

RESEARCH

Open Access



Cognitive-level analysis of dentomaxillofacial radiology questions in the Turkish dentistry specialization examination: a Bloom's revised taxonomy analysis

Melisa Ocbe¹, Hulya Cerci Akcay², Birsay Gumru^{3*} and Sebnem Ercalik Yalcinkaya³

Abstract

Background Bloom's revised taxonomy is widely used to evaluate the cognitive level of examination questions. With the growing use of large language models (LLMs), interest has increased in their potential to support cognitive classification. However, evidence comparing LLM-based classification with expert consensus in dental specialty examinations remains limited. This study aimed to analyze the cognitive levels of dentomaxillofacial radiology (DMFR) questions in the Turkish Dentistry Specialization Examination (DUS) and to evaluate the agreement between expert consensus and ChatGPT v5.2.

Methods A total of 130 text-based DMFR questions from the DUS were classified according to Bloom's revised taxonomy using expert consensus and ChatGPT v5.2. Agreement between classifications was assessed using exact agreement and quadratic-weighted Cohen's kappa.

Results "Analyze" was the most frequent cognitive level in both expert consensus (41.5%) and ChatGPT classification (44.6%), followed by "Remember." Exact agreement between ChatGPT and expert consensus was 80.0%, with almost perfect ordinal agreement (kw = 0.851; 95% bootstrap CI: 0.769–0.923).

Conclusions ChatGPT v5.2 showed high agreement with expert consensus for Bloom-level classification in this dataset. These findings suggest that LLM-based classification may support Bloom-level mapping as an assistive tool; however, they should be interpreted as preliminary evidence of feasibility rather than practical validation. Expert oversight remains necessary, and further research is needed to assess performance across different models, settings, and educational contexts.

Keywords Dentomaxillofacial radiology, Dentistry specialization examination, Bloom's revised taxonomy, Cognitive demand, Multiple-choice questions, Curriculum alignment, Large language models

*Correspondence:

Birsay Gumru
bgumru@marmara.edu.tr

¹Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Kocaeli Health and Technology University, Kocaeli, Türkiye

²Department of Pedodontics, Faculty of Dentistry, Kocaeli Health and Technology University, Kocaeli, Türkiye

³Department of Oral and Maxillofacial Radiology, Faculty of Dentistry, Marmara University, Basibuyuk Mah. Basibuyuk Yolu Sok. No: 9/3, Basibuyuk-Maltepe, Istanbul 34854, Türkiye



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Background

The Dentistry Specialization Examination (DUS) is a standardized, national, high-stakes examination used to determine postgraduate training positions in dental specialty areas in Türkiye. The DUS consists of 40 multiple-choice questions (MCQs) in the Basic Sciences Test and 80 in the Clinical Sciences Test, with dentomaxillofacial radiology (DMFR) represented by 10 questions in the clinical test of each examination [1]. High-stakes examinations have a significant influence on students' priorities and study strategies in health professions education, which has been referred to as "assessment drives learning" [2, 3]. Similarly, in Türkiye, since the DUS strongly shapes how students prepare for the examination and what they prioritize during their training, systematic analyses of DUS questions have been suggested as a practical means to support more targeted study and improve the alignment between curriculum and assessment [4].

DMFR is a fundamental branch of modern dentistry that covers radiation physics and radiobiology, imaging principles, conventional and advanced imaging methods, and the interpretation of maxillofacial pathologies. Selecting appropriate imaging modalities, identifying radiographic patterns, integrating multiple findings, and proposing radiologic differential diagnoses are essential competencies in DMFR practice. Therefore, evaluating the cognitive demands of DMFR questions is important to ensure that this high-stakes examination accurately measures these skills.

Bloom's revised taxonomy is a widely used framework for categorizing the cognitive processes required by educational objectives and assessment questions. The six hierarchical levels of the taxonomy - Remember, Understand, Apply, Analyze, Evaluate, and Create - allow differentiation between lower-order cognitive skills (LOCS) and higher-order cognitive skills (HOCS), particularly those related to clinical reasoning and decision-making [5, 6]. Mapping examination questions to Bloom's levels is frequently used in health professions education to assess cognitive complexity; however, reported associations with learning outcomes, examination performance, and clinical reasoning remain inconsistent and context-dependent [7].

In addition to cognitive level, question quality should also be considered in terms of alignment with intended curricular outcomes and scientific accuracy of the question stem and answer key. In Türkiye, competency-based expectations for undergraduate dental education are defined in the National Core Curriculum for Undergraduate Dental Education (DUÇEP) [8]. From an assessment-design perspective, alignment between examination content and the curriculum is essential for content validity and educational coherence, while inaccuracy or ambiguity may reduce educational quality [9].

However, evidence specific to DMFR remains limited regarding the cognitive demand of DUS questions and their alignment with nationally defined undergraduate competencies. Since the DUS plays a crucial role in determining postgraduate training pathways, further domain-specific evaluation is needed. Therefore, the aims of the present study were to: (i) classify DMFR MCQs administered in the DUS between 2012 and 2021 according to Bloom's revised taxonomy; (ii) evaluate curriculum relevance with reference to DUÇEP and assess scientific accuracy for each question; (iii) examine the association between cognitive demand and curriculum relevance; and (iv) compare expert classifications with a large language model (LLM)-based automated classification approach.

Methods

Study design

This study was conducted as a retrospective, descriptive, cross-sectional educational analysis to evaluate the cognitive demands of DMFR questions included in the Turkish DUS. Bloom's revised taxonomy was used as the primary analytical framework to evaluate the cognitive structure of examination items. The study focused on the content and cognitive characteristics of the questions and did not include candidate performance data, examination scores, or psychometric evaluation. The methodological workflow consisted of data identification, eligibility screening, cognitive classification, expert evaluation, LLM-based classification, and statistical analysis.

Data source and study material

The study material consisted of officially released DUS question booklets obtained from the official website of the Measurement, Selection and Placement Center (ÖSYM) [10], where DMFR question booklets are publicly available. DMFR questions administered between 2012 and 2021 were systematically identified and reviewed. The question booklets were retrieved on 15 January 2026. All MCQs categorized under DMFR were screened. No officially cancelled questions were identified. The final sample included all valid DMFR questions published within the specified time frame ($n = 130$).

