Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Classification Fidelity of Large Language Models on Endodontic Items from the Turkish Dental Specialty Examination

Bilge Hakan Şen

hakan.sen@tinaztepe.edu.tr

Izmir Tınaztepe University

Aslıhan Yekeler

Izmir Tınaztepe University

Duygu Kürklü Arpaçay

Izmir Tınaztepe University

Jülide Ocak

Izmir Tınaztepe University

Oğuzhan Akkoçan

Dokuz Eylül University

Research Article

Keywords: Large language models, Bloom's Taxonomy, SOLO Taxonomy, Miller's Pyramid, Endodontics

Posted Date: July 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9913605/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Medical Education on October 30th, 2026. See the published version at https://doi.org/10.1186/s12909-026-10746-1.

Abstract

Background

Large language models (LLMs) are increasingly used in dental education, but research has focused almost exclusively on answer accuracy. Whether LLMs can identify the cognitive demand of examination items, a distinct competency relevant to assessment quality assurance and curriculum mapping, remains untested. This study evaluated LLM classification fidelity for endodontic multiple-choice questions (MCQs) across three established educational taxonomies and compared performance among four publicly accessible models.

Methods

Two hundred endodontic MCQs from the Turkish Dentistry Specialty Examination (DUS, 2012–2025) were classified by ChatGPT, Gemini, Kimi, and DeepSeek according to Revised Bloom's Taxonomy, the SOLO Taxonomy, and Miller's Pyramid. A three-layer expert consensus procedure (calibration training, procedural audit, and structured joint deliberation) established the gold standard, with two endodontists rating items and a medical education professor serving as external auditor and arbitrator. LLMs received a standardised prompt in dedicated incognito browser sessions to eliminate cross-session memory and personalisation effects. Agreement was assessed using weighted Cohen's kappa (κw) with bootstrap 95% confidence intervals (2,000 iterations); classification performance was evaluated using accuracy and macro-averaged F1; inter-model differences in item-level accuracy were tested using Cochran's Q with Kendall's W as effect size, followed by Holm–Bonferroni-corrected pairwise McNemar tests (□=0.05).

Results

For Bloom's Taxonomy, Gemini (κw = 0.676) and DeepSeek (κw = 0.667) achieved substantial agreement; ChatGPT (κw = 0.365) and Kimi (κw = 0.428) performed lower. For the SOLO Taxonomy, Gemini reached substantial agreement (κw = 0.622), while the remaining models achieved moderate agreement. For Miller's Pyramid, agreement did not exceed fair for any model, with ChatGPT highest (κw = 0.372). Cochran's Q confirmed significant inter-model differences for all three taxonomies (all $p < 0.001$). ChatGPT ranked lowest for Bloom and SOLO but highest for Miller, a complete reversal of model rankings across frameworks. Gold standard Bloom levels increased significantly across examination years (Spearman $\rho = 0.301$, $p < 0.001$), indicating that DUS endodontic questions have become progressively more cognitively demanding over time.

Conclusions

LLMs can classify cognitive complexity with meaningful fidelity for Bloom and SOLO taxonomies, particularly Gemini and DeepSeek, but performance for Miller's Pyramid remains weak across all models. Bloom, SOLO, and Miller assess distinct constructs and should not be treated as interchangeable; LLM selection should be tailored to the specific taxonomy and assessment purpose. Answer accuracy and taxonomy classification fidelity represent separate capabilities that warrant independent evaluation.

Introduction

Large language models (LLMs) are a class of artificial intelligence systems capable of processing and generating human language at scale through transformer-based deep learning architectures. Since the public release of ChatGPT in late 2022, their potential applications in health professions education have attracted rapid and widespread interest, encompassing clinical decision support, patient communication, and knowledge assessment [1].

