)	3.14 ± 0.62	3.0 (3.0–3.6)	1.41			
	Perplexity Pro	4.31 ± 0.53	4.0 (4.0–5.0)	3.51			
Overall quality	ChatGPT-5.5 Plus	3.64 ± 0.51	4.0 (3.0–4.0)	2.89	99.10	<.001	0.619
	Claude Opus 4.8	4.17 ± 0.66	4.0 (4.0–5.0)	3.76			
	Gemini 3.1 Pro	4.29 ± 0.71	4.0 (4.0–5.0)	4.12			
	Grok 4 (SuperGrok)	2.40 ± 0.51	2.0 (2.0–3.0)	1.11			
	Perplexity Pro	3.76 ± 0.64	4.0 (3.0–4.0)	3.11			
FKGL	ChatGPT-5.5 Plus	16.31 ± 3.87	16.4 (14.6–19.2)	2.20	92.30	<.001	0.577
	Claude Opus 4.8	22.88 ± 5.15	22.7 (18.9–26.6)	4.03			
	Gemini 3.1 Pro	24.75 ± 4.61	26.1 (20.7–27.9)	4.35			
	Grok 4 (SuperGrok)	14.53 ± 4.62	14.1 (11.2–18.7)	1.50			
	Perplexity Pro	18.11 ± 3.31	17.9 (15.4–20.2)	2.92			
FRES	ChatGPT-5.5 Plus	8.91 ± 25.86	10.0 (−10.5–23.0)	3.77	40.98	<.001	0.256
	Claude Opus 4.8	−14.15 ± 27.04	−11.6 (−27.8–3.5)	1.95			
	Gemini 3.1 Pro	−9.68 ± 16.72	−7.5 (−21.2–0.1)	2.35			
	Grok 4 (SuperGrok)	7.61 ± 33.04	10.1 (−14.5–34.6)	3.40			
	Perplexity Pro	5.90 ± 25.26	10.2 (−11.5–24.4)	3.52			

QAMAI, Quality Analysis of Medical Artificial Intelligence; GQS, Global Quality Score; FKGL, Flesch-Kincaid Grade Level; FRES, Flesch Reading Ease Score; IQR, interquartile range; SD, standard deviation; χ^2, Friedman test statistic; W, Kendall's coefficient of concordance (effect size). Quality scores are the mean of the two evaluators (n = 40 questions per platform). Readability indices were computed with textstat from the raw response text without truncation; negative FRES values indicate text more complex than the nominal 0–100 range. Pairwise post-hoc comparisons are given in Table 4.

4. Discussion

This guideline-based evaluation demonstrated substantial differences among five contemporary chatbot platforms in responding to clinician-oriented questions on open TMJ surgery. Perplexity achieved the highest QAMAI total score, an advantage attributable largely to its retrieval-based source provision; Claude achieved the highest GQS and usefulness scores, and Gemini the highest five-point overall quality rating. Claude and Gemini produced the least readable text. Grok consistently received the lowest scores across the three quality measures. The large effect sizes observed for QAMAI, GQS, and overall quality (Kendall's W 0.46–0.75) indicate that platform choice materially influenced the quality of the generated clinical information. Importantly, no single platform performed best across all evaluation measures, highlighting the multidimensional nature of performance in this clinically complex domain.

The divergence in platform rankings across QAMAI, GQS, and overall quality is clinically relevant because these measures capture different aspects of response quality: QAMAI evaluates six distinct components [21], GQS provides a global assessment of quality, flow, and usefulness

Table 4
Bonferroni-adjusted pairwise post-hoc comparisons between platforms (Wilcoxon signed-rank test).

Comparison	QAMAI total	Accuracy	Source provision	Usefulness	GQS	Overall quality
ChatGPT vs Claude	<.001 (0.78)	<.001 (0.71)	1.000 (0.17)	<.001 (0.79)	.033 (0.46)	<.001 (0.64)
ChatGPT vs Gemini	<.001 (0.65)	.003 (0.58)	.402 (0.32)	.033 (0.47)	1.000 (0.21)	<.001 (0.67)
ChatGPT vs Grok	.007 (0.53)	<.001 (0.75)	1.000 (0.26)	.008 (0.53)	<.001 (0.79)	<.001 (0.88)
ChatGPT vs Perplexity	<.001 (0.87)	<.001 (0.77)	<.001 (0.89)	.002 (0.60)	1.000 (0.15)	1.000 (0.14)
Claude vs Gemini	1.000 (0.20)	1.000 (0.01)	1.000 (0.19)	.038 (0.46)	.033 (0.46)	1.000 (0.11)
Claude vs Grok	<.001 (0.87)	<.001 (0.87)	.132 (0.39)	<.001 (0.82)	<.001 (0.83)	<.001 (0.86)
Claude vs Perplexity	<.001 (0.80)	1.000 (0.17)	<.001 (0.89)	.128 (0.39)	.535 (0.31)	.112 (0.40)
Gemini vs Grok	<.001 (0.85)	<.001 (0.83)	.011 (0.51)	<.001 (0.67)	<.001 (0.72)	<.001 (0.88)
Gemini vs Perplexity	<.001 (0.83)	1.000 (0.15)	<.001 (0.91)	1.000 (0.09)	.276 (0.35)	.022 (0.48)
Grok vs Perplexity	<.001 (0.87)	<.001 (0.87)	<.001 (0.89)	<.001 (0.75)	<.001 (0.82)	<.001 (0.86)