Eligibility criteria

The inclusion criteria comprised DMFR questions published in the DUS, specifically MCQs with a single best answer. Eligible items included questions assessing radiological principles, image interpretation, recognition of maxillofacial pathologies, selection of appropriate imaging modalities, and radiobiological concepts. Questions were excluded if they had been officially cancelled or duplicated, belonged to dental specialties other than

DMFR, or lacked sufficient information to determine the dominant cognitive process required for classification.

Cognitive classification using Bloom's revised taxonomy

All included questions were classified according to Bloom's revised taxonomy, which organizes cognitive processes into six hierarchical levels: Remember, Understand, Apply, Analyze, Evaluate, and Create. Each question was assigned to a single dominant cognitive level, defined as the highest level of cognitive processing required to answer the question correctly. Classification was based on the complexity of the clinical or radiographic scenario, the degree of reasoning required, the need to integrate multiple radiologic findings, and the requirement for differential diagnosis or imaging modality selection. Questions requiring recognition of isolated factual knowledge were classified as LOCS (Remember, Understand), whereas items involving interpretation of radiographic patterns, synthesis of findings, or differential diagnosis were classified as HOCS (Apply, Analyze, Evaluate, Create). This classification was adopted in line with commonly used applications of Bloom's taxonomy in health professions education [5–7].

Rater training and evaluation process

Cognitive-level classification was performed independently by three DMFR specialists with experience in undergraduate and postgraduate radiology education. Before evaluation, all raters reviewed the theoretical framework and operational definitions of Bloom's revised taxonomy to ensure conceptual consistency. A pilot calibration was performed on a subset of questions not included in the final analysis to harmonize classification criteria. Each rater evaluated all questions independently and was blinded to the other raters' classifications.

Pre-consensus inter-rater agreement among the three experts was assessed using Fleiss' kappa. In addition, descriptive agreement patterns (full agreement, partial disagreement, and complete disagreement) were calculated to provide further insight into rating consistency. Consensus discussions were then conducted among the three raters to resolve disagreements.

Consensus rules were defined a priori as follows: if at least two experts agreed, the majority level was accepted. In cases of complete three-way disagreement, the median ordinal Bloom level was assigned. Expert consensus was used as the reference classification for comparison with ChatGPT classifications.

ChatGPT-based classification

To enable comparison with expert classification, an independent cognitive-level assessment was performed using ChatGPT v5.2. ChatGPT was provided with the six levels of Bloom's revised taxonomy and instructed to assign the

highest cognitive level required to answer each question correctly. The classifications generated by ChatGPT were recorded for all questions and compared with the expert consensus classifications.

To support judgments about curriculum relevance, the DUÇEP framework was introduced to the model through a standardized instructional prompt summarizing the scope of undergraduate DMFR competencies, including radiation physics and radiobiology, imaging modalities, interpretation of radiographic findings, and diagnostic reasoning. In addition to Bloom-level classification, ChatGPT was instructed to rate curriculum relevance and scientific accuracy for each item using a 5-point Likert scale (1 = very poor, 5 = excellent). All questions were presented to the model in Turkish, and each item initially received a single recorded output. To maintain consistency across evaluations, all questions were assessed independently using the same standardized prompt structure, as follows:

"You are an expert in DMFR education and dental curriculum design. I will provide you with MCQs from the Turkish DUS. For each question, please perform the following tasks:

1. Determine the single highest cognitive level required to correctly answer the question using Bloom's revised taxonomy: Remember, Understand, Apply, Analyze, Evaluate, and Create.
2. Evaluate the relevance of the question to the undergraduate DMFR curriculum defined in DUÇEP, which was provided as an attachment to this prompt. Consider whether the item addresses competencies such as radiation physics and radiobiology, conventional radiographic techniques, panoramic imaging, advanced imaging modalities (cone-beam computed tomography [CBCT], computed tomography [CT], magnetic resonance imaging [MRI], ultrasonography [US]), interpretation of radiolucent and radiopaque lesions, anatomical variations, systemic diseases affecting the maxillofacial region, soft tissue calcifications, and radiologic diagnostic reasoning.
3. Rate curriculum relevance on a 5-point Likert scale (1 = very poor, 5 = excellent).
4. Evaluate the scientific accuracy of the question stem and answer key based on current radiological knowledge and accepted clinical guidance, and rate it on the same 5-point Likert scale.
5. Return your answer using this template: Bloom level; Curriculum relevance (1–5); Scientific accuracy (1–5)."

To evaluate the reproducibility of LLM-based classification, a randomly selected subset of the dataset ($n = 38$

questions) was re-evaluated using the same standardized prompt in two additional independent runs. Thus, each item in this subset was classified three times under identical input conditions. Agreement across runs was assessed using exact agreement, and consistency patterns were examined to evaluate the stability of model outputs.

Assessment of curriculum relevance and scientific accuracy

In addition to cognitive-level classification, each question was independently evaluated for curriculum relevance and scientific accuracy. Both domains were rated using a 5-point Likert scale (1 = very poor, 5 = excellent).

Curriculum relevance was defined as the degree to which the question aligned with the national undergraduate DMFR curriculum, including radiation physics, radiobiology, conventional radiographic techniques, panoramic imaging, advanced imaging modalities (CBCT, CT, MRI, US), interpretation of radiolucent and radiopaque lesions, systemic disease manifestations in maxillofacial radiology, soft tissue calcifications, and anatomical variations. Curriculum relevance was evaluated using DUÇEP published by the Turkish Council of Higher Education (YÖK) [8]. Each question was reviewed and mapped to the relevant learning outcomes and competency domains defined in DUÇEP.

Scientific accuracy was defined as the correctness and clarity of the question stem and answer key and its consistency with current evidence-based radiologic knowledge and accepted clinical guidelines.

These expert ratings were performed independently by the three experts and finalized by consensus using the same procedure described above.

Statistical analysis

All statistical analyses were performed using Python (Pandas, SciPy, and StatsModels libraries). Statistical significance was set at $p < 0.05$. Bloom level distributions were summarized using frequencies and percentages. Exact agreement between ChatGPT and expert consensus classifications was calculated as the proportion of identical ratings. Ordinal agreement across Bloom categories was assessed using quadratic-weighted Cohen's kappa (κ_w), with 95% confidence intervals (CIs) obtained through 5000 bootstrap resamples. Overall asymmetry in the 5×5 contingency table was evaluated using Bowker's

test of symmetry; cell pairs with structural zeros (0–0 pairs) were excluded from the effective degrees of freedom calculation. After dichotomization into LOCS and HOCS, paired comparisons were conducted using McNemar's exact test. Likert-scale ratings were summarized as mean $\pm$ standard deviation and median (inter-quartile range). The association between Bloom level and curriculum relevance was examined using Spearman's rank correlation coefficient.