A substantial body of evidence has examined LLM performance on multiple-choice questions (MCQs) in dental licensing and specialty examinations. A systematic review and meta-analysis incorporating eleven studies across eight countries found that GPT-4 achieved an integrated accuracy of 72% across dental licensing examinations worldwide, substantially outperforming GPT-3.5 (54%) and Bard (56%), while noting that overall accuracy remained below the threshold required for clinical application [2]. At the level of individual examinations, ChatGPT-4.0 successfully passed both the US and UK dental licensing examinations, correctly answering 80.7% and 62.7% of questions respectively [3]. A recent comparative study evaluated 11 LLMs - including both free and subscription-based versions - on 60 multiple-choice questions derived from American Association of Endodontists and European Society of Endodontology position statements, finding that ChatGPT-4o and Claude Opus 4 achieved the highest all-correct accuracy (95.0%), while DeepSeek performed substantially lower (63.3%) [4]. In addition, intra-model consistency also varied widely, from 98.3% for top performers to 75.0% for DeepSeek, underscoring that reproducibility cannot be assumed even among high-performing models [4]. Studies particularly evaluating the Turkish Dentistry Specialty Examination (DUS) have reported similar patterns across multiple disciplines. A growing body of work has evaluated LLM answer accuracy on Turkish DUS items across multiple specialties, including oral pathology [5] and periodontology [6], with performance varying by model, question type, and clinical content. Most directly relevant to the present study, Çekiç and Tavşan [7] evaluated eight LLMs on 151 Turkish DUS endodontic MCQs and identified ChatGPT-4o as the highest-performing model. However, this body of work shares a fundamental limitation. It evaluates LLM performance exclusively in terms of answer accuracy, whether models select the correct option, without addressing the cognitive demands the questions themselves impose.

The capacity to correctly answer a recall question is qualitatively different from the capacity to reason through a question requiring analysis or clinical judgement. Yet, these distinctions are invisible to accuracy-based metrics alone. Taxonomic classification of examination items addresses this gap directly. Revised Bloom’s Taxonomy, proposed by Anderson et al. [8] as a refinement of the original 1956 framework, hierarchically classifies cognitive demand from lower-order remembering and understanding

to higher-order application, analysis, evaluation, and creation. A scoping review of 24 studies demonstrated that performance on higher-order examination items is consistently lower across health professions, and that taxonomy mapping provides educators with a structured method for understanding and improving the cognitive quality of assessments [9]. The intersection of LLM capability and taxonomic complexity has been examined in only a small number of studies. Herrmann-Werner et al. [10] demonstrated that when GPT-4's errors on psychosomatic medicine MCQs were analysed through the lens of Bloom's Taxonomy, they clustered predominantly at the Remember and Understand levels, revealing qualitative error patterns invisible to overall accuracy rates alone. Similarly, Law et al. [11] found that AI-generated MCQs primarily tested lower-order cognitive skills while human-authored MCQs better assessed higher-order skills ($\chi 2 = 14.27$, $p = 0.003$), underscoring the importance of taxonomic evaluation in AI-assisted assessment design. The Structure of the Observed Learning Outcome (SOLO) taxonomy, introduced by Biggs and Collis [12], complements Bloom's framework by assessing the structural complexity of understanding rather than cognitive process per se, progressing from uni-structural to extended abstract levels of response integration. Sirri et al. [13] demonstrated a significant association between Bloom's Taxonomy and SOLO classifications of MCQs ($\chi 2(1) = 26.12$, $p < .001$), confirming that the two frameworks capture overlapping constructs of cognitive complexity while offering distinct structural perspectives. Miller's Pyramid [14], widely adopted in clinical competency assessment, stratifies knowledge across four levels - Knows, Knows How, Shows How, and Does - directly mapping assessment design to professional performance expectations. ten Cate et al. [15] noted that Miller's framework created worldwide awareness of the need for different assessment approaches at different competency levels, with the lower two levels (Knows, Knows How) characterising the knowledge assessment domain in which written MCQs principally operate.

Despite the growing use of LLMs in dental education contexts and the established role of educational taxonomies in examination quality assurance, no study has systematically evaluated whether LLMs can reliably classify the cognitive complexity of dental specialty examination items simultaneously across multiple taxonomic frameworks, nor compared LLM classifications against an expert consensus gold standard.

The present study addresses this gap using 200 endodontic MCQs drawn from the DUS administered between 2012 and 2025 - a nationally standardised, high-stakes assessment written exclusively in Turkish, spanning 24 distinct endodontic topic categories and providing a comprehensive, authentic non-English evaluation context. Four publicly accessible LLMs (ChatGPT, Gemini, Kimi, and DeepSeek) were tasked with simultaneously classifying each item according to Revised Bloom's Taxonomy, the SOLO Taxonomy, and Miller's Pyramid, using a standardised prompt applied under controlled session conditions. LLM classifications were compared against a three-layer expert consensus gold standard.

Given the absence of prior evidence on LLM taxonomic classification performance in dental specialty examinations, the present study was designed as a comparative exploratory investigation. Three research questions guided the study: (1) To what extent do LLM classifications agree with expert consensus across Revised Bloom's Taxonomy, the SOLO Taxonomy, and Miller's Pyramid? (2) Does

classification fidelity vary systematically across cognitive levels within each taxonomy? (3) Are there significant differences in classification performance among the four LLMs, and do these differences vary by taxonomic framework?