Values are Bonferroni-adjusted *P* values (effect size $r = Z/\sqrt{n}$; 0.1 small, 0.3 medium, 0.5 large). Post-hoc comparisons for relevance, completeness and the readability indices, and uncorrected *P* values, are provided in Supplementary Table S6. Clarity was not tested post hoc because the Friedman test was not significant.

[22], and the five-point rating reflected the evaluators' holistic clinical appraisal. Strength in one dimension therefore does not necessarily translate into superior performance across others, and these complementary measures should be interpreted together rather than relying on a single score.

Our findings should be interpreted alongside recent dental studies showing that LLM performance is model-, prompt-, task-, and evaluation-dependent rather than a fixed property of a given system [9,10,18,23–27]. Hassanein et al. [28] evaluated a single multimodal model (ChatGPT-5.1) on 300 biopsy-confirmed oral lesion vignettes under three prompting strategies: prompt structure was an independent determinant of diagnostic accuracy in multivariable analysis (consistency-optimized versus context-only prompt, $\beta = +0.167$, $P = .002$), the best prompt reached 72% top-1 accuracy compared with 80% for expert clinicians, calibration and explanation quality differed significantly between prompts ($P < .0001$), and the authors proposed a Prompt Performance Index to quantify stability while noting that prompt sensitivity may be model-specific. The same group [29] compared ChatGPT-4 and DeepSeek-3 as rubric-based graders of 2358 clinical short-answer responses against three calibrated experts and found markedly different agreement (ICC 0.64 versus 0.87), a response-style-dependent error pattern in ChatGPT-4 (increasing error from concise to verbose to flawed responses) absent in DeepSeek-3, and an optimistic bias toward incorrect answers in both models. Accordingly, the platform ranking we observed, obtained with a single zero-shot prompt, the QAMAI and GQS instruments, and one point in time, should not be generalized to other prompts, tasks, or evaluation frameworks; structured prompting might have narrowed or changed the differences between platforms, and model-specific prompt sensitivity [28] is one reason multi-platform comparisons remain necessary. The response-style finding [29] motivated our within-platform analysis of length-score associations. Whereas Hassanein et al. [28] accessed the model through an API at temperature 0, consumer chat interfaces do not expose decoding parameters; our study therefore characterizes the products as clinicians actually use them, which strengthens ecological validity but limits reproducibility, a constraint both studies acknowledge with respect to model updating. Previous TMD-focused studies have reported potentially useful chatbot-generated information while emphasizing professional oversight [15,16], but addressed general TMD knowledge or patient education rather than clinician-oriented decision-making in open TMJ surgery, where decisions are highly context dependent and potentially irreversible [2–6,19]. The present study extends this evidence to indications and patient selection, imaging and preoperative planning, surgical approaches, anatomical risks, complications, and reconstruction [2–6,19].

Table 5
Inter-rater reliability between the two evaluators (intraclass correlation coefficients, n = 200 pooled responses).

Outcome measure	ICC(3,1) single measure (95% CI)	ICC(3,2) average measure (95% CI)
QAMAI total	0.970 (0.96–0.98)	0.985 (0.98–0.99)
QAMAI accuracy	0.937 (0.92–0.95)	0.967 (0.96–0.98)
QAMAI clarity	0.864 (0.82–0.90)	0.927 (0.90–0.94)
QAMAI relevance	0.888 (0.85–0.91)	0.941 (0.92–0.95)
QAMAI completeness	0.863 (0.82–0.89)	0.927 (0.90–0.94)
QAMAI source provision	0.961 (0.95–0.97)	0.980 (0.97–0.98)
QAMAI usefulness	0.920 (0.90–0.94)	0.958 (0.94–0.97)
GQS	0.918 (0.89–0.94)	0.957 (0.94–0.97)
Overall quality	0.932 (0.91–0.95)	0.965 (0.95–0.97)

ICC, intraclass correlation coefficient (two-way mixed-effects model, consistency definition); CI, confidence interval. ICC(3,1) quantifies agreement between the two individual raters; ICC(3,2) is the reliability of the averaged score used in the analyses. Interpretation: <0.50 poor, 0.50–0.75 moderate, 0.75–0.90 good, >0.90 excellent [31]. Platform-specific values are given in Supplementary Table S1; for the source-provision item of Perplexity both evaluators assigned identical scores to all 40 responses (ICC = 1.00, CI not estimable).

