International Journal of Dental Hygiene

WILEY

International Journal of Dental Hygiene

ORIGINAL ARTICLE OPEN ACCESS

Comparative Analysis of the Response Accuracies of Large Language Models in the Korean National Dental Hygienist Examination Across Korean and English Questions

Eun Sun Song | Seung-Pyo Lee

Department of Oral Anatomy, Dental Research Institute, School of Dentistry Seoul National University, Seoul, South Korea

Correspondence: Seung-Pyo Lee (orana9@snu.ac.kr)

Received: 10 May 2024 | **Revised:** 29 August 2024 | **Accepted:** 26 September 2024

Funding: This work was supported by the Korea Medical Device Development Fund grant funded by the Korean Government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: 1711196792, RS-2023-00253380).

Keywords: artificial intelligence | ChatGPT | dental hygienist | Gemini | large language models | licensing examination

ABSTRACT

Introduction: Large language models such as Gemini, GPT-3.5, and GPT-4 have demonstrated significant potential in the medical field. Their performance in medical licensing examinations globally has highlighted their capabilities in understanding and processing specialized medical knowledge. This study aimed to evaluate and compare the performance of Gemini, GPT-3.5, and GPT-4 in the Korean National Dental Hygienist Examination. The accuracy of answering the examination questions in both Korean and English was assessed.

Methods: This study used a dataset comprising questions from the Korean National Dental Hygienist Examination over 5 years (2019–2023). A two-way analysis of variance (ANOVA) test was employed to investigate the impacts of model type and language on the accuracy of the responses. Questions were input into each model under standardized conditions, and responses were classified as correct or incorrect based on predefined criteria.

Results: GPT-4 consistently outperformed the other models, achieving the highest accuracy rates across both language versions annually. In particular, it showed superior performance in English, suggesting advancements in its training algorithms for language processing. However, all models demonstrated variable accuracies in subjects with localized characteristics, such as health and medical law.

Conclusions: These findings indicate that GPT-4 holds significant promise for application in medical education and standardized testing, especially in English. However, the variability in performance across different subjects and languages underscores the need for ongoing improvements and the inclusion of more diverse and localized training datasets to enhance the models' effectiveness in multilingual and multicultural contexts.

1 | Introduction

Artificial intelligence (AI) and deep learning have made significant strides in various fields, notably in medicine [1–3]. AI applications now span disease prediction, diagnostic assistance, drug development, and medical data analysis [4–6]. A particular focus has been on large language models (LLMs) like ChatGPT, BingChat, and Gemini. These models, launched by companies like OpenAI and Google, are revolutionizing medical education and practice by processing and generating human language at an unprecedented scale [7, 8].

LLMs have shown promising results in medical licensing exams worldwide, demonstrating their potential to understand complex medical knowledge. Studies have highlighted their ability to pass exams like the United States Medical Licensing Examination (USMLE) and other national medical exams in Japan, and China [9–12]. Davies et al. [13] evaluated ChatGPT 3.5 on the Faculty of Public Health Diplomate exam, finding it passed 3 out of 4 papers, excelling in research methods but with concerns about factual accuracy. Similarly, Wang, Shen, [14] and Chen assessed ChatGPT 3.5 on the Taiwanese Pharmacist Licensing Exam, where it failed but showed better performance in English than in Chinese, suggesting potential for improvement through deep learning. These findings suggest that ChatGPT could be a valuable tool in professional education, while also highlighting its current limitations, particularly in its performance with non-English languages, emphasizing the need for further research in this area.

There is a paucity of research comparing and evaluating the efficacy of LLMs in the context of the national examination for dental hygienists in Korea, indicating a clear need for further investigation in this area. This study aims to address this gap by evaluating and comparing the performance of Gemini, GPT-3.5, and GPT-4 on the Korean National Dental Hygienist Examination. The study assesses the models' response accuracy in both Korean and English, providing insights into their applicability in multilingual environments and their potential to enhance dental hygiene education.

2 | Materials and Methods

This study did not involve human or animal participants, and ethical approval was not required.

2.1 | LLMs

Gemini (Google), Chat GPT-3.5 (OpenAI), and Chat GPT-4 (OpenAI) were used in this study. At the time of the study (March and April 2024), GPT-3.5 was equipped with information accumulated until January 2022, while the information in GPT-4 was updated through April 2023.

2.2 | Input Data Sets From the Korean Dental Hygienist Examination

The data utilized in this research included exam questions from the Korean National Dental Hygienist Examination conducted over 5 years, from December 2019 to December 2023. The questions and answers are publicly available on the official website.

The examination consists of 200 questions designed to evaluate the competence of oral hygiene experts. To pass, candidates must score at least 60% overall and at least 40% in each subject, with each question offering five multiple-choice options and requiring the selection of the most correct answer. The exam covers 21 subjects, primarily grouped into health and medical law, dental morphology, oral prophylaxis, and periodontology; it includes foundational medical sciences and knowledge necessary for actual dental clinical practice. Table 1 shows the subjects and number of questions in the Korean National Dental Hygienist Examination.

