Benchmarking Large Language Models for Dental Clinical Decision Support: A BERT Score Analysis of Claude Opus 4.5

Achmad Zam Zam Aghasy[1], Muhammad Lutfan Lazuardi[2], Hari Kusnanto Josef[3]
Faculty of Medicine-Department of Health Policy and Management-Public Health and Nursing,
Universitas Gadjah Mada, Yogyakarta, Indonesia[1]
Faculty of Dentistry-Department of Preventive and Community Dentistry,
Universitas Gadjah Mada, Indonesia, Yogyakarta, Indonesia[1, 2]
Faculty of Medicine-Department of Family-Community Medicine and Bioethics-Public Health and Nursing,
Universitas Gadjah Mada, Yogyakarta, Indonesia[3]

***Abstract*—The integration of Large Language Models (LLMs) into clinical decision support systems represents a significant advancement in healthcare informatics. This study presents a comprehensive evaluation framework for benchmarking LLM-generated dental treatment recommendations using BERT Score as the primary semantic similarity metric. We evaluated Claude Opus 4.5 as a Clinical Decision Support System (CDSS) across 116 dental case reports extracted from the Case Reports in Dentistry journal (2024-2025), spanning nine dental specialties. The BERT Score was calculated using the RoBERTa-large model to measure semantic alignment between AI-generated treatment plans and gold-standard published treatments. Results demonstrated strong semantic alignment with a mean BERT Score F1 of 0.8199 with a standard deviation of 0.0144 (95 per cent confidence interval: 0.8172-0.8225), significantly exceeding the 0.80 threshold (t = 14.90, p < 0.001, d = 1.38). Cross-specialty analysis revealed consistent performance across all nine dental domains (Kruskal-Wallis H = 3.07, p = 0.879), indicating robust generalizability. A significant negative correlation was observed between BERT Score and response time (ρ = -0.371, p < 0.001), suggesting a speed-accuracy trade-off in LLM reasoning. This study contributes a reproducible benchmarking methodology for evaluating LLM performance in specialized clinical domains and demonstrates the potential of BERT Score as a scalable evaluation metric for AI-generated clinical text.**

***Keywords*—BERT Score; Large Language Models; clinical decision support system; semantic similarity; Claude Opus 4.5**

I. Introduction

The rapid advancement of Large Language Models (LLMs) has catalyzed transformative applications in healthcare, particularly in clinical decision support systems (CDSS) [1]. These models demonstrate remarkable capabilities in processing and generating medical text, offering potential solutions for diagnostic assistance, treatment planning, and clinical documentation [2]. However, the deployment of LLMs in high-stakes clinical environments necessitates rigorous evaluation methodologies that can accurately assess the semantic fidelity and clinical relevance of AI-generated recommendations [3].

Traditional evaluation metrics for natural language generation, such as BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation), rely primarily on n-gram overlap between generated and reference texts [4]. While these metrics provide valuable insights for machine translation and summarization tasks, they exhibit significant limitations in clinical contexts where semantic equivalence often transcends lexical similarity. Clinical terminology frequently admits multiple valid expressions for identical concepts; for instance, "necrotic pulp" and "non-vital tooth" are clinically synonymous yet lexically distinct, resulting in artificially low scores under traditional metrics [5].

BERT Score addresses these limitations by leveraging contextual embeddings from transformer-based models to compute semantic similarity at the token level [6]. By representing words in a high-dimensional vector space that captures contextual meaning, BERT Score can recognize semantic equivalence between different surface forms of the same clinical concept. This capability is particularly valuable in medical domains where paraphrasing is common and clinical correctness should not be penalized by lexical variation [7].

Dentistry presents unique challenges for AI evaluation due to its specialized terminology, procedural complexity, and the critical importance of spatial reasoning [8]. Dental clinical decisions require integration of patient history, clinical examination findings, and radiographic interpretation, a triangulation of evidence that tests the reasoning capabilities of LLMs. Recent benchmarking studies have demonstrated variable performance of LLMs across dental specialties, with models achieving high accuracy on standardized examinations but exhibiting inconsistencies in complex case-based reasoning [9][10].

Claude Opus 4.5, released by Anthropic, represents a reasoning-optimized LLM architecture featuring extended context windows (200,000 tokens) and inference-time compute scaling through its "Thinking" mechanism [11]. These architectural innovations suggest potential advantages for complex clinical reasoning tasks that require maintaining coherent logical chains across extensive patient histories. However, systematic evaluation of Claude Opus 4.5 in dental clinical decision support remains limited in the current literature.

While prior studies have evaluated LLMs on dental examinations using multiple-choice accuracy metrics [9] [10], these approaches cannot capture the nuanced semantic alignment required for open-ended clinical reasoning tasks. Furthermore, existing benchmarks predominantly assess single-specialty performance, leaving cross-specialty generalizability largely unexplored. This study addresses these methodological gaps by introducing a BERT Score-based evaluation framework specifically designed for open-ended dental treatment recommendations—a context where multiple semantically equivalent but lexically distinct responses may be clinically valid.

The primary contributions of this study are threefold. First, we establish a reproducible, training-free semantic evaluation pipeline that enables scalable assessment of LLM-generated clinical text without requiring domain-specific fine-tuning or retrieval augmentation. Second, we provide empirical evidence of cross-specialty generalizability by demonstrating consistent BERT Score performance across nine dental disciplines, addressing the critical question of whether LLM clinical reasoning transfers across subspecialty boundaries. Third, we characterize the speed-accuracy relationship in dental CDSS applications, offering practical insights for real-time clinical deployment where response latency affects usability. These contributions extend beyond dental informatics to inform evaluation methodology for LLM-based clinical decision support systems across medical domains.