No prior power analysis was performed, as the study included all publicly available DMFR questions within the specified period. However, a post hoc sensitivity analysis indicated that the study could detect approximately small-to-moderate effects (e.g., $|\rho| \approx 0.24$ at $\alpha = 0.05$ and 80% power).

Ethical considerations

Ethical approval was not required for this study because the analysis was limited to publicly available examination materials. No human participants, patient data, or identifiable information were included. The study was conducted in accordance with the principles of ethical educational research.

Results

A total of 130 DMFR MCQs from the Turkish DUS between 2012 and 2021 were included in the final dataset. Classifications according to Bloom's revised taxonomy, curriculum relevance, and scientific accuracy ratings were available for expert consensus and ChatGPT v5.2.

Distribution of Bloom's revised taxonomy levels

Of the five Bloom categories observed (Remember, Understand, Apply, Analyze, and Evaluate), "Analyze" was the most frequent level for both expert consensus (54/130, 41.5%) and ChatGPT v5.2 (58/130, 44.6%) (Table 1). The second most frequent level was "Remember" for both expert consensus (43/130, 33.1%) and ChatGPT v5.2 (33/130, 25.4%). There were no items at the "Create" level in either classification (Fig. 1).

Pre-consensus inter-rater agreement among experts

Before consensus, inter-rater agreement among the three expert raters was assessed. Fleiss' kappa demonstrated substantial inter-rater agreement ($\kappa = 0.765$) across the Bloom categories.

Complete agreement among all three raters was observed in 97 of 130 items (74.6%), while partial disagreement, defined as agreement between two raters, occurred in 33 items (25.4%). There were no instances of complete three-way disagreement. All six Bloom levels were available for rating; however, no item was ultimately classified as "Create" in the final dataset.

Table 1 Distribution of Bloom's revised taxonomy levels (n = 130)

Bloom level	Expert consensus n (%)	ChatGPT v5.2 n (%)
Remember	43 (33.1)	33 (25.4)
Understand	11 (8.5)	21 (16.2)
Apply	9 (6.9)	11 (8.5)
Analyze	54 (41.5)	58 (44.6)
Evaluate	13 (10.0)	7 (5.4)

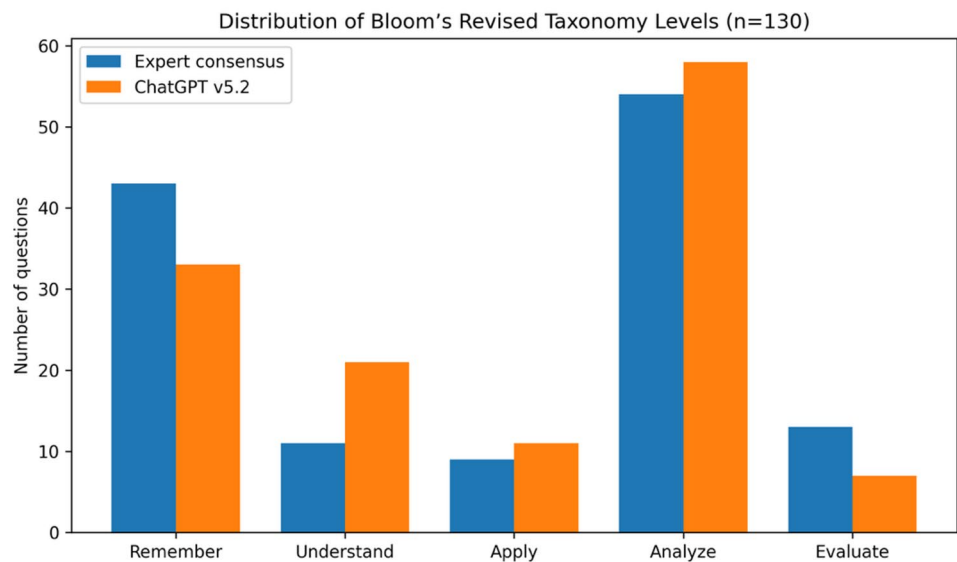


Fig. 1 Distribution of Bloom's revised taxonomy levels for expert consensus and ChatGPT v5.2 across 130 DMFR DUS items

Table 2 Agreement statistics between ChatGPT v5.2 and expert consensus

Metric	Value
Exact agreement, n/N (%)	104/130 (80.0%)
Quadratic-weighted Cohen's κ (ordinal Bloom)	0.851
κ 95% bootstrap CI	0.769–0.923
Bowker's test of symmetry (χ^2 , df, p)	19.2, 10, 0.038

Agreement between ChatGPT v5.2 and expert consensus (Bloom levels)

There was 80.0% (104/130) exact agreement across the five observed Bloom categories. Quadratic-weighted Cohen's kappa indicated almost perfect ordinal agreement ($\kappa_w=0.851$; 95% bootstrap CI: 0.769–0.923) (Table 2).

Despite the high level of agreement, Bowker's test revealed statistically significant asymmetry in the overall multicategory disagreement pattern ($\chi^2 = 19.2$, df= 10, $p=0.038$) (Table 2).

As shown in Fig. 2, asymmetry primarily resulted from directional mismatches between adjacent levels. The most frequent imbalances were transitions from "Remember" to "Understand" ($n=8$; with no reverse cases) and from "Evaluate" to "Analyze" ($n=6$; with no reverse cases). Inconsistencies were generally adjacent-level shifts, with only a few larger transitions (e.g., "Remember" to "Analyze", $n=3$; "Analyze" to "Remember", $n=2$).

Reproducibility of ChatGPT classifications

To assess the stability of LLM-based classification, a subset of 38 questions was re-analyzed across three independent ChatGPT runs using the same standardized prompt. Exact agreement across all three runs was observed in 36

of 38 items (94.2%), indicating a high level of reproducibility. The remaining two items (5.8%) showed minor variability, typically between adjacent Bloom levels. No substantial classification shifts across distant cognitive levels were observed.

Lower- vs. higher-order cognitive skills (LOCS/HOCS)

After dichotomizing Bloom's levels into LOCS (Remember + Understand) and HOCS (Apply + Analyze + Evaluate + Create), paired agreement was 92.3% (120/130). Discordant classifications were symmetric: five items were classified as LOCS by expert consensus but as HOCS by ChatGPT, and five items were classified as HOCS by expert consensus but as LOCS by ChatGPT. McNemar's exact test was not significant ($p=1.000$) (Fig. 3).