TABLE 1 | Subjects in the korean national licensing examination for dental hygienists.

Subject	Number of questions (*n*)
Health and medical law[a]	20
Oral anatomy	7
Dental morphology	7
Oral histology and embryology	7
Oral pathology	7
Oral physiology	7
Oral microbiology	5
Community oral health[a]	12
Oral health administration[a]	10
Dental statistics	8
Oral health education	10
Preventive dentistry	18
Oral prophylaxis	20
Dental radiology	20
Oral and maxillofacial surgery	6
Prosthodontics	6
Conservative dentistry	6
Pediatric dentistry	6
Periodontology	6
Orthodontics	6
Dental materials	6
Total	200[b]

[a]Subjects that include the laws and institutional systems specific to korea.
[b]The total number of questions over a single year.

Figure 1 Illustrates the flowchart and the target questions in the study. Subjects like health and medical law, community oral health, and oral health administration, which deal extensively with Korea's legal and administrative systems, were analyzed separately and evaluated in Korean due to their regional specificity. The remaining 18 subjects were directly reproduced from the Korean national dental hygienist exam questions and input in Korean. For the English version, the questions were translated based on a “Standardized Dental

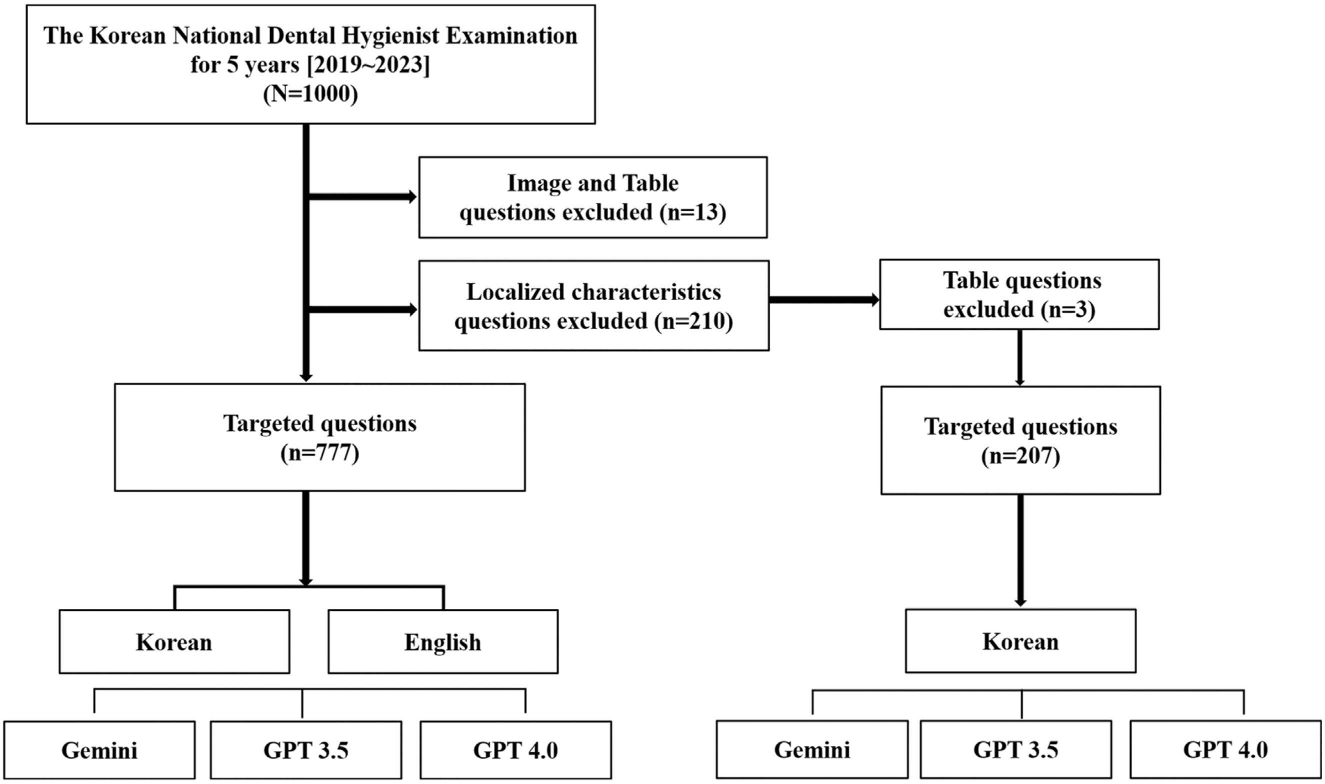

FIGURE 1 | The flow chart and targeted questions used in the study.