The remainder of this study is organized as follows: Section II reviews related work on LLM applications in dental clinical decision support, evaluation metrics for clinical text generation, and existing benchmarking frameworks for healthcare AI. Section III details the methodology, including data collection from the *Case Reports in Dentistry* journal, dental specialty classification, AI configuration with structured prompting, and BERT Score calculation procedures. Section IV presents the results encompassing overall BERT Score performance, cross-specialty analysis, and speed-accuracy correlations. Section V discusses the interpretation of findings, clinical implications, and study limitations. Finally, Section VI concludes with key contributions and directions for future research.

II. Related Work

A. Large Language Models in Dental Clinical Decision Support

The application of Large Language Models in dentistry has expanded rapidly, with multiple model families demonstrating varying capabilities across clinical tasks. Kim et al. evaluated LLM performance on the Korean Dental Licensing Examination, finding that GPT-4 achieved 75.2% accuracy compared to GPT-3.5's 53.8%, establishing a clear generational performance gap [9]. Dashti et al. extended this analysis to U.S. dental examinations, reporting similar patterns with GPT-4 reaching 80.1% accuracy versus 67.3% for GPT-3.5 [12]. Wu et al. conducted multi-dimensional evaluation of LLMs in dental implantology, comparing ChatGPT, DeepSeek, Grok, Gemini, and Qwen across clinical consensus and case analysis tasks [13]. These findings indicate substantial improvement in dental knowledge representation across model generations, though both studies relied exclusively on multiple-choice question formats.

Beyond accuracy metrics, recent studies have begun characterizing operational parameters relevant to clinical deployment. Nguyen et al. quantified the speed-accuracy trade-off in oral and maxillofacial surgery multiple-choice tasks, demonstrating that reasoning-optimized models achieve 12-18% higher accuracy at the cost of 3-5× increased response latency [14]. This trade-off has direct implications for real-time clinical decision support, where response time affects usability and workflow integration. However, whether this relationship holds for open-ended clinical reasoning tasks as opposed to multiple-choice selection remains unexplored.

Comparative benchmarking across model families has revealed distinct performance profiles and highlighted the importance of model selection for specific clinical applications. Wu et al. conducted a multi-dimensional evaluation in dental implantology comparing five major LLMs, finding that GPT-4 achieved the highest accuracy (78.5%) followed by Gemini (72.1%), Claude 3 (69.8%), Qwen (65.4%), and DeepSeek (61.2%) across clinical consensus and case analysis tasks [10]. Notably, performance rankings varied by task type—Claude demonstrated stronger performance on complex case analysis requiring multi-step reasoning, while GPT-4 excelled on factual recall tasks. Fujimoto et al. evaluated LLMs specifically in dental anesthesiology, reporting that ChatGPT-4 achieved 51.2% accuracy compared to Claude 3 Opus at 47.4% and Gemini 1.0 at 43.6% on Japanese board certification questions [15]. Hou et al. benchmarked multiple LLMs on the Dental Admission Test, with GPT-4 achieving 76.8% overall accuracy while demonstrating particular strength in reading comprehension (84.2%) compared to perceptual ability sections (68.1%) [16].

Despite these advances, critical methodological limitations persist in current dental LLM evaluation. First, the predominant reliance on multiple-choice question formats constrains assessment to recognition-based rather than generation-based clinical reasoning [17]. Real clinical decision-making requires synthesizing patient information into coherent treatment plans—a generative task fundamentally different from selecting among predetermined options. Second, existing studies predominantly evaluate single-specialty performance, leaving cross-specialty generalizability largely unexplored. Whether LLM clinical reasoning transfers across dental subspecialty boundaries remains an open question with significant implications for deployment scope. Third, Claude Opus 4.5, Anthropic's latest reasoning-optimized model featuring extended context windows (200,000 tokens) and inference-time compute scaling, remains systematically unevaluated in dental clinical decision support contexts despite architectural innovations potentially advantageous for complex clinical reasoning.

B. Evaluation Metrics for Clinical Text Generation

Traditional NLP evaluation metrics exhibit well-documented limitations in clinical contexts that motivate the adoption of embedding-based alternatives. BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) rely on n-gram overlap, penalizing semantically equivalent paraphrases common in

clinical documentation [17]. For instance, "extraction of non-restorable tooth" and "removal of hopeless dentition" convey identical clinical meaning but share minimal lexical overlap, yielding artificially low BLEU/ROUGE scores. Hanna and Bojar [6] demonstrated through fine-grained analysis that these metrics correlate poorly with human judgment (Pearson $r < 0.4$) for semantic equivalence tasks, particularly when surface-form variation is high—a characteristic inherent to clinical terminology where multiple valid expressions exist for identical concepts.

BERT Score addresses these limitations by leveraging contextual embeddings from transformer-based models to compute semantic similarity at the token level [4]. Zhang et al. introduced the framework demonstrating superior correlation with human judgment (Pearson $r = 0.73$) compared to BLEU ($r = 0.41$) and ROUGE ($r = 0.38$) across multiple natural language generation tasks [4]. The metric computes precision, recall, and F1 scores through greedy matching of contextualized token embeddings, capturing meaning beyond surface-level lexical patterns. Precision measures whether generated content is semantically supported by the reference, recall assesses coverage of reference content, and F1 provides a balanced harmonic mean. This capability is particularly valuable in clinical domains where terminology variation is common and semantic correctness should not be penalized by lexical diversity.