Curriculum relevance and scientific accuracy

Expert-consensus curriculum relevance was high (mean $\pm$ SD, 4.91 ± 0.40), with most items rated 5/5 (Table 3). ChatGPT v5.2 rated curriculum relevance as 5/5 for all items. Expert-consensus scientific accuracy exhibited a ceiling effect, with all items rated 5/5. ChatGPT v5.2 scientific accuracy ratings were slightly lower (mean $\pm$ SD, 4.57 ± 0.63). The association between expert-consensus Bloom level and expert-consensus curriculum relevance was negligible (Spearman's $\rho=0.000$, $p=0.997$) (Figs. 4 and 5).

Discussion

DMFR is a specialized branch of clinical dentistry that requires a comprehensive theoretical foundation and the practical application of radiological knowledge to case-based scenarios. In Türkiye, as with other dental

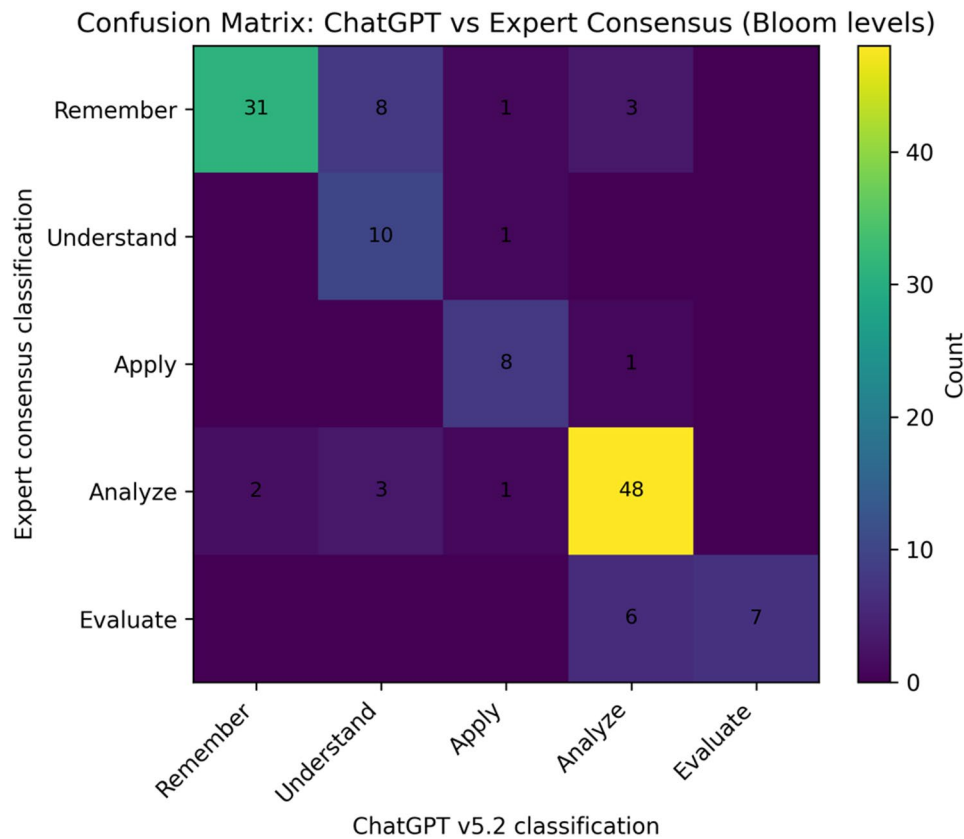


Fig. 2 Confusion matrix showing the cross-classification of Bloom levels assigned by ChatGPT v5.2 versus expert consensus

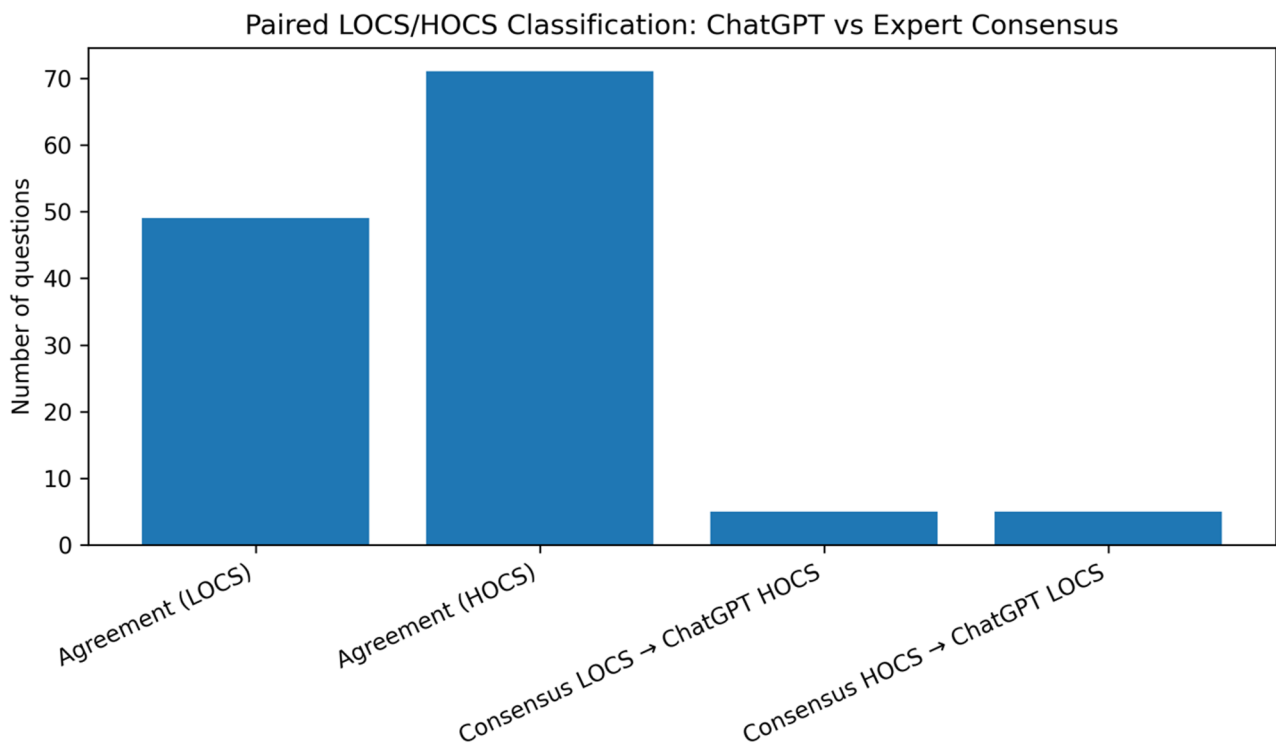
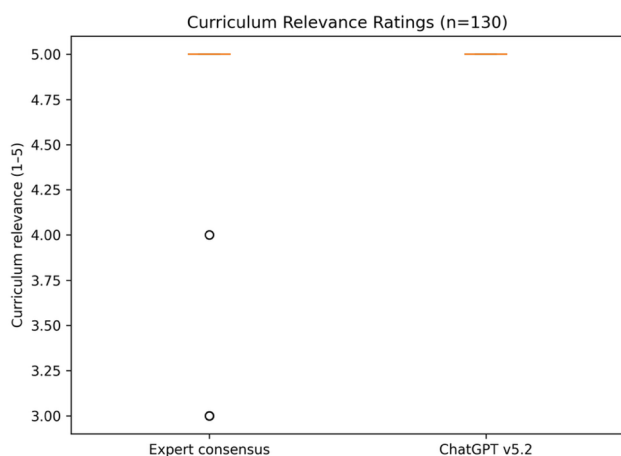