Terminology" glossary. Questions that included multimedia and those with tables and graphs not officially provided due to privacy and copyright protection were excluded. The three subjects with regional characteristics comprised a total of 207 questions, and the remaining 18 subjects comprised 777 questions.

2.3 | Prompt Engineering and Data Input Methods

Prompt engineering, which significantly impacts the output results, involved standardizing prompts that were consistently input for each question as follows: "You are a student taking this exam. The exam consists of multiple-choice questions, and you must select only one answer that is most appropriate. You are free to search and utilize necessary information through the internet. Which of the following best represents the most appropriate answer?" Moreover, chat sessions were reset for every question to prevent memory retention and in-context learning from influencing the results. Examples of prompts and responses input in English for each of the LLMs are shown in Figure 2. The same prompts were used for the Korean inputs, as detailed in Figure S1.

Responses were classified as correct or incorrect. A correct response was defined as providing precisely one correct answer. Incorrect responses included incorrect answers, responses with two or more answers, or a failure to provide an answer. Each question was asked only once, and any incorrect, multiple, or indeterminate responses were categorized as incorrect.

2.4 | Statistical Analysis

The obtained data were entered into Microsoft Excel. All statistical calculations were performed using the SPSS software package (SPSS Inc., Chicago, IL, USA). A two-way analysis of variance (ANOVA) was used to examine the impacts of LLMs (Gemini, GPT-3.5, and GPT-4) and language (Korean and English) on accuracy. Tukey's Honestly Significant Difference (HSD) post hoc test was applied to compare the mean accuracy scores between the AI models. The statistical significance was set at $p < 0.05$.

3 | Results

Table 2 Demonstrates the annual accuracy results of each LLM. GPT-4 achieved the highest accuracy rates annually, surpassing the pass threshold each year in both language versions. In particular, it showed high accuracy in the English version, with over 70% accuracy on average. In the Korean version, Gemini and GPT-3.5 did not fulfill the passing criteria, whereas, in the English version, they nearly fulfilled the established passing criteria. Notably, the accuracy in the English version was consistently higher than in the Korean version across all models, with both GPT-3.5 and Gemini showing significant improvements in English. However, in the Korean version, GPT-3.5 had the lowest accuracy among the models tested. Statistical analysis indicated significant effects of model type ($p < 0.001$) and language ($p < 0.001$) on accuracy, with a notable interaction between the two factors ($p < 0.001$). Tukey's HSD post hoc comparisons revealed significant differences in accuracy across all models, with