Domain adaptation techniques have further enhanced semantic evaluation accuracy for medical text, though their application to open-ended clinical text generation remains limited. BioBERT [18], pretrained on over 18 billion words from PubMed abstracts and PMC full-text articles, demonstrates improved performance on biomedical named entity recognition, relation extraction, and question answering tasks compared to general-domain BERT. ClinicalBERT [4], trained on approximately 2 million clinical notes from the MIMIC-III database, achieves superior results on clinical concept extraction and hospital readmission prediction by capturing the distinctive linguistic patterns of clinical documentation. PubMedBERT [4] employs domain-specific vocabulary and pretraining exclusively on biomedical literature, outperforming mixed-domain models on the Biomedical Language Understanding Evaluation (BLUE) benchmark. More recently, Koroleva et al. [19] demonstrated that domain-adapted embeddings improve semantic similarity measurement for clinical trial outcomes, achieving higher correlation with expert annotations than general-domain alternatives.

However, these domain-adapted models have been primarily validated on classification, extraction, and structured prediction tasks rather than open-ended text generation evaluation. Whether the advantages of biomedical pretraining transfer to semantic similarity assessment of generated clinical recommendations remains an open empirical question. Furthermore, no systematic comparison exists between the general-domain BERT Score (using RoBERTa-large) and domain-adapted variants (BioBERTScore, ClinicalBERTScore) for dental clinical text specifically. This study employs RoBERTa-large as the backbone model based on its status as the recommended default in the BERT-score library and its robust cross-domain performance [20]. This choice provides a conservative baseline—success with general-domain embeddings suggests even stronger potential with domain adaptation, while failure would not be attributable to domain mismatch.

C. Benchmarking Frameworks for Healthcare AI

The development of standardized benchmarking datasets has accelerated healthcare AI evaluation while revealing persistent methodological gaps. Zhu et al. [20] introduced DentalBench, a bilingual benchmark comprising over 36,000 questions across 16 dental subfields, including operative dentistry, prosthodontics, orthodontics, and oral surgery. This comprehensive dataset enables systematic cross-specialty comparison but relies exclusively on multiple-choice formats. Hou et al. [16] benchmarked LLMs on the Dental Admission Test, providing comparative performance data across model families with a detailed analysis of performance variation across test sections. Huang et al. [5] presented a comprehensive survey on evaluating LLM applications in medical settings, identifying accuracy, safety, clinical relevance, and explainability as key evaluation dimensions that should be assessed in combination rather than in isolation.

Recent methodological advances emphasize multidimensional evaluation frameworks that extend beyond single accuracy metrics. Sivaramakrishnan et al. [21] evaluated LLMs for dental patient education materials using BERT Score alongside readability indices (Flesch-Kincaid, SMOG) and clinical relevance assessments by domain experts, demonstrating that semantic similarity alone incompletely captures communication quality. Zheng et al. [22] proposed hierarchical divide-and-conquer approaches for fine-grained alignment in LLM-based medical evaluation, decomposing complex clinical reasoning into assessable sub-components. These frameworks recognize that no single metric fully captures clinical utility, advocating for complementary quantitative and qualitative assessments tailored to specific use cases.

Despite these advances, existing benchmarking frameworks exhibit limitations that constrain their applicability to real-world clinical decision support evaluation. The predominant MCQ format assesses knowledge recall rather than clinical reasoning synthesis—a clinician must not only identify correct answers but generate coherent, contextually appropriate treatment plans from unstructured patient information [17]. Current benchmarks lack standardized semantic similarity metrics for open-ended response evaluation, relying instead on binary correctness judgments that cannot capture partial credit, near-miss responses, or alternative valid approaches. It is uncommon to systematically evaluate cross-specialty generalizability; models may perform well in knowledge-intensive domains but poorly in procedural or spatial reasoning, but single-specialty benchmarks are unable to identify this variation [8]. Finally, speed-accuracy trade-offs critical for real-time clinical deployment are inconsistently reported across studies, limiting practical implementation guidance.

C. Research Gap and Study Positioning

The preceding review identifies three convergent gaps that motivate the present study. First, an evaluation methodology gap: existing dental LLM benchmarks predominantly employ MCQ-based accuracy metrics that cannot assess open-ended

clinical reasoning quality, where multiple semantically equivalent responses may be valid. Semantic similarity metrics like BERT Score offer a principled alternative for evaluating generated clinical text but remain underutilized in dental AI evaluation. Second, a cross-specialty evidence gap: whether LLM clinical reasoning generalizes across dental subspecialties is largely unexplored, yet this question has direct implications for deployment scope, safety boundaries, and clinical workflow integration. Third, a model coverage gap: Claude Opus 4.5, featuring architectural innovations including extended context windows and inference-time reasoning optimization, potentially advantageous for complex clinical decision-making, lacks systematic evaluation in dental clinical decision support.