Fig. 3 Paired comparison of lower-order versus higher-order cognitive skills between ChatGPT v5.2 and expert consensus

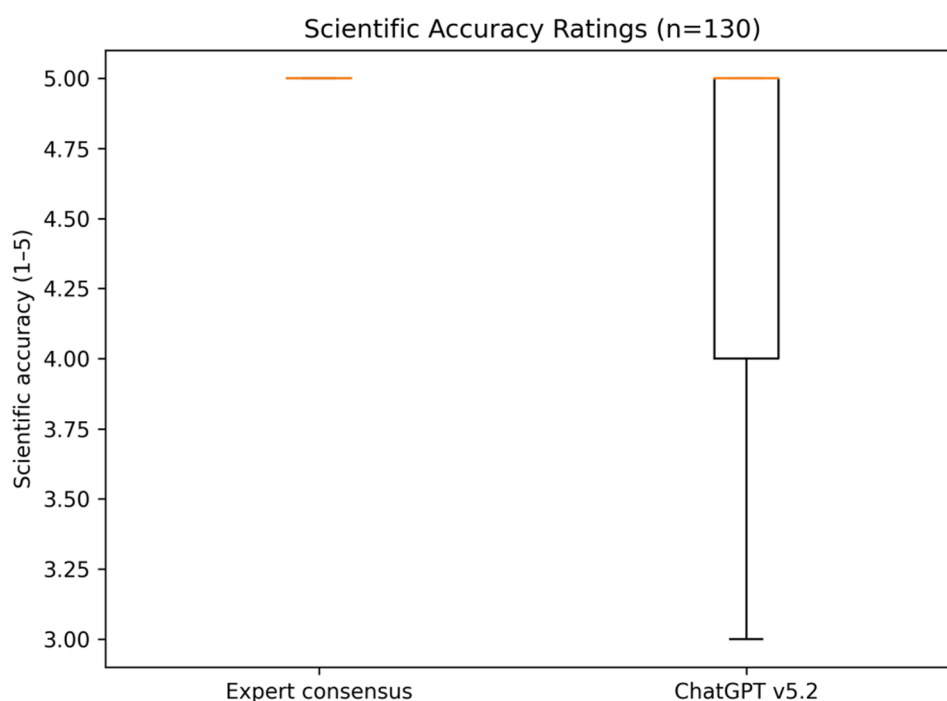
Table 3 Likert-scale ratings for curriculum relevance and scientific accuracy ($n = 130$)

Domain	Rater	Mean $\pm$ SD	Median (IQR); Min–Max
Curriculum relevance (1–5)	Expert consensus	4.91 $\pm$ 0.40	5 (5–5); 3–5
Curriculum relevance (1–5)	ChatGPT v5.2	5.00 $\pm$ 0.00	5 (5–5); 5–5
Scientific accuracy (1–5)	Expert consensus	5.00 $\pm$ 0.00	5 (5–5); 5–5
Scientific accuracy (1–5)	ChatGPT v5.2	4.57 $\pm$ 0.63	5 (4–5); 3–5

**Fig. 4** Boxplot of curriculum relevance ratings (Likert 1–5) for expert consensus and ChatGPT v5.2

specialties, admission to specialist training in DMFR is determined by performance in the DUS. This national examination plays a crucial role in selecting candidates for postgraduate dental education and is therefore an important indicator of the knowledge and skills expected of future specialists [11–14]. While psychometric properties such as item difficulty and discrimination are important indicators of examination quality, these aspects have already been evaluated in previous studies [4, 11, 12, 14–18]. Therefore, the present study focused specifically on the cognitive dimension of assessment using Bloom's revised taxonomy, together with curriculum relevance and scientific accuracy. Since DUÇEP defines national minimum standards for undergraduate dental education, this study provides one of the first DUÇEP-referenced, Bloom-aligned content analyses of DUS DMFR questions.

The distribution of cognitive levels showed that “Analyze” was the most frequent category in both expert consensus (41.5%) and ChatGPT v5.2 classification (44.6%). This finding is consistent with the diagnostic nature of DMFR, where interpretation of radiographic findings and integration of diagnostic clues are central competencies. The second most frequent level was “Remember” (33.1% vs. 25.4%), indicating that many questions still rely on factual recall, such as imaging terminology, radiologic characteristics, or anatomical landmarks. In contrast, “Understand” and “Apply” levels were relatively infrequent in both classifications, suggesting that questions requiring conceptual explanation or direct application

**Fig. 5** Boxplot of scientific accuracy ratings (Likert 1–5) for expert consensus and ChatGPT v5.2

of knowledge in basic clinical scenarios are less common than analytical interpretation tasks. The “Evaluate” level accounted for only a small number of questions, and no items were classified at the “Create” level. This is not unexpected and primarily reflects the structural limitations of MCQ-based examination formats, which typically require selecting the best answer among predefined options rather than generating new diagnostic strategies or problem-solving approaches.

