

Performance of Large Language Models as Assistive Tools for Clinical Pharmacists in the Stomatological Hospital: A Multi-dimensional Comparative Study

Jiatong Zhang

Xi'an Jiaotong University

Siyuan Hao

School of Stomatology, The Fourth Military Medical University

Bin Feng

School of Stomatology, The Fourth Military Medical University

Fang Jin

School of Stomatology, The Fourth Military Medical University

Shuibing Liu

School of Stomatology, The Fourth Military Medical University

Hui Li

Xi'an Jiaotong University

Jia Ju

jujia143@163.com

School of Stomatology, The Fourth Military Medical University

Research Article

Keywords: Large language models, Artificial intelligence, Clinical pharmacy, Stomatological Hospital, Drug information

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10798049/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Performance of Large Language Models as Assistive Tools for Clinical Pharmacists in the Stomatological Hospital: A Multi-dimensional Comparative Study

Jiatong Zhang², Siyuan Hao³, Bin Feng¹, Fang Jin³, Shuibing Liu¹, Hui Li^{2,*}, Jia Ju^{1,*}

Correspondence: jujia143@163.com (J Ju); huili16@xjtu.edu.cn (H Li)

(¹ National Clinical Research Center for Oral Diseases, State Key Laboratory of Oral & Maxillofacial Reconstruction and Regeneration, Shaanxi Engineering Research Center for Dental Materials and Advanced Manufacture, Department of Pharmacy, School of Stomatology, The Fourth Military Medical University, Xi'an 710032, China;

² School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an, 710049, China;

³ State Key Laboratory of Oral & Maxillofacial Reconstruction and Regeneration, National Clinical Research Center for Oral Diseases, Shaanxi Clinical Research Center for Oral Diseases, Department of Orthodontics, School of Stomatology, The Fourth Military Medical University, Xi'an, 710032, China.)

Abstract

Background Clinical pharmacists play an important role in managing pharmacotherapy safety. However, insufficient efficiency of manual drug consultation, heavy repetitive workloads, and delayed updating of professional knowledge commonly pose major challenges. Large language models (LLMs) have shown great capabilities in professional knowledge retrieval, text analysis, and intelligence-assisted decision making. Therefore, this study aimed to systematically evaluate the performance of LLMs (ChatGPT-5.5 Instant, Gemini-3.5 Flash-Lite and DeepSeek-V3) as a tool for clinical pharmacists in dental specialty hospital.

Methods Thirty questions were curated from daily pharmacotherapy practice in stomatological hospital and classified into three levels, easy, moderate and difficult. The response quality of three LLMs was blindly and independently assessed by five senior clinical pharmacists by global quality scale (GQS). The scores were reported as

median and interquartile range (IQR). Inter-model differences were assessed Kruskal–Wallis tests with Bonferroni-corrected post-hoc comparisons. The correlation between question difficulty level and response quality were evaluated by Fisher-Freeman-Halton exact test. The quality of information sources and readability of the responses were also evaluated.

Results ChatGPT reached the highest response quality median score (4.30, IQR: 4.00–4.60), followed by DeepSeek (3.70, IQR: 3.25–4.20) and Gemini (3.40, IQR: 3.00–3.40). The difference between each LLMs reached statistical significance by Kruskal–Wallis tests with Bonferroni-corrected post-hoc comparisons. Fisher-Freeman-Halton exact test indicated that response quality was not significantly correlated with question difficulty level. And the quality of information sources and readability were not significant among three LLMs.

Conclusions This study indicated that LLMs, especially ChatGPT, have the potential to serve as an assistive tool for clinical pharmacist in dental specialty hospital, with acceptable response quality across diverse pharmacotherapy domains. LLMs should be regarded as a promising assistive tool for clinical pharmacist rather than an independent decision maker. Further studies are still warranted since the rapid evolution and refinements of LLMs.

Keywords: Large language models, Artificial intelligence, Clinical pharmacy, Stomatological Hospital, Drug information

Introduction

Clinical pharmacists are important for pharmaceutical care system in stomatological hospital, which play a crucial role in ensuring the safety and effectiveness of drug treatment for oral diseases[1]. The responsibilities of clinical pharmacist include prescription rationality review, individualized medication consultation, monitoring and prevention of adverse drug events, etc. In particular, it is necessary to carry out professional pharmaceutical intervention for patients with combined systemic underlying diseases, multiple drug regimens and special drug treatments during the perioperative period[2]. Although the professional value of clinical pharmacists has been widely recognized, their daily clinical practice still faces many practical challenges. First, individualized medication consultation needs to be comprehensively assessed based on the patient's basic disease status, comorbidities and systemic organ function[3]. Second, the continuous update of evidence-based pharmacy evidence and the dynamic iteration of clinical guidelines have brought a heavy lifelong learning burden to pharmacists, and the professional threshold continue to rise[4]. Third, junior clinical pharmacists generally have insufficient clinical practice experience, and their professional ability growth is highly dependent on the teaching guidance of senior pharmacists, while the existing resources are difficult to cover the normal and immediate clinical decision support needs[5].

In recent years, artificial intelligence-based large language models (LLMs) have shown excellent performance in the fields of natural language understanding, professional knowledge retrieval and logical reasoning, and are expected to be used as an auxiliary tool to alleviate these challenges[6]. Among the current mainstream models, ChatGPT is one of the most widely validated generative large language models in the field of healthcare. It has been confirmed that it has good application potential in drug information query, clinical question answering and other tasks[7]. Gemini relies on a comprehensive global medical knowledge base with advanced multimodal information integration and reasoning capabilities[8]. As a representative of open-source LLM, DeepSeek has attracted attention due to its stable performance and high cost-

effectiveness[9].

An increasing number of studies have explored the application of LLMs in clinical pharmacy, reporting encouraging outcomes in answering drug-related questions and identifying adverse reactions[10]. But research within the specialized field of dental pharmacotherapy remains scarce and fragmented. Cirakoglu et al. explored the performance of ChatGPT 3.5 and Gemini 1.0 in providing knowledge about dental antibiotics use. They suggested that the reliability of LLMs remained limited in influencing dental clinical decision-making[11]. Isleyen et al. assessed the quality of responses generated by leading commercial LLMs including ChatGPT 5.3, Gemini 3.1 Pro, Claude 4.6 Sonnet in oral and maxillofacial surgery pharmacotherapy[12]. Published studies have focused on limited area such as dental antibiotics use and oral and maxillofacial surgery, yet multi-dimensional assessments spanning the full daily workflow of clinical pharmacists are conspicuously absent. It is highly necessary to systematically evaluate the performance of different LLMs as a tool for clinical pharmacists in daily pharmacotherapy practices[13].

Therefore, this study conducted a multi-dimensional comparative evaluation of ChatGPT-5.5 Instant, Gemini-3.5 Flash-Lite and DeepSeek-V3 as potential assistive tools for clinical pharmacists in dental specialty hospitals. The design and the research process for this study was illustrated in Figure 1. Based on dental pharmacotherapy guidelines and real-world clinical scenarios, 30 questions were curated, covering six domains: dosage and administration, drug interactions, contraindications, contraindications for drug combination, route of administration, and liver and kidney dysfunction. Responses from the three models were blindly and independently assessed by five experienced clinical pharmacists using a validated global quality scale (GQS). The associations between question difficulty and response quality, information source and readability of responses was also evaluated, thus aiming to provide multi-dimensional evidences for assistive utility of LLMs in clinical dental pharmacy practice.