Materials and Methods

This study employed a cross-sectional observational design to evaluate the capacity of four large language models (LLMs) to classify endodontic multiple-choice questions (MCQs) according to established educational taxonomies. LLM-generated classifications were compared against a consensus-derived expert gold standard. The study is reported in accordance with current methodological standards for LLM performance evaluation in health professions education research.

A total of 200 endodontic MCQs administered in the Turkish Dentistry Specialty Examination (Diş Hekimliği Uzmanlık Sınavı, DUS) between 2012 and 2025 were obtained from DUSDATA (https://www.dusdata.com), which holds copyright over the archived DUS item bank, under a written permission agreement for academic research use. DUS is administered by the Assessment, Selection and Placement Center (Öğrenci Seçme ve Yerleştirme Merkezi, ÖSYM) of Türkiye as a national entrance examination through which dentists compete for placement into postgraduate specialty programs based on their overall examination score. Questions are prepared based on internationally recognised dental textbooks selected by ÖSYM and are written in Turkish.

MCQs were presented in their original Turkish language and included the question stem and all answer options. Items were distributed across 14 examination years (2012–2025), with 10–20 items per year depending on the number of endodontic questions included in each administration cycle. No items were excluded or modified. Two endodontists (AY and JO) independently classified the 200 MCQs according to a pre-established topic coding scheme comprising 18 broad topic domains, several of which contained subcategories yielding 24 distinct thematic codes in total. Each item was assigned a primary topic code, with a secondary code permitted when an item substantively addressed two domains. Inter-rater agreement was assessed using Cohen's unweighted kappa.

A three-layer expert consensus procedure was used to establish the gold standard taxonomy classifications. Two raters (AY and DKA) performed the gold standard classification: One endodontist and one prosthodontist, both academic clinicians with active roles in dental education and formal involvement in advanced teaching programmes. Taxonomy classification is an educational measurement task that depends on expertise in cognitive assessment design rather than on subject specialty; the panel was therefore composed to ensure that both raters held substantive educational expertise across the three frameworks, while the endodontist additionally contributed direct content familiarity with the item domain.

Calibration training

Both raters underwent structured calibration training prior to classification. The training was delivered by an external advisor with expertise in medical education and educational taxonomy. Before any items were classified, this process ensured that both raters shared a theoretically grounded and consistent understanding of Revised Bloom's Taxonomy, the SOLO Taxonomy, and Miller's Pyramid. This directly addressed a common criticism of consensus panels: that agreement may reflect shared misconceptions rather than genuine expertise.

Procedural audit

The external advisor reviewed the raters' initial classifications against the taxonomic frameworks and confirmed methodological adequacy before the full classification procedure proceeded. This checkpoint is functionally analogous to a validity check by an independent expert.

Structured joint consensus

Final item classifications were determined through structured joint deliberation between the endodontist and the prosthodontist. Where disagreement arose, items were discussed until consensus was reached. The external advisor served as an independent procedural endorser throughout the process. This three-layer procedure - calibration, external audit, and structured consensus - constituted a stronger methodological safeguard than independent parallel rating designs, which yield inter-rater statistics but do not include external validation of the rating process itself.

Each item was classified according to three established educational taxonomic frameworks:

Revised Bloom’s Taxonomy [8]: Six cognitive levels - Remember, Understand, Apply, Analyse, Evaluate, and Create. Level 6 (Create) was included in the framework presented to both the expert panel and LLMs but was not expected to occur in standard MCQ formats.

SOLO Taxonomy [12]: Five levels - Pre-structural, Uni-structural, Multi-structural, Relational, and Extended Abstract. All five levels were included; however, Pre-structural is defined as applicable only when a question cannot be engaged with at any relevant level and was not expected to be assigned to items drawn from a standardised examination.

Miller’s Pyramid [14]: Four levels - Knows, Knows How, Shows How, and Does. All four levels were presented; however, Miller’s ‘Does’ requires direct performance observation and was not expected to be assigned to MCQ-based items.

Four large language models representative of the current landscape of publicly accessible generative AI systems were selected: ChatGPT, Gemini, Kimi, and DeepSeek. Selection criteria included broad adoption in academic and educational contexts, availability via consumer-facing web interfaces, and support for Turkish-language input. The model versions active during the data collection period were: ChatGPT (GPT-

5.2 Instant; chat.openai.com), Gemini (Gemini 3.1 Pro; gemini.google.com), Kimi (K2.5; kimi.ai), and DeepSeek (V3.2; chat.deepseek.com), all accessed via their consumer web interfaces between February 23–26, 2026. Generation parameters including temperature were platform-managed and not modifiable through these interfaces, which is acknowledged as a limitation precluding full output reproducibility. API-based access with fixed temperature settings would allow greater methodological standardisation in future studies.