This study addresses these gaps by establishing a BERT Score-based benchmarking framework for evaluating open-ended dental treatment recommendations. Unlike MCQ-based approaches that assess recognition accuracy, our framework evaluates the semantic alignment between AI-generated and expert-authored treatment plans—capturing clinical reasoning quality through continuous similarity scores rather than binary correctness judgments. By systematically evaluating performance across nine dental specialties using 116 published case reports, we provide empirical evidence regarding cross-specialty generalizability that informs appropriate deployment boundaries. The methodology offers a reproducible, training-free evaluation pipeline applicable to any LLM without requiring domain-specific fine-tuning or retrieval augmentation, facilitating standardized comparison across model families and enabling longitudinal tracking of model performance across version updates.

III. Methodology

A. Study Design and Data Collection

This cross-sectional study evaluated Claude Opus 4.5 as a clinical decision support system using published dental case reports as the evaluation corpus. The overall study workflow is illustrated in Fig. 1, comprising four sequential phases: 1) data collection and preprocessing, 2) AI-generated treatment recommendation, 3) BERT Score calculation, and 4) statistical analysis. This design enables systematic comparison between AI-generated and expert-authored treatment plans while controlling for case complexity through the use of published cases with documented outcomes for the period 2024 to 2025.

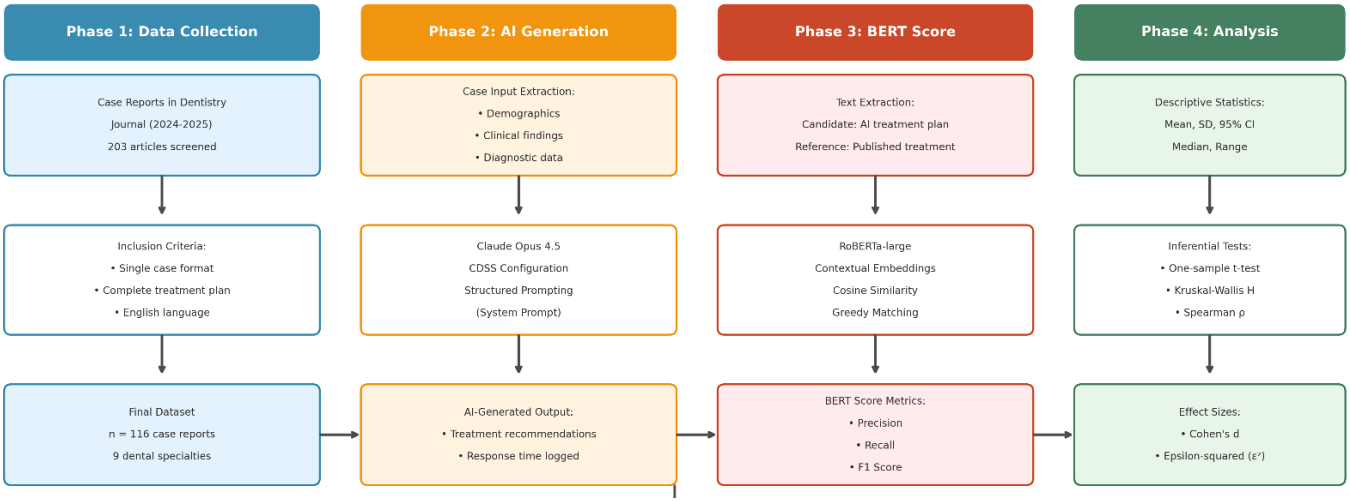

Fig. 1. Study workflow of benchmarking LLM for CDSS using BERT score.

Inclusion criteria comprised: 1) publication type classified as "case report" in journal metadata; 2) complete case presentation including patient demographics, chief complaint, clinical findings, diagnostic workup, and treatment plan sections; 3) single-case format with clearly identifiable treatment outcomes; 4) English language publication. Exclusion criteria included: a) case series presenting multiple patients, as these confound treatment-case matching; b) reports lacking explicit treatment plan descriptions; c) cases focused primarily on diagnostic imaging or laboratory findings without therapeutic intervention; d) letters to the editor, commentaries, or review articles misclassified as case reports. The final dataset comprised 116 case reports meeting all inclusion criteria.

B. Dental Specialty Classification

Cases were classified into dental specialties using a keyword-based algorithm with title weighting. The classification system assigned cases to nine specialty categories: Oral and Maxillofacial Surgery, Orthodontics, Endodontics, Prosthodontics, Implantology, Pediatric Dentistry, Periodontics, Oral Medicine and Pathology, and Conservative Dentistry. Keywords were derived from established dental specialty nomenclature and weighted by occurrence in article titles (2×) versus body text (1×). Cases matching multiple specialties were assigned to the category with the highest weighted keyword frequency.

C. AI Configuration and Prompt Engineering

Claude Opus 4.5 was configured as a Clinical Decision Support System through structured system prompting. The prompt established: 1) Role definition as CDSS for dental healthcare professionals; 2) Explicit limitations stating recommendations require clinical validation; 3) Reference standards including FDI World Dental Federation notation and ICD-10/ICD-11 classification systems; 4) Structured response format comprising clinical findings analysis, differential

diagnosis, treatment recommendations, referral considerations, and diagnostic codes. The model was accessed via the Claude Website without fine-tuning, retrieval augmentation, or embedding modifications to evaluate baseline capabilities.