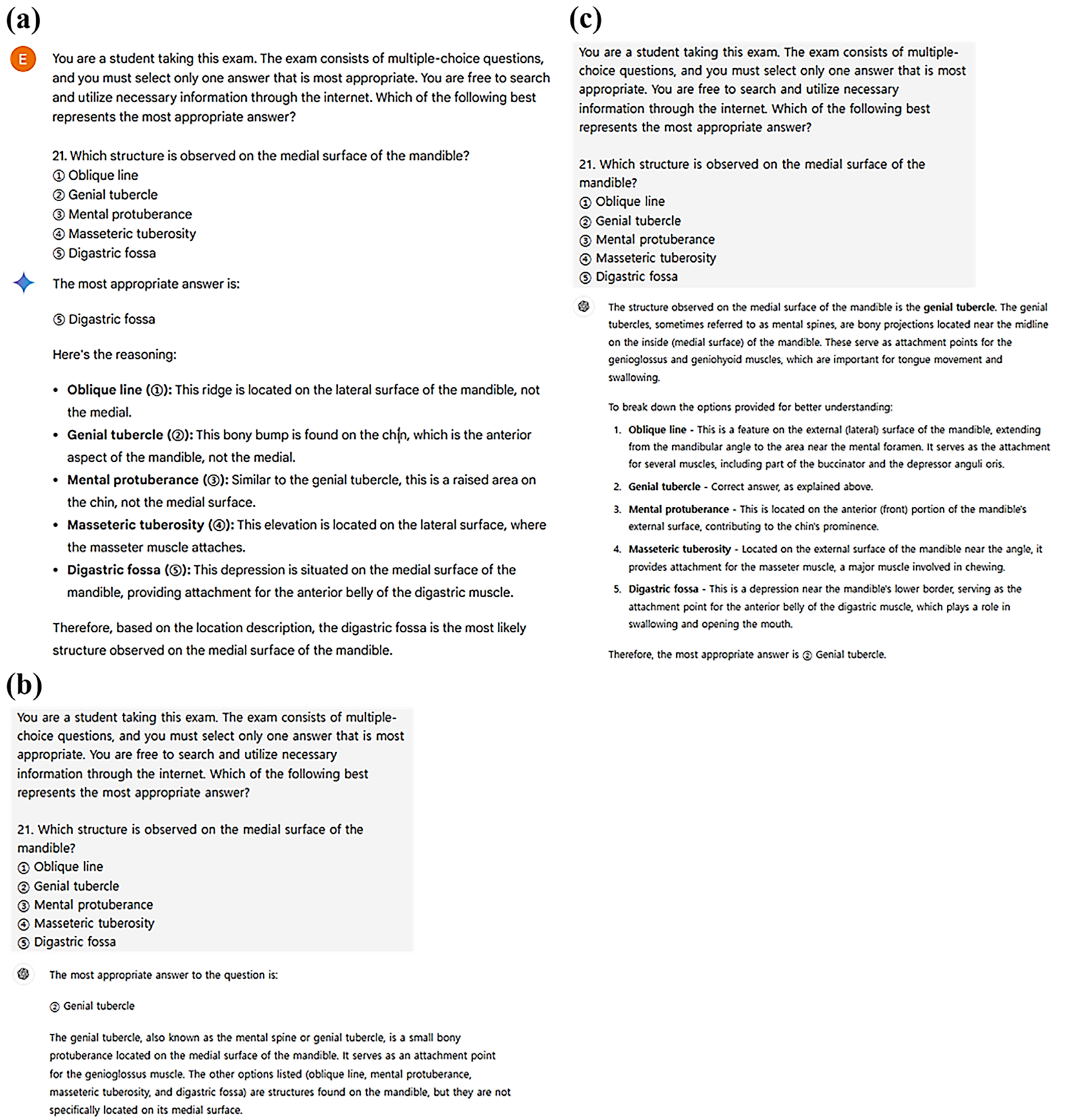

FIGURE 2 | Examples of the prompts and corresponding responses input into the LLMs in English version. (a) Gemini, (b) ChatGPT-3.5, and (c) ChatGPT-4.

GPT-4 consistently demonstrating higher accuracy than Gemini ($p<0.001$) and GPT-3.5 ($p<0.001$). Furthermore, Gemini showed a significant improvement over GPT-3.5 ($p=0.002$).

Table 3 Shows the accuracy rates of LLMs in the Korean version of the Korean National Dental Hygienist Examination compared to the passing rates of Korean dental hygiene students. GPT-4 achieved the highest average accuracy at 65.9%, peaking at 70.3% in 2023, yet it remained below the students' average passing rate of 81.9%. Gemini averaged 51.1%, with its highest accuracy at 58.2% in 2023, while GPT-3.5 had the lowest average at 37.7%, with a significant low of 30.3% in 2020.

Table 4 Shows the subject-specific accuracy results of each LLM. GPT-4 surpassed the passing criterion of 40% in all subjects except dental statistics. Conversely, Gemini and GPT-3.5 failed to meet the passing threshold in several subjects. Overall, the English version showed higher accuracy across all subjects. The discrepancy in performance among the AI models was less pronounced in the English version. The highest

TABLE 2 | Accuracy comparison in the performance of each of LLM.

	Average	2019	2020	2021	2022	2023
Korean version						
Gemini (%)	51.1	49.4	40.1	54.8	53.2	58.2
GPT 3.5 (%)	37.7	45.5	30.3	36.3	37.2	39.2
GPT 4 (%)	65.9	63.6	67.1	61.8	66.7	70.3
English version						
Gemini (%)	61.7	59.7	58.6	58	64.7	67.7
GPT 3.5 (%)	61.3	54.5	67.8	59.2	62.2	62.7
GPT 4.0 (%)	74.2	74.7	69.7	74.5	73.7	78.5

TABLE 3 | Comparative analysis of the LLM accuracy rates and Korean Dental Hygiene Students National Examination passing rates (2019–2023).

	Accuracy rate of Gemini (%)	Accuracy rate of GPT-3.5 (%)	Accuracy rate of GPT-4 (%)	Passing rate of korean dental hygiene students (%)
Average	51.1	37.7	65.9	81.9
2019	49.4	45.5	63.6	84.6
2020	40.1	30.3	67.1	74.1
2021	54.8	36.3	61.8	80.8
2022	53.2	37.2	66.7	82.1
2023	58.2	39.2	70.3	88

accuracy in the Korean version was achieved by GPT-4 in oral microbiology at 92%, followed by oral physiology at 88.6%. In the English version, GPT-4 also led in oral pathology with 94.3% accuracy and in oral histology and embryology, as well as oral physiology, with 91.4% accuracy each. The lowest accuracy was observed in dental statistics, where GPT-3.5 scored 14.7% in the Korean version and Gemini scored 23.5% in the English version.

Table 5 Presents the performance results of the LLMs across three subjects with localized characteristics. Contrary to previous results where GPT-4 consistently exceeded the pass threshold, this time, it fell short, except for the year 2022. In some instances, Gemini and GPT-3.5 demonstrated higher accuracy than GPT-4. Furthermore, in 2023, all models failed to surpass the pass threshold. The subject-specific accuracy results revealed particularly low performances in health and medical law, with Gemini displaying higher accuracy than GPT-4.