A standardised prompt was developed and applied identically across all four LLMs to ensure comparability (Supplementary Appendix I). The prompt assigned each LLM a multi-domain expert role encompassing endodontics, dental education, medical assessment, and educational taxonomy. Contextual background information regarding the DUS examination (including its purpose, administration by ÖSYM, Turkish-language format, and textbook-based item construction) was provided to anchor classifications within the appropriate institutional context.

To minimise cross-session contamination and personalisation effects, each annual examination batch was submitted in a dedicated incognito browser session, ensuring that no prior classification history or session memory influenced subsequent responses. This approach was intended to preserve classification independence across examination years and represents a methodological refinement compared with previous studies that did not explicitly report session isolation procedures.

Questions were presented with all answer options included, in annual batches corresponding to the original examination structure. Depending on the year, each batch contained 10 or 20 questions. For each item, LLMs were instructed to classify cognitive demand according to all three frameworks simultaneously: Revised Bloom's Taxonomy (six levels), SOLO Taxonomy (five levels), and Miller's Pyramid (four levels).

Explicit methodological instructions were embedded within the prompt to promote conservative and consistent classification, including directives to avoid artificial inflation of cognitive levels and to apply careful differentiation between adjacent taxonomy categories (e.g., Apply vs. Analyse; Knows vs. Knows How). Although questions were submitted in annual batches, all analyses and classifications were performed at the individual item level.

No replicate sessions were conducted, as the study was designed as a single-session exploratory evaluation reflecting real-world LLM usage conditions.

Statistical Analysis

LLMs were used exclusively for taxonomy classification tasks. All statistical analyses were subsequently performed independently in Python (version 3.12) using the pandas, numpy, scipy, scikit-learn, scikit-posthocs, and openpyxl libraries. The statistical significance threshold was set at $\alpha = 0.05$. The analytical framework addressed three levels of inquiry: (1) agreement between LLM classifications and the gold standard, (2) classification performance, and (3) inter-LLM comparison.

Agreement analysis

Agreement between each LLM and the gold standard was assessed using weighted Cohen's kappa (κw). The weighting scheme was determined by the ordinal structure of each taxonomy rather than by the LLM evaluated: quadratic weights were applied to Bloom's Taxonomy and SOLO Taxonomy, where larger ordinal discrepancies carry proportionally greater cognitive significance; linear weights were applied to the Miller Pyramid, where the progression across levels reflects a more uniform cognitive gradient. Bootstrap 95% confidence intervals for κw were computed using 2,000 resampling iterations to quantify the precision of each agreement estimate. Kappa values were interpreted according to the benchmarks proposed by Landis and Koch [16]: <0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and > 0.80 almost perfect agreement. Spearman's ρ was additionally computed as an ordinal correlation measure between LLM ratings and gold standard labels.

Classification performance

Overall accuracy and macro-averaged precision, recall, and F1 were computed for each LLM–taxonomy combination. Macro-averaging was selected to ensure equal weighting of all cognitive levels regardless of class frequency - a critical consideration given the marked imbalance across Bloom levels (Level 1: n = 63; Level 5: n = 5). Level-specific precision, recall, and F1 were additionally computed to characterise per-level classification fidelity. Confusion matrices were examined to characterise the directionality of misclassification errors.

Multi-LLM comparison

Differences in item-level accuracy among the four LLMs were evaluated using the Cochran's Q test, with Kendall's W reported as the corresponding effect size. Where the Cochran's Q test indicated significant differences, pairwise McNemar tests were conducted with p-values adjusted using the Holm–Bonferroni correction.

Results

Question Dataset Characteristics

Topic classification of the 200 DUS endodontic MCQs by two independent expert endodontists yielded almost perfect inter-rater agreement (Cohen's κ = 0.936, 95% CI: 0.898–0.968; percentage agreement: 94.0%). Of the 12 discordant items, nine represented prioritisation differences, one rater's primary code matched the other's secondary code, rather than genuine content disagreements. The remaining three fully discordant items were resolved through structured panel discussion with a third experienced endodontist [BHŞ] as an arbitrator. The final classification identified 24 distinct thematic categories, subsequently grouped into 18 broad topic domains (Table 1).