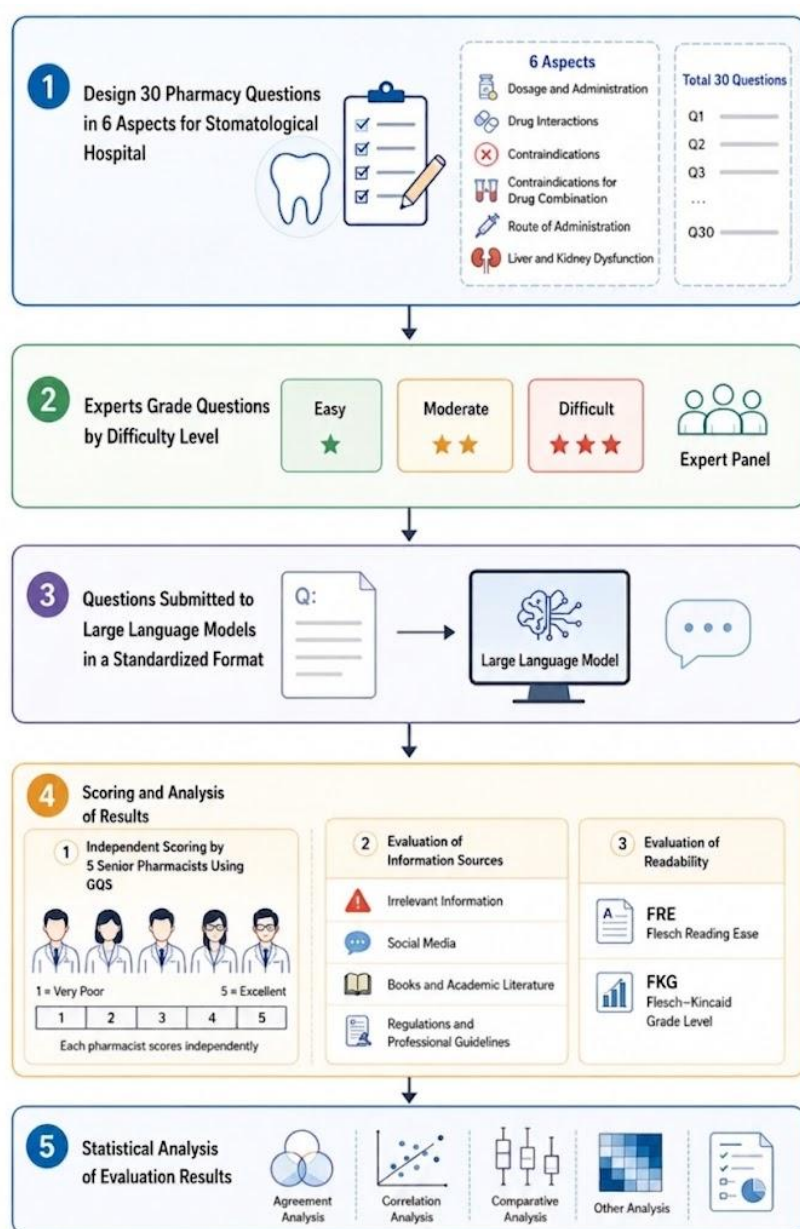


Figure 1. The flowchart of research design and process

Methods

Large Language Models

Three different large language models were selected to generate responds in this study: ChatGPT-5.5 Instant, Gemini3.5 Flash-Lite and DeepSeek-V3. To ensure consistency and reproducibility, all 30 curated questions were submitted to the respective models via their official web interfaces. All models were utilized in the standard configurations via their respective free tiers with advanced reasoning or extended-thinking modes disabled. All of the questions were queried in a new, empty

chat session to prevent context carryover or bias from previous prompts.

Questions

The questions addressed in this study were collected from the past routine pharmaceutical services provided by clinical pharmacists at stomatological hospital. Six common domains including dosage and administration, drug interactions, contraindications, drug incompatibilities, routes of administration, and liver and kidney dysfunction were predefined based on daily practice in the stomatological hospital. The questions were shown in Table 1, which were constructed as hypothetical cases solely for the purpose of evaluating the performance of different large language models, and did not involve any treatments for real patients. To ensure clinical relevance and comprehensiveness, pharmacists were asked to submit five typical questions from each category, reflecting the most frequent queries encountered in such settings. The difficulty of each question was subsequently rated independently by two senior clinical pharmacists, classified them into three levels: “easy”, “moderate”, or “difficult” according to the criteria as shown in Table 2. Disagreements were resolved through consultation with a third senior pharmacist until consensus was reached.

Table 1. List of the questions

Domain		Specific questions
Dosage and administration		Female, 40-45 years old, weight 50 kg, with recurrence squamous cell carcinoma of the head and neck. What is the recommended dose of pembrolizumab, and what concentration should be used for intravenous infusion?
		Female, 75-80 years old, with a salivary fistula, treated with atropine to suppress salivary secretion. What is the recommended dose?
		Male, 65-70 years old, with hypertrophy of the masseter and trapezius muscles. What is the maximum single dose of botulinum toxin?
		Male, 5-10 years old, weight 23.7 kg, scheduled for submandibular gland cystectomy. What is the recommended dosage of cefazolin?
		Female, 60-65 years old, following maxillofacial mass removal, has serum potassium of 2.7 mmol/L and requires intravenous potassium supplementation. What is the recommended dosage?
Drug interactions		Male, 60-65 years old, with mandibular fracture, reports that he underwent PCI for a myocardial infarction 6 months ago and has been taking

	clopidogrel bisulfate orally since then. Can omeprazole enteric-coated tablets be used concurrently to prevent perioperative stress-induced gastric mucosal lesions? Male, 50-55 years old, with Behçet's disease, taking 10 mg methotrexate tablets once weekly. Recently diagnosed with chronic gastritis due to epigastric burning. Is it safe to prescribe omeprazole enteric-coated tablets for his treatment? Female, 40-45 years old, with oral candidiasis, treated with fluconazole capsules. Has a history of hyperlipidemia and has been taking atorvastatin calcium tablets long-term. Which adverse reactions should be closely monitored during this treatment? Male, 46-50 years old, with squamous cell carcinoma of the left maxilla, receiving preoperative neoadjuvant chemotherapy with paclitaxel and nedaplatin. In what order should these two drugs be administered? Male, 55-60 years old, with trigeminal neuralgia, hypertension, and hyperlipidemia, treated with carbamazepine tablets (200 mg, twice daily), amlodipine besylate tablets (5 mg, once daily), and atorvastatin calcium tablets (20 mg, once nightly). Does this regimen require adjustment?
Contraindications	Male, 56-60 years old, with pemphigus vulgaris complicated by thrombocytopenic purpura. Can he take Tripterygium glycoside tablets? Male, 25-30 years old, undergoing third molar extraction, reports a history of aspirin-induced asthma. Which analgesic should be selected? Female, 25-30 years old, with oral lichen planus, reports macular degeneration in her right eye and is receiving abciximab injections. Which systemic immunosuppressive agents is appropriate for her treatment? Female, 25-30 years old, 25 weeks pregnant, with chronic periodontitis. Can she use povidone-iodine mouthwash? Female, 45-50 years old, with a history of epilepsy and severe periodontitis. Can ornidazole be used for the treatment?
Contraindications for drug combination	Female, 80-85 years old, underwent incision and drainage of an abscess. Cefotaxime was selected as the perioperative antimicrobial agent. Postoperatively, sodium lactate Ringer's solution was given for fluid replacement. Can cefotaxime sodium be added to sodium lactate Ringer's solution? Female, 70-75 years old, underwent tumor resection. Can vitamin C injection and inosine injection be administered in the same intravenous bag postoperatively? Female, 55-60 years old, with drug-induced liver injury. Can polyene