D. BERT Score Calculation

BERT Score was computed using the bert-score Python library (version 0.3.13) with RoBERTa-large as the backbone model. For each case, the candidate text comprised the treatment recommendation section of Claude's response, while the reference text comprised the treatment plan section from the published case report. BERT Score computes precision, recall, and F1 scores by: 1) generating contextual embedding for all tokens in candidate and reference texts; 2) computing pairwise cosine similarity between token embedding; 3) calculating greedy matching to maximize similarity scores; 4) aggregating token-level scores into sentence-level metrics. The F1 score, representing the harmonic mean of precision and recall, served as the primary evaluation metric.

E. Statistical Analysis

Statistical analyses were performed using JAMOVI Desktop 2.6.44. Descriptive statistics included mean, standard deviation, median, and 95% confidence intervals for BERT Score metrics. Distribution normality was assessed using the Shapiro-Wilk test. One-sample t-test evaluated whether the mean BERT Score significantly exceeded the 0.80 threshold, with Cohen's d calculated for effect size. Cross-specialty comparison employed the Kruskal-Wallis H test due to unequal group sizes, with epsilon-squared (ε^2) as the effect size measure. Spearman's rank correlation assessed the relationship between BERT Score and response time. Statistical significance was set at $\alpha = 0.05$.

IV. Results

A. Dataset Characteristics

The dataset comprised 116 dental case reports distributed across nine specialties. Oral and Maxillofacial Surgery represented the largest category (n = 56, 48.3%), followed by Orthodontics and Endodontics (n = 14 each, 12.1%), Prosthodontics (n = 13, 11.2%), Implantology (n = 7, 6.0%), Pediatric Dentistry (n = 5, 4.3%), Periodontics (n = 3, 2.6%), Oral Medicine and Pathology (n = 3, 2.6%), and Conservative Dentistry (n = 1, 0.9%). Table I presents the complete distribution.

TABLE I. Distribution of Case Reports by Dental Specialty

Dental Specialty	n	%
Oral and Maxillofacial Surgery	56	48.3
Orthodontics	14	12.1
Endodontics	14	12.1
Prosthodontics	13	11.2
Implantology	7	6.0
Pediatric Dentistry	5	4.3
Periodontics	3	2.6
Oral Medicine and Pathology	3	2.6
Conservative Dentistry	1	0.9
Total	**116**	**100.0**

B. Overall BERT Score Performance

Claude Opus 4.5 demonstrated strong semantic alignment with published treatment plans. The mean BERT Score F1 was 0.8199 ± 0.0144 (95% CI: 0.8172-0.8225), with a median of 0.8191 and a range of 0.7782-0.8618. The Shapiro-Wilk test confirmed normal distribution ($W = 0.994$, $p = 0.856$). Precision (mean = 0.8140 ± 0.0161) and recall (mean = 0.8252 ± 0.0198) values indicated balanced performance in capturing reference content while avoiding extraneous information. BERT Score F1 ≥ 0.80 was attained in 107 (92.2%) of the 116 cases.

One-sample t-test confirmed that the mean BERT Score significantly exceeded the 0.80 threshold ($t = 14.90$, $df = 115$, $p < 0.001$). Cohen's $d = 1.38$ indicated a large effect size, demonstrating substantial performance above the acceptability benchmark. Table II summarizes the overall performance metrics.

TABLE II. Overall BERT Score Performance Metrics

Metric	Value	95% CI
BERT Score F1 (Mean ± SD)	0.8199 ± 0.0144	[0.8172, 0.8225]
BERT Score Precision	0.8140 ± 0.0161	[0.8110, 0.8169]
BERT Score Recall	0.8252 ± 0.0198	[0.8216, 0.8289]
Median F1	0.8191	-
Range (Min-Max)	0.7782 - 0.8618	-
Cases ≥ 0.80 threshold	107 (92.2%)	-

C. Cross-Specialty Performance Analysis

BERT Score F1 demonstrated consistent performance across dental specialties. Mean scores ranged from 0.8088 (Pediatric Dentistry) to 0.8273 (Implantology), with all specialties exceeding the 0.80 threshold. Kruskal-Wallis H test revealed no statistically significant differences among specialties ($H = 3.07$, $df = 8$, $p = 0.879$), with epsilon-squared (ε^2) = 0.026 indicating a negligible effect size. This finding suggests robust generalizability of Claude Opus 4.5 across diverse dental clinical scenarios. Table III presents specialty-specific performance.

TABLE III. BERT Score F1 Performance by Dental Specialty

Specialty	n	Mean	SD	95% CI
Oral & Maxillofacial Surgery	56	0.8200	0.0155	[0.8159, 0.8241]
Orthodontics	14	0.8211	0.0122	[0.8147, 0.8275]
Endodontics	14	0.8173	0.0121	[0.8110, 0.8236]
Prosthodontics	13	0.8200	0.0088	[0.8152, 0.8248]
Implantology	7	0.8273	0.0181	[0.8139, 0.8407]
Pediatric Dentistry	5	0.8088	0.0186	[0.7926, 0.8251]
Periodontics	3	0.8198	0.0063	[0.8127, 0.8270]
Oral Medicine & Pathology	3	0.8217	0.0232	[0.7954, 0.8480]
Conservative Dentistry	1	0.8255	-	-

D. Response Time and Speed-Accuracy Trade-off

Mean response time was 34.79 ± 18.63 seconds (median = 32.0 seconds, range = 6-98 seconds). Spearman rank correlation revealed a significant negative relationship between BERT Score F1 and response time (ρ = -0.371, $p < 0.001$), indicating that faster responses were associated with higher semantic alignment scores. This counterintuitive finding suggests that cases eliciting clear, well-structured AI responses (higher BERT Scores) also required less processing time, potentially reflecting case complexity rather than a simple speed-accuracy trade-off. Table IV summarizes the overall statistical analysis.