Recent studies have explored the performance of LLMs in examination settings, suggesting increasing potential in educational contexts, particularly as supportive learning tools [19, 20]. In the present study, agreement between ChatGPT v5.2 and expert consensus was high, with an exact agreement rate of 80.0% and almost perfect ordinal agreement according to quadratic-weighted Cohen’s kappa ($\kappa_w = 0.851$). These findings indicate that ChatGPT v5.2 can approximate expert judgement for Bloom-level categorization in this dataset. However, Bowker’s test of symmetry revealed statistically significant asymmetry in the pattern of disagreements ($p = 0.038$), indicating that discrepancies were not randomly distributed across Bloom categories. As shown in Fig. 2 and detailed in the Results, this asymmetry primarily reflected directional borderline shifts between adjacent Bloom levels. In DMFR, where image interpretation often involves overlapping cognitive processes, slight variations between adjacent levels, particularly “Apply”, “Analyze”, and “Evaluate”, are expected. Therefore, LLM-based classifications should be interpreted cautiously, especially for borderline items where distinctions between cognitive levels are less clearly defined.

In the Turkish dental education system, the third and fourth years are dedicated primarily to practical clinical training, with the third year marking the transition from theoretical learning to hands-on clinical practice. At this stage, students are expected to integrate theoretical knowledge with clinical understanding and decision-making skills. In a previous report from our study group, the importance of practical training during these clinical years was examined, particularly its role in developing practical radiographic technique skills [21]. In line with this educational framework, recent studies evaluating multiple LLMs in dental specialty examinations have reported relatively high performance in knowledge-based tasks [22–24], but reduced accuracy in clinically oriented questions requiring more complex reasoning. Similarly, in clinical applications such as differential diagnosis of oral lesions, LLMs have shown promising performance when broader diagnostic ranges are considered, but they still underperform compared with experts in precise, top-level decisions [25, 26]. Overall, current evidence suggests that LLM performance is highly context-dependent, with limitations in tasks requiring detailed morphological

interpretation, clinical judgement, and complex decision-making [24–26]. These findings are consistent with the present study, where strong agreement with expert cognitive classification was observed alongside minor variability in higher-order or borderline cases. Therefore, the role of LLMs in dental education and clinical practice should be considered supportive rather than substitutive, requiring expert validation and oversight.

Previous work in medical and dental education has shown that MCQ-based examinations may overrepresent LOCS. The cognitive level of assessment items can also influence students’ learning strategies: LOCS-oriented assessments are more likely to promote surface learning and memorization, whereas HOCS-oriented items may encourage deeper conceptual learning and reasoning, particularly in high-stakes settings [27–31]. Table 4 provides a narrative overview of representative studies relevant to Bloom-based item analysis and educational applications. Due to the emerging and multidisciplinary nature of this field, the included studies are heterogeneous and are not intended to constitute a systematic review. In the context of DMFR, where clinical competence requires application of imaging principles, interpretation of radiographic findings, and diagnostic reasoning, Bloom’s taxonomy may help inform assessment design by supporting more intentional alignment between learning objectives and examination content. Accordingly, increasing the proportion of items targeting higher-order cognitive processes, especially Apply, Analyze, and Evaluate, may better reflect the cognitive demands of radiologic decision-making and support educational coherence in DMFR training.

Expert consensus indicated very high curriculum relevance, with most items receiving the maximum rating of 5/5. Similarly, ChatGPT v5.2 assigned the highest possible curriculum relevance score to all items. However, these findings should be interpreted in light of a clear ceiling effect in both curriculum relevance and scientific accuracy ratings, suggesting measurement saturation and limited discriminatory capacity of these secondary outcomes. Several factors may explain this pattern, including the limited sensitivity of a 5-point Likert scale for distinguishing subtle differences among high-quality examination items, the high-stakes nature of the DUS and the likelihood of prior expert validation, and the broad scope of the DMFR curriculum, where topic-level alignment may be captured more readily than differences in depth or competency level [8]. Therefore, these results should not be interpreted as strong confirmatory evidence of optimal curriculum alignment, but rather as an indication that these measures had limited analytical value in this context. This pattern may reflect measurement saturation when high-quality standardized examination items are evaluated using relatively coarse rating scales [11, 22,

Table 4 Overview of representative studies on Bloom's taxonomy and LLM applications in dental and medical education

Authors	Year	Country	Context / Exam Type	Sample	LLM	Bloom Taxonomy	Key Findings	Educational Implication
Gonzalez-Cabezas et al.	2015	USA	Dental education – student- and instructor-generated exam questions	104 intervention students and 106 historical controls	None	Modified Bloom taxonomy (3-level) used to classify cognitive level	Student-generated questions were classified at higher cognitive levels than instructor-generated questions; the intervention cohort performed as well as or better than controls	Student item-writing may promote higher-order learning and support critical thinking
Rossettini et al.	2024	Italy	Healthcare sciences university entrance exam	820 MCQs	ChatGPT-4, Copilot, Gemini	No Bloom classification	ChatGPT-4 and Copilot outperformed Gemini; high overall accuracy	LLMs show strong performance in structured standardized tests but require cautious educational integration
Turunç-Oğuzman & Yurdabakan	2024	Türkiye	Dental Specialty Entrance Exam (DUS)	123 MCQs	ChatGPT-3.5, Bard	Lower vs. Higher Bloom classification	LLMs showed limited agreement with authors' Bloom-level categorization; Bloom level did not significantly affect accuracy	LLMs are not yet reliable for high-stakes assessment use but may serve supportive roles
Wei & Peng	2025	China	Undergraduate endodontics education (Bloom-based flipped classroom)	32 students (16 lecture vs. 16 flipped)	None	Bloom taxonomy used to structure learning objectives	Flipped classroom group: theoretical knowledge self-rating + 10.9%, exam score + 17.1%	Bloom-based flipped classroom may improve learning outcomes
Geduk et al.	2026	Türkiye	DUS – Oral & Maxillofacial Radiology	208 MCQs	ChatGPT 4.0, ChatGPT 4.5, DeepSeek-R1, Google Gemini, Microsoft Copilot, Grok, Claude	Not applied	ChatGPT 4.0 showed highest accuracy (91.3%); image-based performance significantly lower	LLMs useful as supplementary educational tools but not stand-alone decision systems
Nguyen et al.	2026	Vietnam	AI-generated dental MCQs	300 questions ($n=60$ for each model)	ChatGPT-4o, Copilot Pro, Claude Sonnet 4, Grok 3, DeepSeek R1	Full Bloom (Likert-based evaluation)	Claude Sonnet 4 showed strongest higher-order alignment; most LLMs stronger at lower levels	LLMs can assist in MCQ generation but vary in higher-order cognitive alignment
Present Study	2026	Türkiye	DUS-DMFR questions	130 questions	ChatGPT v5.2	Bloom taxonomy classification compared with expert consensus	ChatGPT v5.2 showed high agreement with experts ($\kappa_w=0.851$) in Bloom-level classification	LLM may support cognitive analysis of DMFR exam questions, but expert evaluation remains essential.