	phosphatidylcholine injection be added to sodium chloride injection?
	Male, 65-70 years old, with elevated transaminases. Can sodium diammonium glycyrrhizate injection be dissolved in 0.9% sodium chloride solution?
	Male, 65-70 years old, with marked swelling of the floor of the mouth following surgery for oropharyngeal squamous cell carcinoma, complicated by diabetes. Can dexamethasone sodium phosphate injection be dissolved in 0.9% sodium chloride solution?
Route of administration	Male, 35-40 years old, with erosive oral lichen planus. Can triamcinolone injection be administered via intravenous infusion?
	Male, 65-70 years old, with candidal stomatitis. Can nystatin tablets (1 million units) be administered orally three times daily?
	Female, 35-40 years old, with herpetic recurrent aphthous ulcers. Can dexamethasone sodium phosphate injection (5 mg) be administered via intravenous infusion once daily?
	Female, 65-70 years old, with trigeminal neuralgia, treated with a combination of dexamethasone (5 mg) and vitamin B12 injection. Can this treatment be administered via intravenous injection?
	Female, 60-65 years old, with hemangioma, treated with bleomycin hydrochloride for injection (5 mg). Can this treatment be given intravenously?
Liver and kidney dysfunction	Female, 70-75 years old, with infection in the right masseteric space, creatinine clearance 11- 34 mL/min. What is the recommended dose of cefazolin?
	Male, 65-70 years old, developed lower-extremity venous thrombosis following surgery for left-sided squamous cell carcinoma of the tongue. Creatinine clearance 20 mL/min. Which anticoagulant should be selected?
	Male, 45-50 years old, with erosive oral lichen planus and hepatic insufficiency (ALT 121, AST 98). Is the treatment with prednisolone appropriate?
	Male, 55-60 years old, admitted for maxillofacial space infection. Physician prescribed piperacillin-tazobactam for empirical treatment. Has a history of diabetic nephropathy, creatinine clearance 38 mL/ min. How should the dosage be adjusted?
	Female, 55-60 years old, with mandibular fracture and hyperglycemia, estimated glomerular filtration rate (eGFR) 19 mL/min. What is the recommended dose of metformin?

Table 2. Criteria for evaluating the difficulty level of questions

Difficulty level	Evaluation criteria
Easy	Involves basic concepts or direct application, and may require simple knowledge association or elementary analysis. Examples including recalling factual information or combining several known concepts to solve routine problems.
Moderate	Require integration and comprehensive application of knowledge, potentially involving interdisciplinary content or moderately complex logical reasoning. Example include multi-step problems solving or analyzing the causes and consequences of specific phenomena.
Difficult	Demand higher-order critical thinking and creative problem-solving skills, typically concerning cutting-edge topics, unresolved issues, or complex challenges that necessitate interdisciplinary approaches. Examples include designing experimental protocols or optimizing complex systems.

Generating Responses

The questions were submitted to the large language models in a standardized format, which comprises patient age, gender, drug name, dosage, and indication. The prompt was framed as follows: “You are a pharmacist at a stomatological hospital. Please answer the following questions, XXX. Please omit greetings and unnecessary filler, return the answer in a single paragraph without a title, and provide the link to the medication reference.” In addition, the original questions were regenerated (i.e., rephrased while preserving clinical content) and resubmitted to evaluate the effect of question wording on response accuracy across the predefined domains.

Evaluation of Model-Generated Responses

The quality of responses generated by the three AI models (ChatGPT, Gemini, and Deepseek) was independently assessed by five senior clinical pharmacists at stomatological hospitals. To minimize bias, each model was assigned a numeric code (model 1, model 2, model 3), ensuring that evaluators remained blinded to the identify of the corresponding model. The modified global quality scale (GQS) was used as the assessment tool, with scoring criteria encompassing factual accuracy, completeness, relevance, fluency, and readability (Table 3). This multi-dimensional and independent evaluation, based on established gold standards, was designed to

comprehensively reflect both the correctness and the clinical applicability of the responses by ChatGPT, Gemini, and Deepseek.

Table 3. Modified global quality scale (GQS)

Score	Evaluation criteria
1	Low quality, provides minimal information, with poor readability and flow, and insufficient content.
2	Poor overall quality, offers some information but lacks fluency and readability, with many essential aspects missing.
3	Moderate quality, covers key points adequately, with good readability and logical flow, and provides some practical value for clinical pharmacists.
4	Good quality, includes most relevant information, with clear readability and coherent flow, and is useful in clinical pharmacists practice.
5	Excellent quality, rich in information, with outstanding readability and logical flow, and provides substantial practical assistance to clinical pharmacists.

Evaluation of Information Sources

As per the standardized prompt, models were required to cite information sources within their generated responses. To ensure the clinical validity of the provided evidence, the reliability of these cited sources was systematically evaluated. The grading criteria utilized for this assessment are detailed in Table 4. If the model provided multiple sources for a single response, the mean reliability score across all cited references was calculated and assigned as the final score for this response.

Table 4. Evaluation criteria of information source

Score	Criteria
0	Irrelevant information
1	Social media
2	Academic books and literature
3	Professional regulations and guidance documents

Evaluation of Readability

To assess the linguistic complexity and comprehensibility of the model-generated responses, the readability was also evaluated. The linguistic accessibility of each response was quantified using two established metrics: the Flesch Reading Ease (FRE) score and the Flesch-Kincaid Grade (FKG) level. These metrics evaluate text complexity based on syntactic and semantic characteristics.

The FRE score measures overall reading difficulty on a standard scale from 0 to 100, where higher scores indicate text that is easier to read and comprehend. The FRE was calculated using the Equation 1:

$$FRE = 206.835 - 1.015 \left(\frac{N_w}{N_s} \right) - 84.6 \left(\frac{N_{syl}}{N_w} \right)$$

where N_w is the total number of words, N_s is the total number of sentences, and N_{syl} is the total number of syllables in the evaluated text.

The FKG score translates the reading difficulty into an equivalent U.S. academic grade level, providing a practical estimate of the educational background required to understand the text. Lower FKG scores signify greater readability. The FKG was determined by the Equation 2:

$$FKG = 0.39 \left(\frac{N_w}{N_s} \right) + 11.8 \left(\frac{N_{syl}}{N_w} \right) - 15.59$$

Statistical Analysis

Descriptive statistics were first calculated to summarize the Global Quality Scale (GQS) scores assigned by the clinical experts. The data were aggregated to determine the mean, median, variance, and interquartile range (IQR) for the model-generated responses.

Prior to conducting comparative analyses, the Shapiro-Wilk test was performed to assess the normality of the GQS score distributions. The formula can be shown in Equation 3:

$$W = \frac{\left(\sum_{i=1}^n a_i x_{(i)} \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where W represents the Shapiro-Wilk test statistic, n denotes the total sample size, $x_{(i)}$ is the i -th order statistic (i.e., the i -th smallest value in the sorted sample), and $\bar{x}$ is the sample mean. The constant a_i in the numerator represents the weight

coefficients derived from the expected values and the covariance matrix of the order statistics of a standard normal distribution. A calculated W value closer to 1 indicates that the sample data more closely conforms to a normal distribution.

The results of Shapiro-Wilk test indicated that the score distributions violated the assumptions of normality. Thus, non-parametric tests were employed for all comparative analyses. The Kruskal-Wallis test was applied to evaluate statistical differences in response quality. The Kruskal-Wallis test statistic (H) is determined by the Equation 4:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1)$$

where N is the total number of observations across all groups, k is the number of groups being compared, R_i is the sum of ranks for the i -th group, and n_i is the sample size of the i -th group.

In case where statistically significant differences were observed in the omnibus Kruskal-Wallis test, post-hoc pairwise comparisons were conducted utilizing the Mann-Whitney U test. The Mann-Whitney U statistic (U) was calculated using Equation 5:

$$\begin{aligned} U_1 &= n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \\ U_2 &= n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - R_2 \\ U &= \min(U_1, U_2) \end{aligned}$$

where n_1 and n_2 represent the sample sizes of the two independent groups being compared, and R_1 and R_2 denote the sum of ranks for the respective groups in the combined ranking. The final test statistic U is defined as the smaller value between U_1 and U_2 .