Table 6 Illustrates the distribution of incorrect responses for each LLM. Incorrect answers were categorized as indeterminate or multiple answers. Gemini had a higher proportion of multiple answers, whereas GPT-4 showed a significantly higher rate of indeterminate answers than the other LLMs. Table 7 provides a distribution of incorrect responses by subject. Incorrect answers were relatively evenly distributed across various subjects, and GPT-4 exhibited a significantly higher rate of incorrect responses in health and medical law and dental statistics.

4 | Discussion

In this study, we assessed the performance of Gemini, GPT-3.5, and GPT-4 on the Korean National Dental Hygienist Examination questions presented in both Korean and English. Analysis of 5 years of data revealed that GPT-4 consistently outperformed the other models, meeting the passing criteria across all years. In contrast, Gemini and GPT-3.5 fell short in the Korean version but reached or nearly reached passing scores in the English version. The improvement in accuracy for all LLMs when using English inputs, particularly for GPT-3.5, suggests that these models may be more adept at processing English due to more comprehensive training data in that language. Statistical analysis showed significant differences in accuracy among the models ($p < 0.001$), with GPT-4 significantly outperforming Gemini and GPT-3.5 across all subjects, highlighting its more refined language comprehension and processing capabilities.

The interaction between language and model performance was also significant ($p < 0.001$), indicating that LLM performance varies depending on the language used. GPT-4's higher performance in English likely reflects the influence of the more extensive training data available for English processing. Consistent with the findings of Yamaguchi et al. [15] GPT-4's superior performance across both languages suggests advancements in AI training algorithms and data quality. This aligns with Lee et al. [16] results, where GPT-4 demonstrated higher accuracy in emergency medicine, further emphasizing its potential in

TABLE 4 | Comparative analysis of the subject-specific accuracies of the LLMs.

Subject	Gemini (%)	GPT-3.5 (%)	GPT-4.0 (%)
Korean version			
Oral anatomy	42.9	17.1	51.4
Dental morphology	37.5	28.1	53.1
Oral histology and embryology	57.1	45.7	77.1
Oral pathology	65.7	42.9	82.9
Oral physiology	65.7	54.3	88.6
Oral microbiology	80	68	92
Dental statistics	23.5	14.7	29.4
Oral health education	68	58	70
Preventive dentistry	49.4	32.6	55.1
Oral prophylaxis	43.9	37.8	49
Dental radiology	40	34	72
Oral and maxillofacial surgery	70	50	76.7
Prosthodontics	48.3	37.9	72.4
Conservative dentistry	44.8	31	69
Pediatric dentistry	50	30	70
Periodontology	44.8	34.5	75.9
Orthodontics	58.6	37.9	72.4
Dental materials	75.9	41.4	86.2
English version			
Oral anatomy	77.1	80	88.6
Dental morphology	46.9	46.9	59.4
Oral histology and embryology	80	77.1	91.4
Oral pathology	85.7	74.3	94.3
Oral physiology	82.9	85.7	91.4
Oral microbiology	84	80	88
Dental statistics	23.5	32.4	35.3
Oral health education	68	70	74
Preventive dentistry	47.2	47.2	58.4
Oral prophylaxis	42.9	50	62.2
Dental radiology	58	60	81
Oral and maxillofacial surgery	86.7	73.3	80
Prosthodontics	79.3	58.6	82.8
Conservative dentistry	55.2	69	82.8
Pediatric dentistry	53.3	50	73.3
Periodontology	93.1	82.8	82.8
Orthodontics	62.1	48.3	72.4
Dental materials	72.4	69	89.7

TABLE 5 | Accuracy comparison in the performance analysis of LLMs on localized characteristics questions.

Localized characteristics questions					
	2019	**2020**	**2021**	**2022**	**2023**
Gemini (%)	53.7	59.5	62.5	45.2	19
GPT 3.5 (%)	41.5	52.4	52.5	35.7	21.4
GPT 4.0 (%)	58.5	35.7	57.5	64.3	28.6

Subject-specific accuracy evaluation			
	Health and medical law	**Community oral health**	**Oral health administration**
Gemini (%)	43.9	47.5	47.5
GPT 3.5 (%)	32.7	44.1	44.1
GPT 4.0 (%)	39.8	59.3	59.3

complex decision-making scenarios within medical education and diagnostics.

However, the study raises concerns about potential biases in AI models predominantly trained on English data, which could limit their effectiveness in non-English speaking regions. Addressing these biases requires more linguistically diverse datasets and innovations in model training to better accommodate different languages and cultures [17, 18].