TABLE IV. STATISTICAL ANALYSIS SUMMARY

Analysis	Statistic	p
Normality (Shapiro-Wilk)	$W = 0.994$	0.856
One-sample t-test (vs. 0.80)	$t = 14.90$, $df = 115$	< 0.001
Effect size (Cohen's d)	$d = 1.38$ (large)	-
Cross-specialty (Kruskal-Wallis)	$H = 3.07$, $df = 8$	0.879
Effect size (Epsilon-squared)	$\varepsilon^2 = 0.026$ (negligible)	-
Correlation (Spearman)	$\rho = -0.371$	< 0.001

V. DISCUSSION

A. Interpretation of BERT Score Performance

The observed mean BERT Score F1 of 0.8199 represents strong semantic alignment between AI-generated and expert-authored treatment recommendations. This performance significantly exceeds the 0.80 threshold typically considered acceptable for clinical text evaluation [19]. The large effect size (Cohen's $d = 1.38$) indicates that Claude Opus 4.5 consistently generates treatment recommendations that capture the semantic content of published clinical decisions, supporting its potential utility as a clinical decision support tool.

The balanced precision (0.8140) and recall (0.8252) scores indicate that the model neither omits critical treatment components, which would diminish recall, nor introduces clinically unsubstantiated or extraneous recommendations—often characterized as hallucinations—which would compromise precision. This metrical equilibrium is particularly imperative in clinical contexts where both informational completeness and semantic accuracy are essential for maintaining patient safety [23]. Furthermore, the narrow confidence interval of 0.8172–0.8225 reflects consistent performance across diverse case presentations.

B. Cross-Specialty Generalizability

The absence of significant performance differences across dental specialties ($p = 0.879$) represents a notable finding. Prior benchmarking studies have documented substantial variation in LLM performance across medical subspecialties, with models often excelling in knowledge-intensive domains while struggling with procedural or spatial reasoning tasks [24]. The consistent performance observed in this study suggests that Claude Opus 4.5's reasoning architecture may provide more robust generalization across diverse clinical scenarios than earlier model generations.

However, the unequal distribution of cases across specialties (48.3% in Oral and Maxillofacial Surgery) warrants caution in generalizing these findings. Smaller specialty subgroups may have insufficient statistical power to detect meaningful differences. Future studies should employ stratified sampling to ensure adequate representation across all dental specialties [25].

C. Speed-Accuracy Relationship

The significant negative correlation between BERT Score and response time (ρ = -0.371) presents an interesting departure from the expected speed-accuracy trade-off documented in other LLM benchmarking studies [14]. This relationship may reflect that straightforward cases with clear treatment pathways elicit both faster responses and higher semantic alignment with published recommendations. Complex or ambiguous cases requiring extended reasoning may produce lower BERT Scores due to legitimate treatment variability rather than model error.

D. Limitations of BERT Score in Clinical Evaluation

While BERT Score provides valuable insights into semantic similarity, several limitations must be acknowledged. First, semantic similarity does not guarantee clinical correctness; a response may be semantically similar to a reference while containing factually incorrect recommendations [26]. Second, BERT Score exhibits reduced sensitivity to numerical values; dosage errors (e.g., "5 mg" vs "500 mg") may not be adequately penalized due to similar contextual embeddings [27]. Third, the metric does not capture reasoning quality or logical consistency in treatment sequencing.

These limitations underscore the importance of multidimensional evaluation frameworks that combine automated semantic metrics with expert clinical review, safety assessments, and guideline concordance checks [28]. BERT Score should be viewed as a scalable screening tool rather than a definitive measure of clinical validity.

E. Implications for Clinical Decision Support Implementation

The results support the potential deployment of Claude Opus 4.5 as a clinical decision support tool in dental practice, with appropriate safeguards. The model demonstrates the capacity to generate semantically appropriate treatment recommendations across multiple specialties, potentially assisting clinicians with treatment planning, documentation, and educational applications [29]. However, the model should function as a "copilot" requiring human oversight rather than an autonomous diagnostic system [30].

Future implementation should incorporate 1) explicit uncertainty quantification in model outputs; 2) integration with domain-specific knowledge bases and clinical guidelines; 3) mandatory clinician review before treatment execution; and 4) continuous monitoring for performance degradation or emerging error patterns [31]. The ethics and governance framework proposed by Rokhshad et al. provides a valuable foundation for the responsible deployment of LLMs in dental clinical settings [32].

F. Study Limitations

Several limitations should be considered when interpreting these results. First, the use of published case reports as ground truth assumes that reported treatments represent optimal clinical

decisions, which may not always hold. Second, the study evaluated text-based clinical reasoning without incorporating radiographic or other imaging data essential to dental diagnosis. Third, the cross-sectional design cannot assess longitudinal performance stability or model drift. Fourth, the evaluation relied solely on automated BERT Score metrics without expert clinical review of individual outputs. Finally, the unequal specialty distribution may limit generalizability to underrepresented domains.