Table 1
Broad topic category summary (n = 200)

Broad Category	Codes Included	n	%
Pulpal & Periapical Diseases	C1	18	9.0%
Diagnosis & Imaging	C2	14	7.0%
Endodontic Anatomy & Access	C3	16	8.0%
Instrumentation	C4	19	9.5%
Irrigation & Disinfection	C5	16	8.0%
Intracanal Medication	C6	3	1.5%
Obturation	C7	9	4.5%
Retreatment & Surgery	C8a, C8b	8	4.0%
Regenerative & Vital Therapy	C9	11	5.5%
Endodontic Materials	C10a, C10b	6	3.0%
Restorative-Endodontic Rel.	C11	4	2.0%
New Technologies & AI	C12	2	1.0%
Emergencies, Pain & Anaesthesia	C13	7	3.5%
Complications	C14	18	9.0%
Bleaching of Root-Filled Teeth	C15	6	3.0%
Traumatic Injuries & Resorption	C16	13	6.5%
Education & Bibliometrics	C17	0	0%
Basic Sciences	C18a, C18b, C18c, C18d, C18e	30	15.0%
TOTAL		**200**	**100.0%**

Note: C8a: Non-surgical retreatment; C8b: Endodontic surgery. C10a: Biomaterials and calcium silicate cements; C10b: Other endodontic materials. C18a: Microbiology; C18b: Pathology; C18c: Pharmacology; C18d: Histology. Basic Sciences subcategories (C18a–C18d) are reported separately within the Basic Sciences domain.

The most frequently represented topics were Basic Sciences (n = 30, 15.0%), Instrumentation (n = 19; 9.5%), Pulpal and Periapical Diseases (n = 18; 9.0%), and Complications (n = 18; 9.0%), collectively accounting for 42.5% of the item bank. Intracanal Medication (n = 3; 1.5%), Digital Technologies (n = 2; 1.0%), and Biochemistry (n = 1; 0.5%) were the least represented categories. Sixteen items (8.0%) were assigned a secondary topic code. The most frequently assigned secondary codes were Pulpal and Periapical Diseases (n = 3), Emergencies, Pain and Anaesthesia (n = 2), and Biomaterials and Calcium Silicate (n = 2).

Gold Standard Label Distribution

Of the 200 DUS endodontic MCQs, the gold standard panel assigned Bloom's Taxonomy levels ranging from 1 to 5; Level 6 (Create) was not assigned by either panellist, consistent with the theoretical incompatibility of synthesis-level demands with MCQ formats. The dataset was markedly imbalanced: Level 2 (Understand) accounted for nearly half of all items (n = 95, 47.5%), while Levels 4 and 5 together comprised only 5.5% (n = 11). Accordingly, 158 items (79.0%) were classified as lower-order thinking (LOT; Bloom Levels 1–2) and 42 items (21.0%) as higher-order thinking (HOT; Bloom Levels 3–5) (Table 2).

For the SOLO Taxonomy, items were distributed across only three levels (Uni-structural: n = 69, 34.5%; Multi-structural: n = 95, 47.5%; Relational: n = 36, 18.0%). Pre-structural and Extended Abstract were not assigned. For the Miller Pyramid, 68 items (34.0%) were classified at Level 1 (Knows) and 132 (66.0%) at Level 2 (Knows How); Levels 3 (Shows How) and 4 (Does) were not assigned, consistent with the theoretical non-applicability of performative levels to MCQ-based assessment.

Table 2
Gold standard label distribution across taxonomies (N = 200).

Taxonomy	Level	n	%
Bloom	Level 1 - Remember	63	31.5
	Level 2 - Understand	95	47.5
	Level 3 - Apply	31	15.5
	Level 4 - Analyze	6	3.0
	Level 5 - Evaluate	5	2.5
	Level 6 - Create	0	0.0
	LOT (Levels 1–2)	158	79.0
	HOT (Levels 3–5)	42	21.0
SOLO	Level 1 - Pre-structural	0	0.0
	Level 2 - Uni-structural	69	34.5
	Level 3 - Multi-structural	95	47.5
	Level 4 - Relational	36	18.0
	Level 5 - Extended Abstract	0	0.0
Miller	Level 1 - Knows	68	34.0
	Level 2 - Knows How	132	66.0
	Level 3 - Shows How	0	0.0
	Level 4 - Does	0	0.0

LOT: lower-order thinking (Bloom Levels 1–2); HOT: higher-order thinking (Bloom Levels 3–5). Levels not assigned by the expert panel are shown as n = 0.

A significant positive correlation was observed between examination year and gold standard Bloom level (Spearman $\rho = 0.301$, $p < 0.001$), indicating that DUS endodontic MCQs became progressively more cognitively demanding over the 2012–2025 study period.