The analysis of LLMs on the Korean version of the dental hygiene examination revealed significant inconsistencies in performance compared to the more stable and higher passing rates of Korean dental hygiene students (Table 3). While GPT-4 maintained the highest average performance among the LLMs at 65.9%, its accuracy fluctuated annually, ranging from 63.6% in 2019 to 70.3% in 2023. Gemini and GPT-3.5 also exhibited significant variability, with Gemini's performance oscillating between 49.4% and 58.2%, and GPT-3.5 ranging from 45.5% to 39.2%. These variations starkly contrast with the consistent passing rates of human examinees, which remained above 74%. This gap highlights the challenges in developing AI systems that can match human understanding and reasoning in specialized fields. The inconsistencies of the models in this study underscore the difficulty of equipping LLMs with the depth of professional knowledge required in the field of dental hygiene.

Table 4 Illustrates the considerable variations in the performances of different LLMs. While GPT-4 met the passing criteria across all subjects except dental statistics, Gemini and GPT-3.5 did not reach the threshold in certain areas. The highest accuracies were seen with GPT-4 in oral microbiology, oral physiology, and dental materials, suggesting these subjects may be less affected by linguistic or regional conditions. Conversely, all models exhibited lower accuracy in dental statistics, which might be due to the requirement for higher-order abstract reasoning and interpretation of specialized indices within dentistry. The generally low performance in dental morphology could be due to the complex nature of dental anatomical terms and specialized terminology that may not translate precisely from Korean to English. The recent utilization of LLMs as supplementary learning tools for educational purposes is experiencing an upward trend [19–22]. Although LLMs offer the potential to function like private tutors, enhancing the student's ability to pass examinations, caution must be exercised because passing an exam does not necessarily mean that the correct answers are being provided. The effectiveness of LLMs varies significantly across different subjects within the dental curriculum, necessitating a subtle approach to their application. For instance, as shown in Table 4, the LLMs performed well in areas relying on memorization, but their utility was limited in subjects that required critical thinking and data interpretation. These findings indicate that although LLMs can enhance traditional educational methods by providing immediate, accurate information in certain areas, they have inherent limitations. Considering the potential for students to use LLMs as private teachers in the future, they must be integrated into dental education alongside traditional methods, thus playing a complementary role. Developing guidelines for the effective use of LLMs will be essential to this process.

The study also highlights the potential of LLMs like GPT-4 as educational support tools in standardized testing [23, 24]. However, it emphasizes the need for continuous improvement, particularly for nuanced details and questions in languages other than English. Furthermore, the cost of accessing advanced LLMs like GPT-4 could be a barrier for students, especially in regions with limited financial resources, potentially exacerbating educational disparities [25]. The future development and implementation of LLMs in education must consider these economic issues and investigate ways to make these sophisticated tools more accessible to a wider variety of users. Furthermore, for educators, LLMs can enhance educational activities by helping create instructional materials, such as quizzes and interactive content, and by assisting in assessing student performance to identify areas needing further attention [26]. This enables more targeted and effective teaching strategies. However, educators must address potential biases to ensure the accuracy and reliability of the data and remain mindful of privacy considerations [27, 28].

TABLE 6 | Distribution of incorrect responses for each LLM.

	Incorrect answers (*n*)	Indeterminate answers (*n*)	Percentage of indeterminate answers (%)	Multiple answers (*n*)	Percentage of multiple answers (%)
Gemini	800	2	0.3	20	2.5
GPT-3.5	892	3	0.3	1	0.1
GPT-4	571	81	14.2	4	0.7

Abbreviation: *n*, number.

TABLE 7 | Distribution of incorrect responses based on the subject category.

Subject	Gemini (*n*)	GPT-3.5 (*n*)	GPT-4 (*n*)
Health and medical law	2		35
Oral anatomy	1		
Oral histology and embryology	2		
Oral physiology	1		1
Oral microbiology			1
Community oral health		2	3
Oral health administration			1
Dental statistics	1	1	20
Oral health education			1
Preventive dentistry	4		5
Oral prophylaxis	6		6
Dental radiology	2	1	3
Oral and maxillofacial surgery			1
Prosthodontics	1		1
Conservative dentistry			2
Pediatric dentistry			2
Periodontology	2		1
Orthodontics			2
Total	22	4	85

Abbreviation: *n*, number.

Analysis of the performances on localized characteristic questions (Table 5) provided a nuanced view of AI capabilities, showing variability across all models over the years, with GPT-4 generally exhibiting higher accuracy, except in 2020 and 2021. This raises questions about the adaptability of models to changes in exam content or format. LLMs struggled with health and medical law, which requires a deep understanding of national regulations, though GPT-4 showed a slight advantage. In subjects like community oral health and oral health administration, GPT-4 performed better, although inconsistently. These findings emphasize the need for careful application of such models, as current training may not fully encompass the depth needed for understanding localized topics.