The between-platform differences in source provision, and consequently in the QAMAI total, should be interpreted in the light of the platforms' retrieval configurations. Perplexity is a retrieval-augmented search engine that cites web sources by default, whereas under the default settings of the other platforms source citation is inconsistent or absent. Its superior source-provision score therefore reflects a product-level design feature rather than greater medical knowledge, and platform-level differences in browsing and citation behavior may have contributed to the ranking. When source provision was excluded, Perplexity's lead disappeared and Claude, Perplexity, and Gemini converged, whereas the lower scores of ChatGPT and Grok persisted, indicating that the accuracy and usefulness differences are not attributable to citation behavior alone. Our findings should accordingly be understood as a comparison of chatbot platforms as used by clinicians, not of the intrinsic performance of the underlying language models.

Response length is a potential source of bias. The moderate pooled correlation between word count and total score was largely explained by platform identity: within the three platforms producing longer answers the correlation was weak, the longest-answering platform did not obtain the highest scores, and completeness was only weakly related to length. Very short responses, as typically produced by Grok, omitted guideline key elements and were penalized on accuracy and usefulness rather than on length per se. The evaluators therefore did not appear to reward verbosity, although the possibility that longer answers were more likely to contain the reference key elements cannot be fully separated from accuracy in this design.

Claude and Gemini produced the most linguistically complex text, whereas ChatGPT, Grok, and Perplexity were comparably more

readable; even the most readable responses remained far more complex than public-facing benchmarks. Although specialized terminology was expected for clinician-oriented questions, excessive complexity may hinder information appraisal, teaching, interdisciplinary communication, and adaptation for patient counseling. Readability should, however, be interpreted independently from clinical accuracy, as simpler language does not indicate clinically correct or complete information [30,30].

These findings support a constrained and supervised role for LLMs in open TMJ surgery: they may assist with preliminary information synthesis, topic organization, and identification of issues requiring verification [7,8], but should not be used independently to determine indications, select approaches, interpret imaging, or guide the management of complications. Hallucinated information, unverifiable citations, omission of clinically important qualifications, automation bias, and variability across repeated queries remain important concerns [11,12,17,31,32], and LLM-generated information should be verified against current guidelines and primary sources before informing clinical decisions.

It should be emphasized that the QAMAI, GQS, and overall quality instruments measure the quality, completeness, and guideline concordance of written responses as judged by specialists; they do not measure clinical safety, diagnostic accuracy in real patients, or the effect of chatbot use on surgical decision-making or patient outcomes. A response rated accurate and complete may still be unsafe if applied to an individual patient without a surgeon's contextual judgement, and readability describes linguistic accessibility rather than correctness. No patient-level outcomes were evaluated; our conclusions are restricted to the informational quality of the responses, and the requirement for specialist verification follows directly from this distinction.

Strengths include a 40-question set developed independently by two oral and maxillofacial surgeons from contemporary AAOMS guidance and finalized by consensus across eight domains [19]; identical wording, new chat sessions, verbatim prompting, and no corrective prompts, which standardized response collection across platforms; and blinded, independent specialist assessment guided by a priori key elements and written scoring anchors, with multidimensional quality measures and readability indices, consistent with the reporting priorities emphasized by CHART [33,33].

Several limitations should be acknowledged. First, each question was submitted once to each platform on a single day (16 July 2026). Because commercial chatbots generate responses stochastically and are updated without notice, the results represent a snapshot at the stated access date; within-platform response variability was not assessed, and repeated querying under identical conditions is needed to establish the consistency of the differences reported here [32]. Decoding parameters (temperature, top-p) and platform-side content filters are not user-configurable or visible in consumer chat interfaces and could not be equalized across platforms, so part of the between-platform variability may reflect default sampling or filtering rather than underlying model capability, and a sensitivity analysis with varied decoding parameters was not possible. Platform labels may also not fully reflect underlying model revisions or system-level instructions. Second, the question set was not subjected to external psychometric validation, and the number of questions varied across domains. Third, text-only prompts did not assess image interpretation, multimodal integration, operative performance, or patient outcomes, and readability metrics quantify linguistic complexity rather than comprehension, accuracy, or safety. Finally, the same two surgeons developed the questions and reference answers and rated the responses; although responses were coded, randomized, and assessed blind to the generating platform, and reference answers were derived directly from the guidance document [19], this dual role may introduce evaluator-related bias, and independent adjudication by an external panel would strengthen future studies.