The mean gold standard Bloom level was 1.93 (SD = 0.82), reflecting the predominance of lower-order items. Mean SOLO and Miller levels were 2.84 (SD = 0.70) and 1.66 (SD = 0.47), respectively.

LLM Agreement with Gold Standard

Weighted kappa values and Spearman correlations for all LLM–taxonomy combinations are presented in Table 3. For Bloom's Taxonomy, Gemini achieved the highest agreement ($\kappa w = 0.676$, 95% CI: 0.575–

0.753, substantial), closely followed by DeepSeek (κw = 0.667, 95% CI: 0.551–0.752, substantial). Kimi (κw = 0.428, moderate) and ChatGPT (κw = 0.365, fair) performed considerably lower. For the SOLO Taxonomy, Gemini again led (κw = 0.622, substantial), while ChatGPT, Kimi, and DeepSeek achieved moderate agreement (κw = 0.433–0.549). The Miller Pyramid produced a notably different pattern: ChatGPT achieved the highest agreement (κw = 0.372, fair), while Gemini (κw = 0.235, fair), Kimi (κw = 0.146, slight), and DeepSeek (κw = 0.127, slight) performed substantially lower. Across all taxonomies, Miller κw values were uniformly the lowest, indicating that clinical applicability classification was more difficult for all LLMs than cognitive complexity classification.

Table 3
Weighted Cohen's kappa (κw) and Spearman's ρ by LLM and taxonomy.

Taxonomy	LLM	κw	95% CI	Interpretation	Spearman ρ	Significance
Bloom	**ChatGPT**	0.365	0.279–0.456	Fair	0.519	***
	Gemini	**0.676**	0.575–0.753	Substantial	0.650	***
	Kimi	0.428	0.313–0.523	Moderate	0.454	***
	DeepSeek	**0.667**	0.551–0.752	Substantial	0.637	***
SOLO	**ChatGPT**	0.433	0.334–0.531	Moderate	0.515	***
	Gemini	**0.622**	0.522–0.709	Substantial	0.616	***
	Kimi	0.549	0.434–0.648	Moderate	0.538	***
	DeepSeek	0.505	0.375–0.615	Moderate	0.538	***
Miller	**ChatGPT**	0.372	0.244–0.496	Fair	0.383	***
	Gemini	0.235	0.139–0.332	Fair	0.311	***
	Kimi	0.146	0.048–0.246	Slight	0.199	**
	DeepSeek	0.127	0.072–0.191	Slight	0.246	***

κw: weighted Cohen's kappa. 95% CI: bootstrap confidence interval (2,000 iterations). Interpretation thresholds (Landis & Koch, 1977): <0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial. Bold values indicate substantial agreement (κw ≥ 0.60). *** p < 0.001, ** p < 0.01.

All Spearman correlations were statistically significant (ρ = 0.199–0.650, all p < 0.01), confirming a consistent ordinal relationship between LLM and expert classifications across all conditions. The weakest correlation (Kimi–Miller: ρ = 0.199, p = 0.005) remained statistically significant.

Classification Performance

Overall accuracy and macro-averaged F1 scores are presented in Table 4a. For Bloom's Taxonomy, DeepSeek achieved the highest accuracy (63.5%) and macro F1 (0.411), followed by Gemini (accuracy:

59.0%, macro F1: 0.367). ChatGPT and Kimi performed poorly (accuracy: 37.5% and 42.5%; macro F1: 0.277 and 0.269, respectively). The consistent gap between accuracy and macro F1 across all LLMs reflects the dominance of lower-order items in the dataset - a pattern particularly pronounced for Kimi, whose high Level 1 recall (0.873) inflated accuracy while masking poor performance at higher levels.

Table 4
a. Classification performance metrics by LLM and taxonomy.

Taxonomy	LLM	Accuracy	Macro F1
Bloom	ChatGPT	0.375	0.277
	Gemini	0.590	0.367
	Kimi	0.425	0.269
	DeepSeek	0.635	0.411
SOLO	ChatGPT	0.450	0.458
	Gemini	0.660	0.654
	Kimi	0.605	0.605
	DeepSeek	0.665	0.621
Miller	ChatGPT	0.695	0.682
	Gemini	0.550	0.374
	Kimi	0.510	0.508
	DeepSeek	0.460	0.435

Accuracy: proportion of exact matches with gold standard. Macro F1: macro-averaged F1 across all observed levels, with equal weight per level regardless of frequency. See Table 4b for per-level F1 values.