23, 28]. Consistent with this interpretation, expert consensus rated scientific accuracy as fully accurate (5/5) for all items, which is expected for standardized examination materials, whereas ChatGPT v5.2 provided slightly lower scientific accuracy ratings. Future studies may benefit from more granular scoring systems or alternative frameworks to better capture variability among high-quality items [11, 18, 28].

Strengths of this study include the use of independent expert ratings followed by a structured consensus process supported by substantial pre-consensus agreement, and the inclusion of a reproducibility assessment of ChatGPT classifications across repeated runs. Several limitations should also be acknowledged. First, only one LLM (ChatGPT v5.2) was evaluated; therefore, the findings are model- and version-specific and may not generalize to

other LLMs or future updates. Second, agreement with expert classification does not necessarily translate into pedagogical validity or assessment effectiveness. Third, expert consensus represents a pragmatic reference standard and may introduce classification bias in the absence of an objective gold standard for Bloom-level assignment. Fourth, the uniformly high curriculum relevance and scientific accuracy ratings indicate ceiling effects, limiting the discriminatory value of these secondary outcomes in the current design. Finally, two authors previously participated in early, officially organized DUS question-preparation workshops around the initial implementation of the DUS in 2012. However, this study analyzed only publicly released DMFR booklets from 2012 to 2021, the authors had no access to unreleased items, and to the best of our knowledge, none of the analyzed questions were authored by the study team.

Conclusions

This study examined the cognitive complexity of DMFR questions in the Turkish DUS using Bloom's revised taxonomy and evaluated agreement between expert consensus and ChatGPT v5.2 classifications. "Analyze" was the most frequent cognitive level, reflecting the interpretive nature of radiology assessments, while a substantial proportion of items still targeted lower-order cognitive processes. ChatGPT v5.2 showed high agreement with expert consensus cognitive-level classification in this dataset. These findings provide preliminary evidence that LLM-based classification may support Bloom-level mapping as an assistive tool; however, they should be interpreted as evidence of feasibility within a single domain, examination context, language setting, and model version. Therefore, LLM outputs should be used with expert oversight, and further research is needed to assess multi-model performance, stability over time, and practical impact in real assessment workflows.

Abbreviations

CBCT	Cone-beam computed tomography
CI	Confidence interval
CT	Computed tomography
DUÇEP	National Core Curriculum for Undergraduate Dental Education
DUS	Turkish Dentistry Specialization Examination
HOCS	Higher-order cognitive skills
LLM	Large language model
LOCS	Lower-order cognitive skills
MCQs	Multiple-choice questions
MRI	Magnetic resonance imaging
DMFR	Dentomaxillofacial Radiology
US	Ultrasonography

Acknowledgements

Not applicable.

Authors' contributions

Material preparation, data collection and analysis were performed by M.O., H.C.A. The manuscript was written by M.O., B.G., and S.E.Y. The final version of the manuscript was approved by M.O., H.C.A., B.G., and S.E.Y.

Funding

No external funding was obtained for this study.

Data availability

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

Declarations

Ethics approval and consent to participate

Not applicable. Open-source public data was used in this study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 14 March 2026 / Accepted: 11 May 2026

Published online: 21 May 2026

References

- Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM). 2012 yılında uygulanacak olan Tıpta Uzmanlık Eğitimi Giriş Sınavı (TUS) ile Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavı (DUS) hakkında açıklama. Ankara (TR): ÖSYM; 2011. Available from: <https://www.osym.gov.tr/TR,5818/2012-yilinda-uygulanacak-olan-tipta-uzmanlik-egitimi-giris-sinavi-tus-ile-dis-hekimligi-uzmanlik-egitimi-giris-sinavi-du-s-hakkinda-aciklama-19102011.html>. Cited 2026 Mar 8.
- Gibbs G, Simpson C. Conditions under which assessment supports students' learning. *Learn Teach High Educ*. 2005;(1):3–31.
- Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: an unavoidable truth? *Anat Sci Educ*. 2009;2(5):199–204. <https://doi.org/10.1002/ase.102>.
- Yalçın O, Topal CG. Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavında Sorulan Ağız, Diş ve Çene Radyolojisi Sorularının Retrospektif Analizi [Retrospective analysis of oral and maxillofacial radiology questions asked in the Dentistry Specialization Examination]. *J Int Dent Sci*. 2025;11(3):193–201. <https://doi.org/10.21306/dishekimligi.1776983>. Turkish.
- Anderson LW, Krathwohl DR, editors. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York (NY): Longman; 2001.
- Krathwohl DR. A revision of Bloom's taxonomy: an overview. *Theory Pract*. 2002;41(4):212–8. https://doi.org/10.1207/s15430421tip4104_2.
- Ray ME, Rudolph MJ, Daugherty KK. Bloom's taxonomy in health professions education: Associations with exam scores, clinical reasoning, and instructional effectiveness. *Curr Pharm Teach Learn*. 2025;17(11):102444. <https://doi.org/10.1016/j.cptl.2025.102444>.
- Yükseköğretim Kurulu. Mezuniyet Öncesi Diş Hekimliği Ulusal Çekirdek Eğitim Programı-2024 (DUÇEP-2024) . Ankara (TR): Yükseköğretim Kurulu; 2024. Available from: <https://egitim.yok.gov.tr/page/499>. Cited 2026 Mar 8.
- Van Der Vleuten CP. The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ Theory Pract*. 1996;1(1):41–67. <https://doi.org/10.1007/BF00596229>.
- Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM). DUS çıkmış sorular. Ankara (TR): ÖSYM; n.d. Available from: <https://www.osym.gov.tr/TR,15070/dus-cikmis-sorular.html>. Cited 2026 Mar 8.
- Turunç-Oğuzman R, Yurdabakan ZZ. Performance of Chat Generative Pre-trained Transformer and Bard on the Questions Asked in the Dental Specialty Entrance Examination in Türkiye Regarding Bloom's Revised Taxonomy. *Curr Res Dent Sci*. 2024;34(1):25–34. <https://doi.org/10.5152/CRDS.2024.23261>.
- Peker RB. ChatGPT, ChatGPT, Plus, Gemini ve Microsoft Copilot Büyük Dil Modellerinin Türkiye'deki Diş Hekimliği Uzmanlık Eğitimi Giriş Sınavında Sorulan Ağız, Diş ve Çene Radyolojisi Sorularını Cevaplama Performanslarının Değerlendirilmesi [Evaluation of the performance of the large language models ChatGPT, ChatGPT, Plus, Gemini, and Microsoft Copilot in answering oral and maxillofacial radiology questions from the Turkish Dentistry Specialization Examination]. *Yeditepe J Dent*. 2025;21(3):130–5. <https://doi.org/10.5505/yeditepe.2025.93265>. Turkish.