To strictly control the inflation of the Type I error probability arising from multiple comparisons, the Bonferroni correction was applied to the results of the pairwise tests. The adjusted significance level (p_{adj}) was determined by Equation 6:

$$\begin{aligned} m &= \frac{k(k-1)}{2} \\ p_{adj} &= \min(p_{raw} \times m, 1) \end{aligned}$$

where m denotes the total number of post-hoc pairwise comparisons (determined by the number of groups k), p_{raw} is the original, unadjusted significance value derived from the Mann-Whitney U test, and p_{adj} is the final corrected P -value. To maintain statistical validity, p_{adj} is capped at an upper limit of 1. Differences between any two evaluated groups were considered statistically significant only when $p_{adj} < 0.05$.

Finally, the Fisher-Freeman-Halton exact test was performed to investigate the categorical association between question difficulty and model accuracy. This extension of Fisher's exact test for $r \times c$ contingency tables was specifically used to determine whether an increase in difficulty level was significantly associated with a decline in response quality. All statistical analyses were computationally executed, and the two-tailed threshold for statistical significance was established at $\alpha = 0.05$.

Ethical Statement

This study exclusively evaluated artificial intelligence models and did not involve human participants or animal subjects, therefore, ethical approval and informed consent were not required.

Results

Descriptive Statistics

The response quality and statistical results were presented in Table 5. Descriptive statistics, including the Mean, Standard Deviation (SD), Median, Interquartile Range (IQR), Maximum, and Minimum values, were calculated for the responses generated by the three different LLMs. Preliminary descriptive analysis indicated that GPT-5.5 Instant yielded the most superior performance, recording the highest mean (4.307) and median (4.300) scores.

Table 5. Descriptive statistics of three different LLMs

	Mean	SD	Median	IQR	Min	Max
ChatGPT	4.307	0.367	4.300	4.000-4.600	3.600	4.800

Gemini	3.247	0.366	3.400	3.000-3.400	2.600	3.800
DeepSeek	3.673	0.631	3.700	3.250-4.200	2.400	4.600

Normality Assessment

Prior to comparing the generation quality across the evaluated LLMs the Shapiro-Wilk test was utilized to assess the normality of the subjective evaluation scores for the three models across all 30 questions (with the significance level set at $\alpha=0.1$). The results indicated that although the score distribution for Deepseek did not significantly deviate from a normal distribution ($W=0.9450, P=0.1238$), the scores for ChatGPT ($W=0.9113, P=0.0161$) and Gemini ($W=0.8858, P=0.0038$) both violated the assumption of normality. Given that the datasets did not uniformly satisfy the prerequisite conditions for parametric testing, the Kruskal-Wallis test was employed for the subsequent inter-group comparative analyses to ensure the validity of the statistical inferences.

Comparison of Different LLMs

To verify the overall differences in response quality among three LLMs, a Kruskal-Wallis test was initially conducted. The results were shown in Figure 2. The analysis revealed a highly significant overall statistical difference in the evaluation scores across the three models ($H=45.8326, P<0.001$). To further elucidate the specific inter-model differences, post-hoc pairwise comparisons were subsequently performed utilizing the Mann-Whitney U test, with a Bonferroni correction applied to control the Type I error rate associated with multiple comparisons. The post-hoc results demonstrated significant differences across all pairwise combinations: distinct score distributions were observed between ChatGPT and Gemini (Bonferroni-adjusted $P<0.001$), ChatGPT and DeepSeek (Bonferroni-adjusted $P<0.001$), and Gemini and DeepSeek (Bonferroni-adjusted $P=0.0070$). These statistical findings robustly demonstrated that the performance of the three models on this task fell into entirely distinct tiers, each exhibiting a significant gradient in generation quality. ChatGPT achieved the highest median score with a relatively concentrated IQR. DeepSeek ranked second in overall

quality and its wider IQR suggests greater variability in performance across different questions. Conversely, Gemini recorded the lowest median evaluation score.

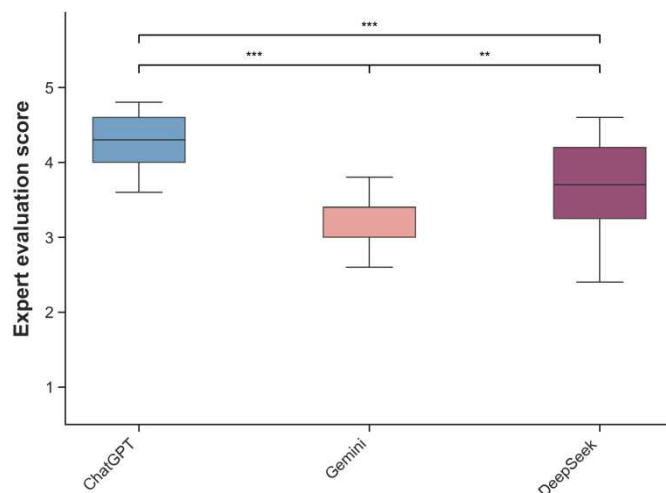


Figure 2. Difference of the evaluation scores across the three models for 30 questions. $n=150$, $**P < 0.01$, $***P < 0.001$.

Comparison of Response Quality to Different Domains of Questions

To evaluate the models' specific capabilities, a Kruskal-Wallis test was conducted to compare the performance of the three models across six distinct domains: Dosage and administration, Drug interactions, Contraindications, Incompatibilities, Route of administration, and Hepatic/Renal impairment. The distributions of the expert evaluation scores and the overall capability profiles were presented in Figure 3. As visually corroborated by the radar chart, ChatGPT achieved the highest overall performance, forming the outermost capability perimeter across all six categories. Furthermore, the omnibus tests and post-hoc pairwise comparisons revealed statistically significant differences in five of the six domains, with incompatibilities being the sole exception where no significant variance was observed among the models. Specifically, the boxplots indicated that ChatGPT scored significantly higher than Gemini in five categories ($P < 0.05$), and significantly outperformed DeepSeek in the "Route of administration" domain ($P < 0.05$). Conversely, DeepSeek consistently occupied the intermediate tier, while Gemini exhibited the lowest median scores across the majority of the evaluated domains.

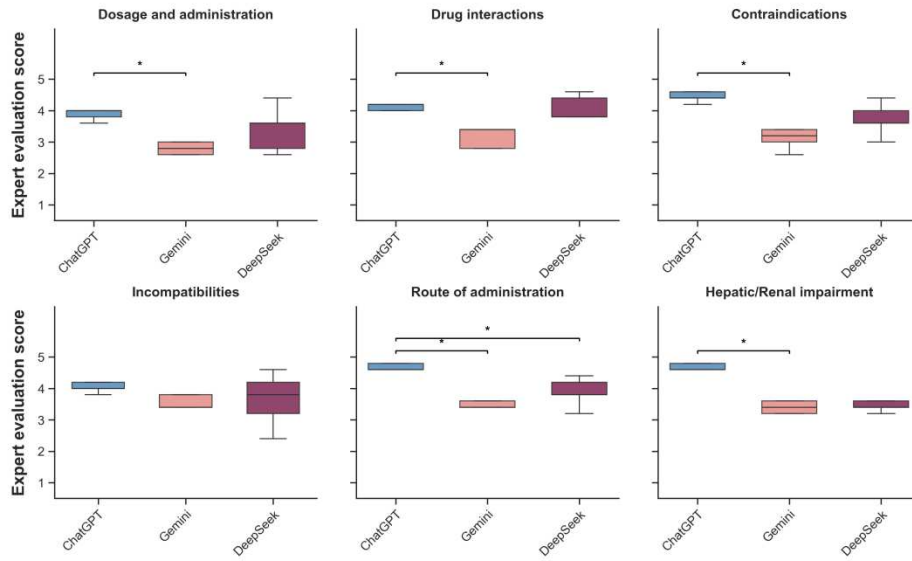
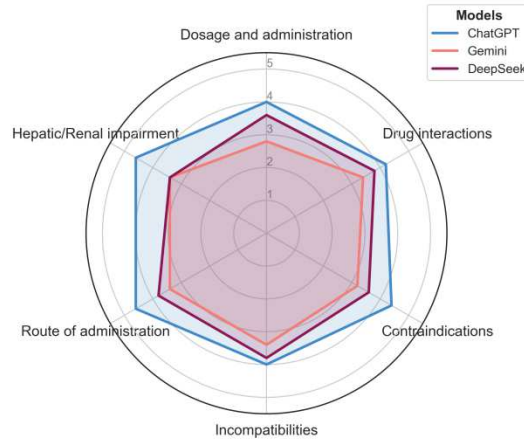
A**B**

Figure 3. Comparison results of the response quality to different domains. A. boxplots. B. Radar plot. $n=25$, $*P < 0.05$.