4.1. Future directions

Future studies should incorporate repeated queries under identical conditions and, where API access permits, controlled decoding parameters, to characterize within-platform variability and temporal stability. Larger, externally validated question sets, multicenter expert panels, independently adjudicated reference answers, systematic source verification, clinically consequential error analyses, and multimodal scenarios incorporating imaging would strengthen rigor and generalizability. Prospective studies are needed to determine whether supervised LLM use improves clinician learning, efficiency, and patient communication without increasing clinical error.

Ethics statement

This study did not involve human participants, human data, animals, or identifiable personal information; it evaluated publicly generated artificial intelligence text outputs. Therefore, ethics committee approval and informed consent were not required.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Claude Fable 5 (Anthropic) in order to improve the language fluency and grammar of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Selin Gaş: Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Erdinç Sulukan:** Writing – review & editing, Validation, Methodology, Investigation. **Büşra Korkmaz:** Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jormas.2026.102972.

References

[1] Valesan LF, Da-Cas CD, Réus JC, Denardin ACS, Garanhani RR, Bonotto D, et al. Prevalence of temporomandibular joint disorders: a systematic review and meta-analysis. Clin Oral Investig 2021;25:441–53. doi: 10.1007/s00784-020-03710-w.

[2] Bouloux GF, Chou J, DiFabio V, Ness G, Perez D, Mercuri L, et al. The contemporary management of temporomandibular joint intra-articular pain and dysfunction. J Oral Maxillofac Surg 2024;82:623–31. doi: 10.1016/j.joms.2024.01.003.