For SOLO Taxonomy, Gemini achieved the highest macro F1 (0.654), with a uniquely balanced per-level F1 profile (Uni-structural: 0.681; Multi-structural: 0.660; Relational: 0.620); the only LLM to demonstrate consistent classification fidelity across all three levels (Table 4b). DeepSeek achieved the highest SOLO accuracy (66.5%) but with marked imbalance, driven by high Level 2 recall (0.855) at the expense of Level 4 (Relational) recall (0.333).

For the Miller Pyramid, ChatGPT achieved the highest accuracy (69.5%) and macro F1 (0.682), with the most balanced per-level performance (Level 1 F1: 0.616; Level 2 F1: 0.747). DeepSeek demonstrated near-total assignment to Level 1 (recall = 0.985), resulting in the lowest accuracy (46.0%) and severely imbalanced Level 2 F1 (0.316).

Table 4
b. Per-level F1 scores by LLM, taxonomy, and cognitive level (supplementary).

Taxonomy	Level	n	ChatGPT	Gemini	Kimi	DeepSeek
Bloom	L1 - Remember	63	0.655	0.677	0.595	0.712
	L2 - Understand	95	0.385	0.674	0.336	0.694
	L3 - Apply	31	0.240	0.293	0.261	0.121
	L4 - Analyse	6	0.106	0.194	0.154	0.000
	L5 - Evaluate	5	0.000	0.000	0.000	0.526
SOLO	L2 - Uni-structural	69	0.610	0.681	0.635	0.702
	L3 - Multi-structural	95	0.329	0.660	0.584	0.681
	L4 - Relational	36	0.437	0.620	0.595	0.480
Miller	L1 - Knows	68	0.616	0.586	0.542	0.554
	L2 - Knows How	132	0.747	0.536	0.473	0.316

F1: per-level F1 score computed as 2 × Precision × Recall / (Precision + Recall) for each cognitive level individually. Values of 0.000 indicate that the LLM assigned no items to that level (zero recall) or all assignments were incorrect (zero precision). n: number of gold standard items at that level. Levels 1 (Pre-structural) and 5 (Extended Abstract) for SOLO, and Levels 3 (Shows How) and 4 (Does) for Miller, were not assigned by the expert panel and are therefore not shown.

Confusion matrix patterns

Examination of confusion matrices revealed systematic directional tendencies. ChatGPT over-assigned Level 4 (Analyse) to 50 of 95 gold standard Level 2 items, inflating higher-order classification. Kimi showed the opposite tendency, assigning Level 1 to 55 of 95 Level 2 items. DeepSeek correctly identified all five Level 5 items, but over-assigned Level 5 to nine additional items rated at Levels 3–4. Gemini showed the most balanced misclassification profile across Bloom levels. For the Miller Pyramid, Gemini uniquely assigned Level 3 (Shows How) to eight items - all corresponding to gold standard Level 2 items with high cognitive demand (SOLO Level 4) - suggesting a broader interpretation of clinical applicability.

Multi-LLM Comparison

Cochran's Q tests revealed statistically significant differences in item-level accuracy among the four LLMs for all three taxonomies (Bloom: $\chi^2(3) = 57.623$, $p < 0.001$; SOLO: $\chi^2(3) = 33.151$, $p < 0.001$; Miller: $\chi^2(3) = 39.894$, $p < 0.001$). Kendall's W indicated moderate effect sizes across all three taxonomies (W = 0.312–0.395, moderate per the 0.30–0.49 interval), confirming meaningful but not strong concordance in item-level accuracy among LLMs (Table 5).

Table 5

Cochran's Q test results and Kendall's W effect sizes.

Taxonomy	Cochran's Q (df = 3)	p-value	Kendall's W	Effect Size	Post-hoc
Bloom	57.623	< 0.001	0.380	Moderate	See Table 6
SOLO	33.151	< 0.001	0.312	Moderate	See Table 6
Miller	39.894	< 0.001	0.395	Moderate	See Table 6

Cochran's Q test for k = 4 related binary samples (n = 200 items). Kendall's W computed as Q / [n(k–1)] serves as the effect size; interpretation: <0.10 negligible, 0.10–0.29 weak, 0.30–0.49 moderate, 0.50–0.69 strong, ≥ 0.70 very strong.

Pairwise McNemar tests with Holm–Bonferroni correction identified the following significant differences (Table 6).

For Bloom's Taxonomy, Gemini and DeepSeek both significantly outperformed ChatGPT (both $p < 0.001$) and Kimi (both $p < 0.001$), while Gemini and DeepSeek did not differ significantly from each other ($p = 0.378$). ChatGPT and Kimi did not differ significantly ($p = 0.378$). For SOLO Taxonomy, all three of Gemini, DeepSeek, and Kimi significantly outperformed ChatGPT ($p < 0.001$, $p < 0.001$, and $p = 0.002$, respectively); no significant differences were observed among Gemini, Kimi, and DeepSeek.