Comparison of Response Quality to Difficulty Level of Questions

The visual representations of the evaluation scores further corroborate the statistical findings. As depicted in Figure 4A, ChatGPT consistently formed the outermost perimeter across all three difficulty levels (Easy, Moderate, and Difficult), visually underscoring its dominant and well-rounded performance. DeepSeek maintained the intermediate triangle, while Gemini formed the innermost boundary, indicating the lowest overall capability profile regardless of query complexity. More granular insights were revealed through the box plot distributions (Figure 4B). The box plots explicitly highlighted the statistically significant disparity between ChatGPT and

Gemini ($P < 0.001$) across every difficulty tier. Notably, while ChatGPT demonstrated tightly clustered IQRs with consistently high medians, DeepSeek exhibited a progressively widening IQR as the difficulty increases. This was particularly evident in the "Difficult" category, where DeepSeek lower whisker drops significantly, suggesting that although its median performance remains superior to Gemini, its response quality becomes increasingly unstable and highly variable when confronted with complex clinical queries. In contrast, Gemini, despite yielding the lowest median scores, maintains a relatively narrow IQR across all levels, indicating a consistent, albeit lower-tier, performance ceiling.

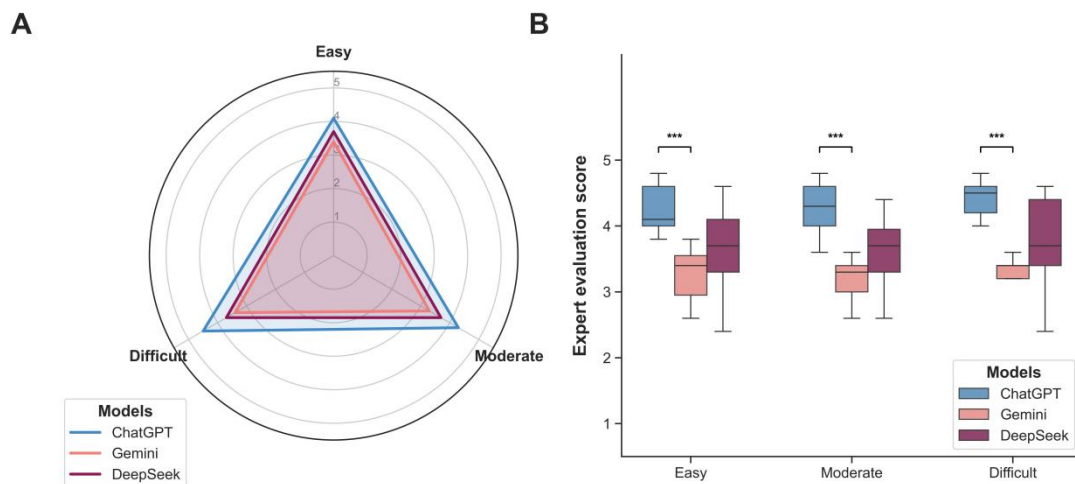


Figure 4. Comparison results of the response quality to the difficulty level. A. boxplots. B. Radar plot. $n=5$, *** $P < 0.001$.

Association between Response Quality and Difficulty Level of Questions

To further investigate whether there is an association between the responses quality and the difficulty level of questions, the Fisher-Freeman-Halton (FFH) exact test alongside Kendall's τ_b trend test were conducted. The overall analysis revealed no statistically significant correlation between the predefined complexity levels and the evaluation scores across all generated responses (FFH $P=0.9318$; Kendall's $\tau_b=-0.0231$, $P=0.8066$). Furthermore, subgroup analyses stratified by individual models demonstrated consistent results: none of three models exhibited any significant monotonic trend in their accuracy as query complexity increased (all $P>0.05$). As

visually corroborated by the heatmap of the accuracy distribution in Figure 5, the response quality for each individual model remained remarkably consistent across different difficulty levels (Easy, Moderate, and Difficult). Rather than fluctuating with question difficulty, the score distributions exhibited strict horizontal stability within each model's inherent capability tier. For instance, ChatGPT maintained a steady output regardless of query complexity, with its scores exclusively and consistently falling into level 5 (4 to 5 instances) and level 4 (5 to 6 instances) across all three difficulty tiers. Similarly, the performance of Gemini showed no sensitivity to increased difficulty, persistently yielding 7 to 8 instances of level 3. DeepSeek also demonstrated an unfluctuating intra-model pattern, peaking uniformly at score 4 (5 to 7 instances per level) irrespective of whether the question was simple or complex. These combined visual and statistical findings strongly indicated that the performance limitations observed in this study were fundamentally driven by the models' inherent reasoning and medical knowledge capacities, rather than the varying complexity of the inputted clinical queries.

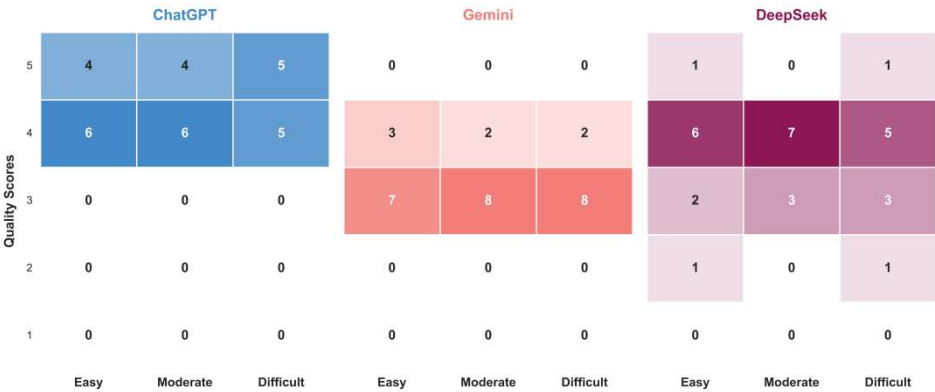


Figure 5. Association between responses quality and difficulty level of questions.

Comparison of the Information Source

To further evaluate the solid foundation of response, the quality and reliability of the information sources provided by the three large language models were also assessed. A Kruskal-Wallis omnibus test was conducted to determine if there were overarching differences in the source evaluation scores among the models. According to the statistical output, the analysis revealed no statistically significant differences across

ChatGPT, Gemini, and DeepSeek ($H=3.3808$, $P=0.1844$). Since the omnibus test did not reach the threshold for statistical significance ($P \geq 0.05$), post-hoc pairwise comparisons were strictly completely bypassed to prevent Type I error inflation. Despite the lack of statistical significance in the overall group comparison, the box plot distribution in Figure 6 reveals distinct descriptive patterns regarding the internal stability of each model. DeepSeek exhibited the most concentrated IQR with a slightly higher median, suggesting relatively consistent behavior when providing information sources. ChatGPT displayed moderate variability with a median score of 2.0. Conversely, Gemini demonstrated extreme dispersion; its IQR spanned the entire possible scoring scale from 0 to 3. This indicated highly unstable performance, meaning Gemini's ability to cite or rely on quality sources fluctuates wildly depending on the query. Ultimately, while visual variations in consistency exist, the statistical evidence concluded that none of the three models demonstrated a significantly superior capability in providing reliable clinical information sources.