[3] Sidebottom AJ. Current thinking in open temporomandibular joint surgery: is this still indicated in the management of articular temporomandibular joint disorder? Br J Oral Maxillofac Surg 2024;62:324–8. doi: 10.1016/j.bjoms.2024.01.006.
[4] Wroclawski C, Mediratta JK, Fillmore WJ. Recent advances in temporomandibular joint surgery. Medicina (Kaunas) 2023;59:1409. doi: 10.3390/medicina59081409.
[5] Peres Lima F, Rios LGC, Bianchi J, Gonçalves JR, Paranhos LR, Vieira WA, et al. Complications of total temporomandibular joint replacement: a systematic review and meta-analysis. Int J Oral Maxillofac Surg 2023;52:584–94. doi: 10.1016/j.ijom.2022.10.009.
[6] Moin A, Shetty AD, Archana TS, Kale SG. Facial nerve injury in temporomandibular joint approaches. Ann Maxillofac Surg 2018;8:51. doi: 10.4103/ams.ams_200_17.
[7] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29:1930–40. doi: 10.1038/s41591-023-02448-8.
[8] Liang EN, Pei S, Staibano P, van der Woerd B. Clinical applications of large language models in medicine and surgery: a scoping review. J Int Med Res 2025;53:03000605251347556. doi: 10.1177/03000605251347556.
[9] Ronsivalle V, Santonocito S, Cammarata U, Lo Muzio E, Cicciù M. Current applications of chatbots powered by large language models in oral and maxillofacial surgery: a systematic review. Dent J (Basel) 2025;13:261. doi: 10.3390/dj13060261.
[10] de Menezes Torres LM, de Morais EF, Fernandes Almeida DRM, Pagotto LEC, de Santana Santos T. The impact of the large language model ChatGPT in oral and maxillofacial surgery: a systematic review. Br J Oral Maxillofac Surg 2025;63:357–62. doi: 10.1016/j.bjoms.2025.03.006.
[11] Azamfirei R, Kudchadkar SR, Fackler JC. Large language models and the perils of their hallucinations. Crit Care 2023;27:120. doi: 10.1186/s13054-023-04393-x.
[12] Jung KH. Large language models in medicine: clinical applications, technical challenges, and ethical considerations. Healthc Inform Res 2025;31:114–24. doi: 10.4258/hir.2025.31.2.114.
[13] Kurt Demirsoy K, Buyuk SK, Bicer T. How reliable is the artificial intelligence product large language model ChatGPT in orthodontics? Angle Orthod 2024;94:602–7. doi: 10.2319/031224-207.1.
[14] Ji K, Wu ZD, Han J, Zhai G, Liu J. Evaluating ChatGPT-4′s performance on oral and maxillofacial queries: chain of thought and standard method. Front Oral Health 2025;6:1541976. doi: 10.3389/froh.2025.1541976.
[15] Kula B, Kula A, Bagcier F, Alyanak B. Artificial intelligence solutions for temporomandibular joint disorders: contributions and future potential of ChatGPT. Korean J Orthod 2025;55:131–41. doi: 10.4041/kjod24.106.
[16] Özdal Zincir Ö, Hatipoğlu Ş, Çifçi Özkan E. Large language model (LLM) in temporomandibular disorder education: a comparative study. BMC Oral Health 2026;26:112. doi: 10.1186/s12903-025-07468-z.
[17] Artsi Y, Sorin V, Glicksberg BS, Korfiatis P, Freeman R, Nadkarni GN, et al. Challenges of implementing LLMs in clinical practice: perspectives. J Clin Med 2025;14:6169. doi: 10.3390/jcm14176169.
[18] Haylaz E, Gumussoy I, Kalabalik F, Say Ş, Eren MC, Geduk G. Evaluation of the performance of four different large language models (ChatGPT, DeepSeek, Copilot, and Gemini) in answering oral and maxillofacial radiology questions: pilot study. BMC Oral Health 2026;26:426. doi: 10.1186/s12903-026-07790-0.
[19] Bouloux GF, Chou J, DiFabio V, Ness G, Perez D, Mercuri L, et al. Management of patients with orofacial pain and temporomandibular disorders. J Oral Maxillofac Surg 2025;83:1078.e1–1078.e83. doi: 10.1016/j.joms.2024.03.018.
[20] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15:155–63. doi: 10.1016/j.jcm.2016.02.012.
[21] Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, et al. Validation of the quality analysis of medical artificial intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. Eur Arch Otorhinolaryngol 2024;281:6123–31. doi: 10.1007/s00405-024-08710-0.
[22] Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol 2007;102:2070–7. doi: 10.1111/j.1572-0241.2007.01325.x.
[23] Uçar SM, Gaş S, Sasany R. Comparative performance of AI chatbots in dental implantology: insights and limitations. BMC Oral Health 2026;26:147. doi: 10.1186/s12903-025-07426-9.
[24] Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA 2025;333:319. doi: 10.1001/jama.2024.21700.
[25] Yilmaz BE, Ozbey F, Gokkurt Yilmaz BN, Akpinar H. Can large language models perform clinical anamnesis? Comparative evaluation of ChatGPT, Claude, and Gemini in diagnostic reasoning through case-based questioning in oral and maxillofacial disorders. J Stomatol Oral Maxillofac Surg 2026;127:102644. doi: 10.1016/j.jormas.2025.102644.
[26] Wu X, Cai G, Guo B, Ma L, Shao S, Yu J, et al. A multi-dimensional performance evaluation of large language models in dental implantology: comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across diverse clinical scenarios. BMC Oral Health 2025;25:1272. doi: 10.1186/s12903-025-06619-6.
[27] Nguyen VA, Ha TBN, Tran MN, Pham NTM, Nguyen TL, Vuong TQT. Quantifying the speed-accuracy trade-off of large language models on oral and maxillofacial surgery multiple-choice questions. Sci Rep 2025;15:40657. doi: 10.1038/s41598-025-27256-7.
[28] Hassanein FEA, Ibrahim SS, Tomo S, Alsahhaf A, Abou-Bakr A. Prompt engineering shapes diagnostic accuracy and explanation quality of LLM in oral lesion diagnosis: a prospective, expert-blinded benchmark study. Odontology 2026. doi: 10.1007/s10266-026-01380-w.
[29] Hassanein FEA, Hussein RR, Ahmed Y, El-Guindy J, Ahmed DE, Abou-Bakr A. Calibration of AI large language models with human subject matter experts for grading of clinical short-answer responses in dental education. BMC Oral Health 2026;26:286. doi: 10.1186/s12903-026-07665-4.
[30] Özdemir ÖT, Kavan MY, Güven Y. Evaluation of the readability, quality, and accuracy of AI chatbot responses to questions about deleterious oral habits. BMC Oral Health 2025;25:1812. doi: 10.1186/s12903-025-07298-z.
[31] Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 2023;13:14045. doi: 10.1038/s41598-023-41032-5.
[32] Zhu L, Mou W, Hong C, Yang T, Lai Y, Qi C, et al. The evaluation of generative AI should include repetition to assess stability. JMIR Mhealth Uhealth 2024;12:e57978. doi: 10.2196/57978.
[33] CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). BMJ 2025;390:e083305. doi: 10.1136/bmj-2024-083305.