VI. Conclusion

This study established a BERT Score-based benchmarking framework for evaluating LLM performance in dental clinical decision support. Claude Opus 4.5 demonstrated strong semantic alignment with published treatment recommendations (mean BERT Score F1 = 0.8199 ± 0.0144), significantly exceeding the 0.80 acceptability threshold ($t = 14.90$, $p < 0.001$, Cohen's $d = 1.38$) with consistent performance across nine dental specialties (Kruskal-Wallis $H = 3.07$, $p = 0.879$). The study contributes a reproducible, training-free semantic evaluation pipeline for open-ended clinical text, provides the first cross-specialty generalizability evidence for dental LLM evaluation, and characterizes the speed-accuracy relationship relevant to clinical deployment.

Several limitations should be acknowledged, including the absence of expert clinical validation, reliance on text-based reasoning without radiographic data, and unequal specialty distribution. The findings support the potential of LLMs as clinical decision support tools in dentistry while highlighting the need for comprehensive evaluation frameworks that extend beyond semantic similarity to encompass clinical accuracy, safety, and guideline adherence. Future research should integrate expert clinical validation, multimodal evaluation incorporating radiographic interpretation, comparative benchmarking across LLM families, and prospective studies in clinical deployment settings. The standardization of evaluation methodologies will be essential as LLM applications in healthcare continue to expand

Acknowledgment

This research was funded by the Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan - LPDP), the Ministry of Finance, Republic of Indonesia. The authors gratefully acknowledge the financial support provided by LPDP that made this research possible. The first author would like to express profound gratitude to the Medical and Health Sciences Doctoral Program, Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada (UGM), where this research was conducted as part of a doctoral dissertation series. The authors also acknowledge the use of Claude Opus 4.5 (Anthropic) for generating clinical decision support responses evaluated in this study. The AI tool was used solely as the subject of evaluation, not in the preparation of this manuscript. All statistical analyses and interpretations were performed by the authors using JAMOVI (Version 2.6.44).

References

[1] H. Huang, O. Zheng, D. Wang, J. Yin, Z. Wang, S. Ding, H. Miao, and B. Shi, “ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model,” *International Journal of Oral Science*, vol. 15, no. 1, p. 29, 2023. doi: https://doi.org/10.1038/s41368-023-00239-y

[2] F. Umer, I. Batool, and N. Naved, “Innovation and application of Large Language Models (LLMs) in dentistry – a scoping review,” *BDJ Open*, vol. 10, no. 1, p. 100, 2024. doi https://doi.org/10.1038/s41405-024-00277-6

[3] S. Bedi, Y. Liu, L. Orr-Ewing, D. Dash, S. Koyejo, A. Callahan, J. Fries, M. Wornow, A. Swaminathan, L. Lehmann, and H. Kwon, “Testing and Evaluation of Healthcare Applications of Large Language Models: A Systematic Review,” *medRxiv*, 2024. doi: https://doi.org/10.1101/2024.04.15.24305869

[4] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” in *Proc. ICLR 2020*. doi: https://doi.org/10.48550/arXiv.1904.09675

[5] Y. Huang, K. Tang, and M. Chen, “A Comprehensive Survey on Evaluating Large Language Model Applications in the Medical Field,” *arXiv:2404.15777*, 2024. doi: https://doi.org/10.48550/arXiv.2404.15777

[6] M. Hanna and O. Bojar, “A Fine-Grained Analysis of BERTScore,” in *Proc. WMT 2021*, pp. 507-517, 2021.

[7] S. Agarwal, D. Wood, B. A. K. Murray, and T. C. Booth, “Impact of hospital-specific domain adaptation on BERT-based models to classify neuroradiology reports,” *European Radiology*, vol. 35, 2025. doi: https://doi.org/10.1007/s00330-025-11500-9

[8] A. Liu, H. Zhou, Y. Hua, O. Rohanian, L. Clifton, and D. A. Clifton, “Large Language Models in Healthcare: A Comprehensive Benchmark,” *arXiv:2405.00716*, 2024. doi: https://doi.org/10.48550/arXiv.2405.00716

[9] W. Kim, B. C. Kim, and H. G. Yeom, “Performance of Large Language Models on the Korean Dental Licensing Examination: A Comparative Study,” *International Dental Journal*, vol. 74, no. 6, pp. 1264-1270, 2024. doi: https://doi.org/10.1016/j.identj.2024.09.002

[10] X. Wu, G. Cai, B. Guo, L. Ma, S. Shao, J. Yu, Y. Zhang, Y. Zheng, and F. Yang, “A multi-dimensional performance evaluation of large language models in dental implantology,” *BMC Oral Health*, vol. 25, p. 232, 2025. doi: https://doi.org/10.1016/j.identj.2025.104193

[11] Anthropic, “Claude Opus 4.5,” 2025. [Online]. Available: https://www.anthropic.com/claude/opus

[12] M. Dashti, S. Ghasemi, N. Ghadimi, D. Ma, and M. Abdullah, “Performance of ChatGPT 3.5 and 4 on U.S. dental examinations,” *Imaging Science in Dentistry*, vol. 54, pp. 235-241, 2024. doi: https://doi.org/10.5624/isd.20240037

[13] Y. Wu, Y. Zhang, M. Xu, C. Jinzhi, and Y. Zheng, “Effectiveness of Various General large language models in Clinical Consensus and Case Analysis in Dental Implantology,” *BMC Medical Informatics and Decision Making*, 2025. doi: https://doi.org/10.1186/s12911-025-02972-2