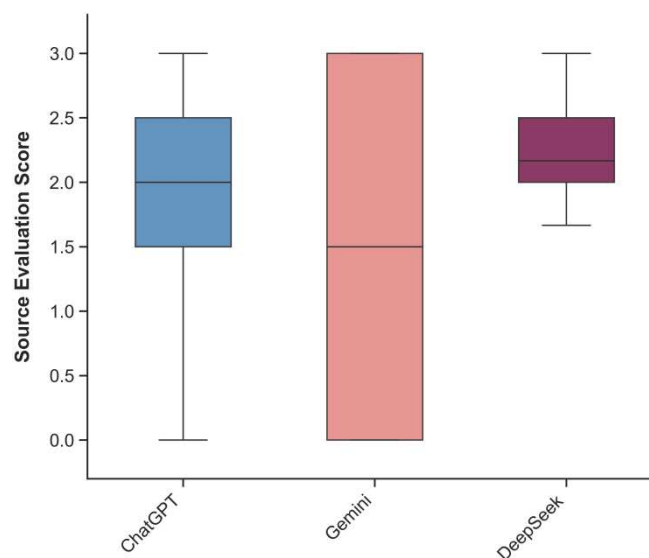


Figure 6. Comparison results of information source between different models. $n=30$.

Comparison of the Readability of Response

To evaluate the linguistic accessibility of the generated responses, the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKG) metrics were applied. A Kruskal-Wallis omnibus test was conducted to determine if the text readability significantly varied among the three models. The statistical analysis revealed absolutely no significant differences across the models for either the FRE metric ($H=0.4361, P=0.8041$) or the FKG metric ($H=1.6269, P=0.4433$). Given that the omnibus *P-values* were substantially above the 0.05 significance threshold, post-hoc pairwise comparisons were bypassed. The visual representations comprehensively corroborate this lack of statistical variance. As illustrated in Figure 7A, there was substantial overlap in the interquartile ranges and whiskers across all three models for both FRE and FKG scores. While Gemini exhibited a slightly wider dispersion of scores (indicating greater variability in text complexity), the median readability scores remained remarkably similar across the board. Furthermore, Figure 7B demonstrated that the readability of the generated text fluctuates more in response to specific clinical queries rather than the choice of the language model. Overall, the data highlighted a consistent trend: the high FKG scores (predominantly ranging between 15 and 22, indicating college or graduate-level reading difficulty) and correspondingly low FRE scores (mostly falling below 30) suggested that all three models inherently generate advanced, academic-level text. Consequently, none of the evaluated models demonstrated a superior capability in providing simpler or more accessible explanations suitable for general patient comprehension.

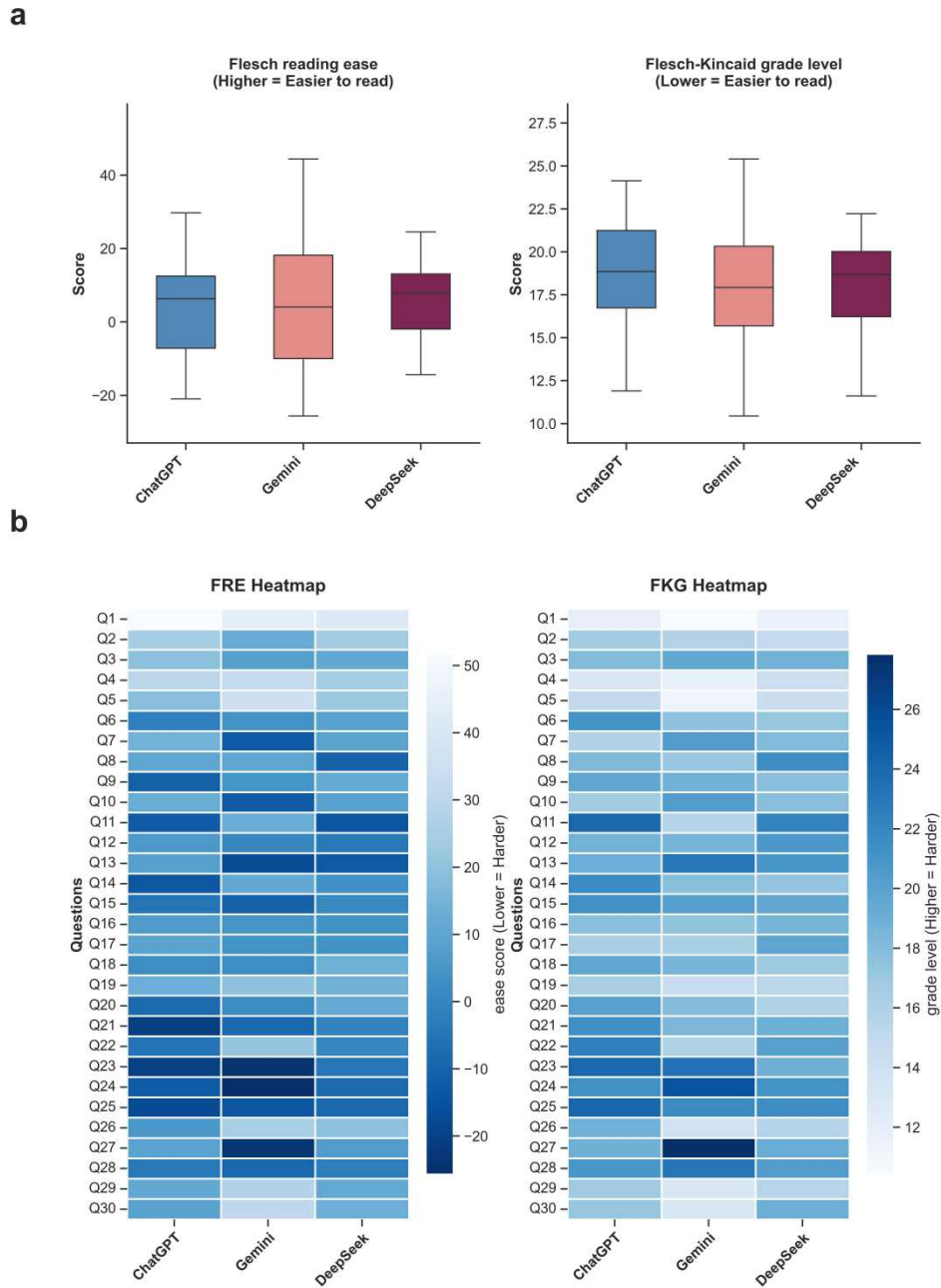


Figure 7. Readability of response generated by different models. A. Boxplot. B. Heatmap. n=30.

Discussion

This study compared the application potential of three LLMs (ChatGPT-5.5 Instant, Gemini-3.5 Flash-Lite and DeepSeek-V3) in assisting clinical pharmacists at stomatological hospital. The study comprised 30 questions spanning six domains: dosage and administration, drug interactions, contraindications, combination

contraindications, administration routes, and hepatic and renal dysfunction. Responses generated by the three LLMs were evaluated by five experienced clinical pharmacists. Statistical analysis of the scores for response quality among three LLMs revealed a distinct performance hierarchy: ChatGPT-5.5 Instant > DeepSeek-V3 > Gemini-3.5 Flash-Lite. Published studies have reported heterogeneous results regarding the comparative quality of LLM responses across various dental specialties. Specifically, Gemini 3 Flash scored significantly higher than ChatGPT-5 when addressing common questions in prosthodontics[14]. In response to open-ended questions concerning dental trauma - including crown-root fractures, dislocation injuries, and avulsion, Gemini Pro achieved the highest scores on most items, substantially outperforming ChatGPT-4o Plus and DeepSeek-V3[15]. Sismanoglu et al. evaluated the accuracy of LLMs in answering queries on dental bleaching, finding that ChatGPT demonstrated significantly greater accuracy than Gemini for textbook-based questions, whereas both models showed comparable performance for patient-facing inquiries[16]. Conversely, Azhari et al. assessed LLM performance on a multiple-choice examination in caries, reporting that Gemini surpassed ChatGPT in both accuracy and pass rates across all mock tests, suggesting that Gemini offers more reliable performance when handling complex caries content[17]. In a real-world diagnostic scenario, Claude and Manus exhibited numerically higher diagnostic performance and greater response stability than ChatGPT, although these differences did not reach statistical significance[18]. Taken together, these performance variations across LLMs are attributable to multiple factors, including the specific subject domain, task formats (e.g., patient generated queries versus pharmacist oriented questions), evaluation metrics, the domain specificity of test datasets, and model version. Within the specific context of pharmacotherapy, our findings indicated that ChatGPT-5.5 Instant exhibited superior performance, whereas Gemini-3.5 Flash-Lite did not confer a significant advantage in addressing these curated questions.