Tables 6 and 7 Show the distribution of incorrect responses, highlighting the challenges AI models face in mastering dental professional knowledge. A significant proportion of GPT-4's incorrect answers were concentrated in health and medical law and dental statistics, while Gemini exhibited a broader distribution of errors. GPT-4 also had a higher rate of indeterminate responses (14.2%) compared to Gemini and GPT-3.5 (0.3%), suggesting uncertainty in areas requiring nuanced understanding.

To assess the impact of indeterminate responses on GPT-4's performance, we conducted a sensitivity analysis by excluding these responses from the dataset. The recalculated accuracy for GPT-4 increased from 42.0% to 45.7%, indicating a 3.7% improvement. This suggests that indeterminate responses contributed significantly to the lower accuracy scores observed in the initial analysis. Despite this increase, GPT-4's performance ranking relative to other models remains consistent, although the overall accuracy is enhanced.

There are several limitations to consider in this study. First, the performance evaluation on questions with localized characteristics was limited to Korean input, restricting comparative language analysis. Second, some questions were excluded from the analysis due to copyright issues, which may have impacted the statistical outcomes. Third, the frequency of model updates and improvements could influence the results, as the current performance and findings are based on the LLMs' responses as of April 2024, which may differ in the future. Additionally, the study did not fully explore the potential impact of cultural and linguistic nuances in other languages, which could further affect the performance of these models.

This study marks a significant step in evaluating LLM performance on Korean National Dental Hygienist Examination questions. The superior performance of GPT-4, regardless of language, suggests its potential applications in dental hygiene and education. However, the results also highlight the need for model training using datasets that reflect national languages and cultural contexts. Future research should focus on developing methodologies for evaluating LLMs in multilingual settings and improving their proficiency in handling specialized content across different languages.

In conclusion, while LLMs like GPT-4 show promise in educational support and standardized testing, their application must be carefully managed, particularly in non-English dominant contexts. Addressing the limitations and biases identified in this study is essential for enhancing the models' effectiveness in global healthcare and education.

5 | Conclusion

This study evaluated the performance of three large language models—Gemini, GPT-3.5, and GPT-4—on the Korean National Dental Hygienist Examination in both Korean and English. The findings indicate that GPT-4 holds significant potential for application in medical education, particularly in standardized testing, due to its superior accuracy, especially in English. However, the variability in performance across different subjects and languages highlights the need for ongoing improvements in these models. These improvements should focus on integrating more diverse and localized training datasets to enhance their effectiveness in multilingual and multicultural contexts.

Scientific Rationale for the Study

Assess the effectiveness of large language models (LLMs) like Gemini, GPT-3.5, and GPT-4 in dental hygiene education by evaluating their performance on the Korean National Dental Hygienist Examination.

Principal Findings

GPT-4 consistently outperformed other models, notably in English, though all models exhibited variability in specialized subjects such as health and medical law.

Practical Implications

The study underscores the potential of LLMs to enhance dental hygiene training but emphasizes the need for targeted improvements. Enhancing their ability to handle specialized content and multilingual environments is crucial for integrating LLMs effectively into diverse educational frameworks.

Author Contributions

E.S.S. contributed to the conception and design, data acquisition, analysis, and interpretation, wrote the first draft of the manuscript, and critically revised the manuscript. S.-P.L. contributed to the conception and design, data acquisition, analysis, and interpretation, and drafted and critically revised the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Acknowledgements

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

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

References

1. A. Haleem, M. Javaid, and I. H. Khan, "Current Status and Applications of Artificial Intelligence (AI) in Medical Field: An Overview," *Current Medicine Research and Practice* 9, no. 6 (2019): 231–237.

2. G. Briganti and O. Le Moine, "Artificial Intelligence in Medicine: Today and Tomorrow," *Frontiers in Medicine* 7 (2020): 509744.

3. H. Lee, "The Rise of ChatGPT: Exploring Its Potential in Medical Education," *Anatomical Sciences Education* 17 (2023): 926–931.

4. M. Mirbabaie, S. Stieglitz, and N. R. Frick, "Artificial Intelligence in Disease Diagnostics: A Critical Review and Classification on the Current State of Research Guiding Future Direction," *Health and Technology* 11, no. 4 (2021): 693–731.

5. J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and P. Q. Consortium, "Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective," *BMC Medical Informatics and Decision Making* 20 (2020): 1–9.

6. A. Bohr and K. Memarzadeh, "The Rise of Artificial Intelligence in Healthcare Applications," in *Artificial Intelligence in Healthcare*, eds. A. Bohr and K. Memarzadeh (Amsterdam, The Netherlands: Elsevier, 2020), 25–60.

7. E. Kasneci, K. Seßler, S. Küchemann, et al., "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education," *Learning and Individual Differences* 103 (2023): 102274.

8. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large Language Models in Medicine," *Nature Medicine* 29, no. 8 (2023): 1930–1940.

9. A. Gilson, C. W. Safranek, T. Huang, et al., "How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment," *JMIR Medical Education* 9, no. 1 (2023): e45312.

10. T. H. Kung, M. Cheatham, A. Medenilla, et al., "Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models," *PLOS Digital Health* 2, no. 2 (2023): e0000198.

11. X. Wang, Z. Gong, G. Wang, et al., "ChatGPT Performs on the Chinese National Medical Licensing Examination," *Journal of Medical Systems* 47, no. 1 (2023): 86.

12. Y. Yanagita, D. Yokokawa, S. Uchida, J. Tawara, and M. Ikusaka, "Accuracy of ChatGPT on Medical Questions in the National Medical Licensing Examination in Japan: Evaluation Study," *JMIR Formative Research* 7 (2023): e48023.

13. N. P. Davies, R. Wilson, M. S. Winder, et al., "ChatGPT Sits the DFPH Exam: Large Language Model Performance and Potential to Support Public Health Learning," *BMC Medical Education* 24, no. 1 (2024): 57.

14. Y.-M. Wang, H.-W. Shen, and T.-J. Chen, "Performance of ChatGPT on the Pharmacist Licensing Examination in Taiwan," *Journal of the Chinese Medical Association* 86, no. 7 (2023): 653–658.

15. S. Yamaguchi, M. Morishita, H. Fukuda, et al., "Evaluating the Efficacy of Leading Large Language Models in the Japanese National Dental Hygienist Examination: A Comparative Analysis of ChatGPT, Bard, and Bing Chat," *Journal of Dental Sciences* 19 (2024): 2262–2267.

16. G. U. Lee, D. Y. Hong, S. Y. Kim, et al., "Comparison of the Problem-Solving Performance of ChatGPT-3.5, ChatGPT-4, Bing Chat, and Bard

for the Korean Emergency Medicine Board Examination Question Bank," *Medicine* 103, no. 9 (2024): e37325.

17. M. Tawkat Islam Khondaker, A. Waheed, E. Moatez Billah Nagoudi, and M. Abdul-Mageed, "GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP," *arXiv* 2023, https://doi.org/10.48550/arXiv.2305.14976.

18. V. D. Lai, N. T. Ngo, A. P. B. Veyseh, et al., "Chatgpt Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning," *arXiv* 2023, https://doi.org/10.48550/arXiv.2304.05613.

19. M.-L. Tsai, C. W. Ong, and C.-L. Chen, "Exploring the Use of Large Language Models (LLMs) in Chemical Engineering Education: Building Core Course Problem Models With Chat-GPT," *Education for Chemical Engineers* 44 (2023): 71–95.

20. T. Phung, V.-A. Pădurean, J. Cambronero, et al., "Generative AI for Programming Education: Benchmarking Chatgpt, gpt-4, and Human Tutors," *arXiv* 2023, https://doi.org/10.48550/arXiv.2306.17156.

21. C. Meaney, R. S. Huang, K. Lu, et al., "Comparing the Performance of ChatGPT and GPT-4 Versus a Cohort of Medical Students on an Official University of Toronto Undergraduate Medical Education Progress Test," *medRxiv* 2023, https://doi.org/10.1101/2023.09.14.23295571.

22. L. Morjaria, L. Burns, K. Bracken, et al., "Examining the Efficacy of ChatGPT in Marking Short-Answer Assessments in an Undergraduate Medical Program," *Journal of Medical Education and Curricular Development* 3, no. 1 (2024): 32–43.

23. L. Yan, L. Sha, L. Zhao, et al., "Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review," *British Journal of Educational Technology* 55, no. 1 (2024): 90–112.

24. A. Abd-Alrazaq, R. AlSaad, D. Alhuwail, et al., "Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions," *JMIR Medical Education* 9, no. 1 (2023): e48291.

25. A. Gangavarapu, *LLMs: A Promising New Tool for Improving Healthcare in Low-Resource Nations* (Radnor, PA: *IEEE*, 2023), 252–255.

26. B. Eager and R. Brunton, "Prompting Higher Education Towards AI-Augmented Teaching and Learning Practice," *Journal of University Teaching and Learning Practice* 20, no. 5 (2023), https://doi.org/10.53761/1.20.5.02.

27. X. Zhang, S. Li, B. Hauer, N. Shi, and G. Kondrak, "Don't Trust ChatGPT When Your Question Is not in English: A Study of Multilingual Abilities and Types of LLMs," *arXiv* 2023, https://doi.org/10.48550/arXiv.2305.16339.

28. G. M. N. Isabwe and F. Reichert, *Revisiting Students' Privacy in Computer Supported Learning Systems* (Radnor, PA: IEEE, 2013), 256–262.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.