13. Sismanoglu S, Capan BS. Performance of artificial intelligence on Turkish dental specialization exam: can ChatGPT-4.0 and gemini advanced achieve comparable results to humans? *BMC Med Educ*. 2025;25(1):214. <https://doi.org/10.1186/s12909-024-06389-9>.
14. Dundar Sari MB, Sezer B. Comparative performance evaluation of ChatGPT-4 Omni and Gemini Advanced in the Turkish Dentistry Specialization Exam. *BMC Med Educ*. 2026;26(1):251. <https://doi.org/10.1186/s12909-026-08621-0>.
15. Gerzeli EA. Dış Hekimliğinde Uzmanlık Sınavında Sorulmuş Ağız, Diş ve Çene Radyolojisi Sorularının Retrospektif Analizi [Retrospective analysis of Oral, Dental and Maxillofacial Radiology questions asked in the specialty exam in dentistry]. *Türkiye Klinikleri J Dent Sci*. 2026;32(1):85–92. <https://doi.org/10.5336/dentalsci.2025-112452>. Turkish.
16. Geduk G, Hasırcı UC, Kusay DD, Aras RÇ, Çapar İ, Altın E, et al. A comparative analysis of the performance of large language models in the dentistry specialty examination. *Sci Rep*. 2026;16(1):6739. <https://doi.org/10.1038/s41598-026-37800-8>.
17. Turan Gökduman C, Arılı Öztürk E, Aktaş Ş, Çanakçı BC. Comparison of chatbots' accuracy in endodontics questions in dentistry specialization exam in Türkiye: ChatGPT-4o, Gemini Advanced, Copilot, and Claude. *BMC Oral Health*. 2025;26(1):28. <https://doi.org/10.1186/s12903-025-07346-8>.
18. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. *BMC Oral Health*. 2025;25(1):573. <https://doi.org/10.1186/s12903-025-05926-2>.
19. Rossetini G, Rodeghiero L, Corradi F, Cook C, Pillastrini P, Turolla A, et al. Comparative accuracy of ChatGPT-4, Microsoft Copilot and Google Gemini in the Italian entrance test for healthcare sciences degrees: a cross-sectional study. *BMC Med Educ*. 2024;24(1):694. <https://doi.org/10.1186/s12909-024-05630-9>.
20. Tassoker M. ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. *BMC Oral Health*. 2025;25(1):173. <https://doi.org/10.1186/s12903-025-05554-w>.
21. Polat Akmansoy B, Gungor T, Cebeci T, Ercalik Yalcinkaya S, Gumru B. Effect of practical demonstration on reducing technical radiographic errors made by dental students. *J Dent Educ* Published online November. 2025;24. <https://doi.org/10.1002/jdd.70109>.
22. Gokkurt Yilmaz BN, Ozbey F, Yilmaz BE. Evaluation of the performance of different large language models on head and neck anatomy questions in the dentistry specialization exam in Turkey. *Surg Radiol Anat*. 2025;47(1):211. <https://doi.org/10.1007/s00276-025-03723-8>.
23. Kandemir M, Sarıbaş EE. Comparative performance of large language models in answering periodontology questions from the Turkish Dental Specialty Examination: a cross-sectional study on accuracy and coverage. *BMC Oral Health*. 2025;25(1):1804. <https://doi.org/10.1186/s12903-025-07241-2>.
24. Hassanein FEA, Hussein RR, Ahmed Y, El-Guindy J, Ahmed DE, Abou-Bakr A. Calibration of AI large language models with human subject matter experts for grading of clinical short-answer responses in dental education. *BMC Oral Health*. 2026;26(1):286. <https://doi.org/10.1186/s12903-026-07665-4>.
25. Abou-Bakr A, El Barbary A, Hassanein FEA. ChatGPT-5 vs oral medicine experts for rank-based differential diagnosis of oral lesions: a prospective, biopsy-validated comparison. *Odontology* Published online November. 2025;17. <https://doi.org/10.1007/s10266-025-01242-x>.
26. Hassanein FEA, Alkabazi M, Tassoker M, Ahmed Y, Alsaeed S, Abou-Bakr A. Multimodal large language models for oral lesion diagnosis: a systematic review of diagnostic performance and clinical utility. *Front Oral Health*. 2026;7:1748450. <https://doi.org/10.3389/froh.2026.1748450>.
27. Wei Y, Peng Z. Application of the flipped classroom model based on Bloom's taxonomy of educational objectives in endodontics education for undergraduate dental students. *PeerJ*. 2025;13:e18843. <https://doi.org/10.7717/peerj.18843>.
28. Alkabazi M, Tassoker M. Comparative performance of AI models on case-based oral medicine questions across Bloom's taxonomy levels and subtopics. *Odontology*. 2026;14. <https://doi.org/10.1007/s10266-026-01335-1>. Epub ahead of print.
29. Thompson AR, Lake LPO. Relationship between learning approach, Bloom's taxonomy, and student performance in an undergraduate Human Anatomy course. *Adv Health Sci Educ Theory Pract*. 2023;28(4):1115–30. <https://doi.org/10.1007/s10459-023-10208-z>.
30. Nguyen TT, Nguyen L, Do HTN, Nguyen HTT, Tong SM. Evaluating cognitive depth of AI-generated multiple-choice questions with Bloom's Taxonomy. *PLoS ONE*. 2026;21(2):e0341317. <https://doi.org/10.1371/journal.pone.0341317>.
31. Gonzalez-Cabezas C, Anderson OS, Wright MC, Fontana M. Association between dental student-developed exam questions and learning at higher cognitive levels. *J Dent Educ*. 2015;79(11):1295–304.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.