We performed the Fisher–Freeman–Halton exact test on the scores of response quality assigned to the LLMs for questions of varying difficulty levels and found no significant correlation between difficulty and the scores. Unexpectedly, the scores did

not decline as the difficulty increased. This result diverged from findings reported in previous studies. For example, Gao et al. observed that, when addressing prosthodontics-related questions, the response accuracy of both standard and optimized LLMs decreased with increasing difficulty[19]. Similarly, a study on routine hospital pharmacy questions revealed a significant inverse relationship between the complexity of questions and the accuracy of ChatGPT's responses[20]. We speculated that this discrepancy may stem from the different pharmacy domains during question design. Given that LLMs possess more extensive knowledge bases, superior information retrieval capabilities, and stronger logical reasoning abilities than humans, they may maintain better performance on difficult questions, thereby attenuating the expected difficulty–accuracy gradient.

Analysis of the information sources revealed that, although the LLMs were capable of providing references, drug package inserts, or FDA approval documents, the provided information frequently contained both irrelevant content and fabricated details. Several studies have examined the authenticity of references generated by LLMs in dentistry and related medical fields, consistently reporting the presence of invented sources. Teke et al. reported that Claude Sonnet 4.5 and Gemini 3 Flash produced the highest proportions of accurate and relevant references (85.7% and 84.2%, respectively), whereas DeepSeek V3.2 exhibited the highest fabrication rate (55.6%)[14]. Liu et al. investigated the verifiability of references provided by different LLMs and found that all models generated reference illusions, with 40% of the cited references being unverifiable. The proportions of verifiable references for Qwen3, Claude 4.0, ChatGPT o3-pro, Gemini 2.5 Pro, and DeepSeek-R1 were 62.3%, 61.9%, 59.3%, 58.0%, and 56.1%, respectively[21]. Other literature suggests that approximately 51% of references supplied by LLMs are fabricated. Although newer versions, such as GPT-4, have reduced this proportion (from 55% for GPT-3.5 to approximately 18% for GPT-4), citation errors remain pervasive. Even among authentic references, 43% of GPT-3.5 citations and 24% of GPT-4 citations contained substantial inaccuracies. While OpenAI recently released GPT-5 with claims of further reduction in false citations, independent evidence supporting this improvement remains

insufficient[22]. The cause of this phenomenon lies in the fundamental working principle of LLMs: they are essentially pattern-based text generators rather than knowledge retrieval systems. When asked for references, an LLM attempts to generate textual patterns that resemble authentic citations—including author names, journal titles, publication years, volume numbers, and other bibliographic elements. However, as it does not genuinely comprehend the meaning of a "reference" nor have access to a live database for verification, it readily produces well-formatted but fabricated citations[23].

Readability assessments of the responses generated by the three LLMs revealed no significant inter-model differences. Nevertheless, both the FRE and FKG scores indicated that all three models produced texts of considerable complexity, characterized by poor readability. This finding aligns with previously published literature[24-26]. This also underscored an inherent tension between the provision of accurate, comprehensive information and the maintenance of good readability. The primary explanation for the generally poor readability of LLM-generated texts lies in their training objectives and evaluation criteria, which heavily prioritize informational richness, professional accuracy, and linguistic coherence over clarity and simplicity. Trained on vast corpora of academic literature and textbooks, these models tend to accumulate specialized terminology while lacking the capacity to accommodate patients' cognitive levels and health literacy, often producing outputs that far exceed the reading comprehension threshold of the general population. Future research should aim to address this deficiency.

Although the performance of the open-source LLM DeepSeek does not yet match that of the closed-source ChatGPT, it surpasses that of the closed-source Gemini. This finding implies that, despite the general perception that open-source LLMs underperform their closed-source counterparts, they still hold considerable potential as supportive tools for clinical pharmacists. Open-source models offer several distinct advantages. First, they enhance data privacy and security, as they can be deployed entirely on local infrastructure, ensuring that patient data never leaves the hospital's internal network and thus fundamentally circumventing the risk of data breaches - a

critical consideration when handling sensitive clinical information, including patient identities and oral imaging data[27]. Second, open-source models are more cost-effective in the long term, eliminating recurring API access fees, the primary expenditures consist of initial hardware acquisition and subsequent maintenance, rendering the total cost of ownership substantially lower than that of closed-source alternatives[28]. Third, they afford a high degree of customizability that extends beyond superficial external knowledge retrieval or prompt engineering. Pharmacists can fine-tune open-source LLMs to encode non-public, heterogeneous clinical data - such as institution-specific post-medication consensus protocols, formulation standards, and even electronic medical record documentation practices - directly into the model's parameter space. In summary, while open-source LLMs still exhibit notable performance limitations, they possess unique strengths that warrant further exploration[29].

Our study indicated that LLMs have the potential to serve as assistive tools for clinical pharmacists. Nevertheless, it is important to emphasize that LLMs have not yet been approved for clinical medication decision-making[30]. Several inherent shortcomings may impede their practical application. First, LLMs are plagued by hallucination. LLMs may generate responses that appear coherent but are factually incorrect, particularly when addressing complex logic, rare disease pharmacotherapy, or extreme dose conversions[31]. Second, LLMs cannot replace real-time drug prescribing information or authoritative databases. Drug labels, clinical guidelines, FDA approvals, and hospital formularies are constantly updated, whereas LLMs are subject to knowledge cutoffs and may not reflect the most current evidence. Consequently, decisions regarding indications, dosing, hepatic or renal impairment adjustments, contraindications, and drug interactions should never rely solely on model-generated outputs[32]. Third, legal and ethical responsibility issues arise, as LLMs lack licensed pharmacist credentials and their generated recommendations carry no legal authority. The ultimate prescribing and dispensing decisions must remain with qualified professionals; in the event of erroneous LLM suggestions, the model cannot bear liability for medical adverse events[33].

Conclusion

This study systematically evaluated the performance of LLMs, including ChatGPT-5.5 Instant, Gemini-3.5 Flash-Lite and DeepSeek-V3, as a tool for clinical pharmacists in dental specialty hospital. Thirty questions were curated from daily pharmacotherapy practice for three LLMs and the response quality was assessed by multi-dimensional comparison. The results indicated that ChatGPT reached the highest response quality, followed by DeepSeek and Gemini. The response quality was not significantly correlated with question difficulty level, and the quality of information sources and readability were not significant among three LLMs. Taken together, our study indicated that LLMs, especially ChatGPT, have potential to act as an assistive tool for clinical pharmacist in stomatological hospitals, with acceptable response quality across diverse pharmacotherapy domains. LLMs should be regarded as a promising assistive tool for clinical pharmacist rather than an independent decision maker.

Acknowledgements

The authors acknowledge the Military Medical Innovation Center, Fourth Military Medical University for the assistance.

Authors' contributions

J Zhang, and S Hao contributed to the study concept and design, collected data, and was involved in drafting the manuscript. B Feng and S Liu contributed to review the questions for accuracy and evaluability, and grad the difficulty levels. F Jin participated in revising the manuscript. H Li contributed to data analysis and interpretation, and participated in revising the manuscript. J Ju performed all asking the questions to language models and participated in revising the manuscript. All authors have approved the final manuscript and agreed to accountable for all aspects of the work. J Zhang and S Hao contributed equally to this work.