[14] V. A. Nguyen, T. B. N. Ha, M. N. Tran, and T. Q. T. Vuong, “Quantifying the speed-accuracy trade-off of large language models on oral and maxillofacial surgery multiple-choice questions,” *Scientific Reports*, vol. 15, 2025. doi: https://doi.org/10.1038/s41598-025-27256-7

[15] M. Fujimoto, H. Kuroda, T. Katayama, and A. Yamaguchi, “Evaluating Large Language Models in Dental Anesthesiology: A Comparative Analysis of ChatGPT-4, Claude 3 Opus, and Gemini 1.0 on the Japanese Dental Society of Anesthesiology Board Certification Exam,” *Cureus*, vol. 16, no. 10, e70302, 2024. doi: https://doi.org/10.7759/cureus.70302

[16] Y. Hou, J. Patel, L. Dai, and R. Zhang, “Benchmarking of Large Language Models for the Dental Admission Test,” *Health Data Science*, vol. 5, p. 0250, 2025. doi: https://doi.org/10.34133/hds.0250

[17] A. S. Tong, K. Chung, A. P. S. Tong, et al., "The pitfalls of multiple-choice questions in generative AI and medical education," *Scientific Reports*, vol. 15, 2025. doi: https://doi.org/10.1038/s41598-024-54332-9

[18] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “BioBERT: a pre-trained biomedical language representation model,” *Bioinformatics*, vol. 36, no. 4, pp. 1234-1240, 2020. doi: https://doi.org/10.1093/bioinformatics/btz682

[19] A. Koroleva, S. Kamath, and P. Paroubek, "Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations," *Journal of Biomedical Informatics: X*, vol. 4, p. 100047, 2019. doi: https://doi.org/10.1016/j.yjbinx.2019.100058

[20] H. Zhu, Y. Xu, Y. Li, Z. Meng, and Z. Liu, "DentalBench: Benchmarking and Advancing LLMs Capability for Bilingual Dentistry Understanding," *arXiv:2508.20416*, 2025. doi: https://doi.org/10.48550/arXiv.2508.20416

[21] G. Sivaramakrishnan, M. Almuqahwi, S. Ansari, and K. Sridharan, "Assessing the power of AI: a comparative evaluation of large language models in generating patient education materials in dentistry," *BDJ Open*, vol. 11, p. 15, 2025. doi: https://doi.org/10.1038/s41405-025-00349-1

[22] S. Zheng, X. Zhang, G. de Melo, and L. Wang, "Hierarchical Divide-and-Conquer for Fine-Grained Alignment in LLM-Based Medical Evaluation," in *Proc. AAAI 2025*. doi: https://doi.org/10.48550/arXiv.2501.06741

[23] M. Raj, V. Ravindran, and A. Arthanari, "Assessing the Utility of Large Language Models in Guiding Dental Practitioners on Pediatric Patient Care," *Journal of Clinical and Experimental Dentistry*, vol. 17, 2025. doi: https://doi.org/10.4317/jced.63136

[24] J. Li, X. He, Y. Wang, Y. Liu, J. Liu, M. Liu, T. He, and Z. Huang, "Clinical decision support of advanced large language models in endodontic disease," *Journal of Dental Sciences*, 2025. doi: https://doi.org/10.1016/j.jds.2025.06.007

[25] K. Termteerapornpimol, S. Kulvitit, S. Prommanee, and T. Porntaveetus, "Comparative Benchmark of Seven Large Language Models for Traumatic Dental Injury Knowledge," *European Journal of Dentistry*, 2025. doi: https://doi.org/10.1055/s-0045-1812064

[26] J. Novikova, "Robustness and Sensitivity of BERT Models Predicting Alzheimer's Disease from Text," in *Proc. W-NUT 2021*, pp. 418-423. doi: https://doi.org/10.48550/arXiv.2109.11888

[27] S.H. Huang, Y.Y. Lee, T.H. Chou, and C.J. Wang, "FinNuE: Exposing the Risks of Using BERTScore for Numerical Semantic Evaluation in Finance," *arXiv:2511.09997*, 2025. doi: https://doi.org/10.48550/arXiv.2511.09997

[28] C. Lafourcade, O. Kérourédan, B. Ballester, and R. Richert, "Accuracy, consistency, and contextual understanding of large language models in restorative dentistry and endodontics," *Journal of Dentistry*, 2025. doi: https://doi.org/10.1016/j.jdent.2025.105764

[29] A. Dermata, A. Arhakis, M. A. Makrygiannakis, and E. G. Kaklamanos, "Evaluating the evidence-based potential of six large language models in paediatric dentistry," *European Archives of Paediatric Dentistry*, 2025. doi: https://doi.org/10.1007/s40368-025-01012-x

[30] M. J. Feldman, E. P. Hoffer, J. J. Conley, and H. C. Chueh, "Dedicated AI Expert System vs Generative AI with Large Language Model for Clinical Diagnoses," *JAMA Network Open*, vol. 8, 2025. doi: https://doi.org/10.1001/jamanetworkopen.2025.12994

[31] T. T. T. D. Huynh, U. K. Wiil, and A. Ebrahimi, "Quality Requirements for Large Language Models in Clinical Settings: A Scoping Review," in *Proc. BigDaCI 2025*.

[32] R. Rokhshad, A. Tichy, M. Ducret, A. Zerman, "The ethics and governance of large language models in dentistry: A framework for research and clinical implementation," *Journal of Dentistry*, 2026. doi: https://doi.org/10.1016/j.jdent.2025.106187