Funding

No funding.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Mahmood S, Mouna S, Jankowski K, Trejo L: **A Pharmacist's Role in a Dental Clinic: Establishing a Collaborative and Interprofessional Education Site.** *Innovations in pharmacy* 2020, **11**(2).
2. Thabit AK, Aljereb NM, Khojah OM, Shanab H, Badahdah A: **Towards Wiser Prescribing of Antibiotics in Dental Practice: What Pharmacists Want Dentists to Know.** *Dentistry journal* 2024, **12**(11).
3. Choi HJ, Stewart AL, Tu C: **Medication discrepancies in the dental record and impact of pharmacist-led intervention.** *International dental journal* 2017, **67**(5):318-325.
4. Hajj A, Hallit S, Azzo C, Abdou F, Akel M, Sacre H, Salameh P, Rabbaa Khabbaz L: **Assessment of knowledge, attitude and practice among community pharmacists towards dental care: A national cross sectional survey.** *Saudi pharmaceutical journal : SPJ : the official publication of the Saudi Pharmaceutical Society* 2019, **27**(4):475-483.
5. Lygre H, Kjome RLS, Choi H, Stewart AL: **Dental providers and pharmacists: a call for enhanced interprofessional collaboration.** *International dental journal* 2017, **67**(6):329-331.
6. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A, Fries JA, Wornow M, Swaminathan A, Lehmann LS *et al*: **Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review.** *Jama* 2025, **333**(4):319-328.
7. Iqbal U, Tanweer A, Rahmanti AR, Greenfield D, Lee LT, Li YJ: **Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis.** *Journal of biomedical science* 2025, **32**(1):45.
8. Arpaci A, Ozturk AU, Okur I, Sadry S: **Evaluation of the accuracy of ChatGPT-4 and Gemini's responses to the World Dental Federation's frequently asked questions on oral health.** *BMC oral health* 2025, **25**(1):1293.
9. Sandmann S, Hegselmann S, Fujarski M, Bickmann L, Wild B, Eils R, Varghese J: **Benchmark evaluation of DeepSeek large language models in clinical decision-making.** *Nature medicine* 2025, **31**(8):2546-2549.
10. Pais C, Liu J, Voigt R, Gupta V, Wade E, Bayati M: **Large language models for preventing medication direction errors in online pharmacies.** *Nature medicine* 2024, **30**(6):1574-1582.

11. Çırakoğlu NY, Doğan M, Duymaz BC: **Comparative evaluation of artificial intelligence language models on knowledge of dental antibiotic use.** *BMC oral health* 2026, **26**(1).
12. Isleyen M, Aydemir A: **Do large language models differ in their pharmacology-related response quality for oral and maxillofacial surgery? a blinded expert benchmark study.** *BMC oral health* 2026.
13. Tordjman M, Liu Z, Yuce M, Fauveau V, Mei Y, Hadjadj J, Bolger I, Almansour H, Horst C, Parihar AS *et al*: **Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning.** *Nature medicine* 2025, **31**(8):2550-2555.
14. Teke E, Borsöken AG: **A comparative evaluation of large language models in diagnosis and treatment planning in restorative dentistry.** *BMC oral health* 2026.
15. Kaya D: **Can artificial intelligence applications be utilised in the management of dental trauma? A comparison of ChatGPT, Gemini, and DeepSeek.** *BMC oral health* 2026.
16. Sismanoglu S, Kotan SS, Isik V: **Accuracy and Temporal Consistency of ChatGPT and Gemini in Responding to Textbook and Patient-Oriented Dental Bleaching Questions: A Multi-Session Comparative Study.** *Journal of esthetic and restorative dentistry : official publication of the American Academy of Esthetic Dentistry [et al]* 2026, **38**(8):1553-1563.
17. Azhari AA, Ahmed WM, Alhamadani A, Alfaraj A, Zhang M, Lu CT: **Assessing the Efficacy of Artificial Intelligence Platforms in Answering Dental Caries Multiple-Choice Questions: A Comparative Study of ChatGPT and Google Gemini Language Models.** *Dentistry journal* 2026, **14**(2).
18. Madfa AA, Alshammari AF, Anazi BA, Alenezi YE, Alkurdi KA: **Accuracy and reliability of Manus, ChatGPT, and Claude in case-based dental diagnosis.** *Frontiers in oral health* 2025, **6**:1686090.
19. Gao S, Wang ZA, Gao Z, Liu J, Zhang H, Pan S, Zhou Y: **Performance of Enhanced Large Language Models on Prosthodontic Multiple-Choice Questions.** *International dental journal* 2026, **76**(2):109441.
20. van Nuland M, Lobbezoo AH, van de Garde EMW, Herbrink M, van Heijl I, Bognår T, Houwen JPA, Dekens M, Wannet D, Egberts T *et al*: **Assessing accuracy of ChatGPT in response to questions from day to day pharmaceutical care in hospitals.** *Exploratory research in clinical and social pharmacy* 2024, **15**:100464.
21. Liu L, Ma K, Wang Y: **Performance of five large language models in oral and maxillofacial surgery exam questions: a comparative study.** *BMC oral health* 2026, **26**(1).
22. Liu H: **Fabricated citations in the age of AI: A wake-up call for editors, reviewers, and authors.** *Journal of dental sciences* 2026, **21**(1):679-680.
23. Alhazmi N, Alshehri A, BaHammam F, Philip M, Nadeem M, Khanagar S: **Can Large Language Models Serve as Reliable Tools for Information in Dentistry? A Systematic Review.** *International dental journal* 2025, **75**(4):100835.
24. Özüdoğru S, Tokgöz Kaplan T: **Evaluation of the accuracy, quality, and readability of large language models on local anesthesia and general anesthesia-sedation in dentistry.** *BMC anesthesiology* 2026, **26**(1).
25. Ozdemir ZM, Yapici E: **Evaluating the Accuracy, Reliability, Consistency, and Readability of Different Large Language Models in Restorative Dentistry.** *Journal of esthetic and restorative dentistry : official publication of the American Academy of Esthetic Dentistry [et al]* 2025, **37**(7):1740-1752.

26. Çolak R, Çolak MK, Cicek O: **Comparative evaluation of AI-driven large language models for periodontal–orthodontic patient questions: accuracy, comprehensiveness, readability, and temporal consistency.** *BMC oral health* 2026.
27. Zhang H, Wang D, Xu Y, Han S, Wang G: **Applications of DeepSeek in Medicine: Bibliometric Analysis and Scoping Review.** *Journal of medical Internet research* 2026, **28**:e93354.
28. Wang Y, Han Y, Han K, Shen J: **Does DeepSeek curb the surge of energy consumption in data centers?** *Innovation (Cambridge (Mass))* 2025, **6**(9):100944.
29. Jin I, Tangsrivimol JA, Darzi E, Hassan Virk HU, Wang Z, Egger J, Hacking S, Glicksberg BS, Strauss M, Krittanawong C: **DeepSeek vs. ChatGPT: prospects and challenges.** *Frontiers in artificial intelligence* 2025, **8**:1576992.
30. Giacobbe DR, Marelli C, La Manna B, Padua D, Malva A, Guastavino S, Signori A, Mora S, Rosso N, Campi C *et al*: **Advantages and limitations of large language models for antibiotic prescribing and antimicrobial stewardship.** *npj antimicrobials and resistance* 2025, **3**(1):14.
31. van Kessel R, Anderson M, McMillan B, Matthews MR, Rust P, Percy P, Nasir K, Mossialos E: **Omission and hallucination prevalence of clinical guidelines in diagnostic large language model outputs.** *BMJ health & care informatics* 2026, **33**(1).
32. Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, Lee Y, Mayol J, Antoniou SA, Thirunavukarasu AJ *et al*: **Large Language Models for Chatbot Health Advice Studies: A Systematic Review.** *JAMA network open* 2025, **8**(2):e2457879.
33. Zhao W, Yang X, Lyu Z, Xu C, Guan Z: **Road of Large Language Model: Source, Challenge, and Future Perspectives.** *Research (Washington, DC)* 2025, **8**:0655.