For the Miller Pyramid, the ranking reversed entirely: ChatGPT significantly outperformed all three other LLMs (vs. Kimi: $p < 0.001$; vs. DeepSeek: $p < 0.001$; vs. Gemini: $p = 0.003$). Gemini performed significantly better than DeepSeek ($p = 0.033$). Kimi and DeepSeek did not differ significantly from each other. This complete reversal - ChatGPT ranking last on Bloom and SOLO but first on Miller - represents the most notable inter-LLM finding of the study.

Table 6
Pairwise McNemar's test results (Holm–Bonferroni corrected).

Taxonomy	Comparison	b_{01}	b_{10}	χ^2	p (adjusted)	Sig.	Direction
Bloom	ChatGPT vs Gemini	55	12	27.597	< 0.001	***	Gemini > ChatGPT
	ChatGPT vs Kimi	34	24	1.724	0.378	ns	-
	ChatGPT vs DeepSeek	68	16	32.190	< 0.001	***	DeepSeek > ChatGPT
	Gemini vs Kimi	17	50	16.254	< 0.001	***	Gemini > Kimi
	Gemini vs DeepSeek	28	19	1.723	0.378	ns	-
	Kimi vs DeepSeek	57	15	24.500	< 0.001	***	DeepSeek > Kimi
SOLO	ChatGPT vs Gemini	65	23	20.045	< 0.001	***	Gemini > ChatGPT
	ChatGPT vs Kimi	56	25	11.864	0.002	**	Kimi > ChatGPT
	ChatGPT vs DeepSeek	69	26	19.463	< 0.001	***	DeepSeek > ChatGPT
	Gemini vs Kimi	24	35	2.051	0.304	ns	-
	Gemini vs DeepSeek	34	33	0.015	0.903	ns	-
	Kimi vs DeepSeek	30	18	3.000	0.250	ns	-
Miller	ChatGPT vs Gemini	23	52	11.213	0.003	**	ChatGPT > Gemini
	ChatGPT vs Kimi	20	57	17.779	< 0.001	***	ChatGPT > Kimi
	ChatGPT vs DeepSeek	19	66	25.988	< 0.001	***	ChatGPT > DeepSeek
	Gemini vs Kimi	20	28	1.333	0.248	ns	-
	Gemini vs DeepSeek	16	34	6.480	0.033	*	Gemini > DeepSeek
	Kimi vs DeepSeek	12	22	2.941	0.173	ns	-

b_{01}: number of items where the first listed LLM was incorrect and the second was correct. b_{10}: number of items where the first listed LLM was correct and the second was incorrect. χ^2: McNemar's chi-square statistic (df = 1), computed as $(b_{01} - b_{10})^2 / (b_{01} + b_{10})$, without continuity correction. p-values are Holm–Bonferroni adjusted across all 18 pairwise comparisons. Sig.: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant. Direction indicates the higher-performing LLM for significant pairs.

Discussion

This study examined whether four LLMs can reliably classify the cognitive complexity of Turkish DUS endodontic MCQs using three established frameworks - Revised Bloom's Taxonomy, SOLO Taxonomy, and Miller's Pyramid - against a three-layer expert consensus gold standard. Three findings emerged consistently. Classification fidelity was substantial for Bloom and SOLO in the best-performing models but did not exceed fair agreement for any model on Miller. Fidelity declined as cognitive level increased across all LLMs. Inter-model differences were statistically significant and framework dependent. Most strikingly, the model ranking was completely reversed between the Bloom/SOLO and Miller domains.

For Bloom's Taxonomy, Gemini (κw = 0.676) and DeepSeek (κw = 0.667) reached substantial agreement with the gold standard. Kimi (κw = 0.428) achieved moderate agreement, and ChatGPT (κw = 0.365) only fair agreement. For SOLO, Gemini again led (κw = 0.622, substantial); the remaining models reached moderate agreement (κw = 0.433–0.549). Lower-order Bloom levels (1 and 2) were classified with higher fidelity across all models. Classification accuracy declined at higher-order levels (3 to 5). This gradient mirrors a pattern observed across different LLM task paradigms: answering accuracy in medical MCQs declines at higher Bloom levels [10, 17], and alignment with higher-order Bloom levels is weaker in LLM-generated dental MCQs [18]. Bajčetić et al. [19] similarly demonstrated that LLM Bloom classifications were inconsistent when the same items were reclassified two months later, supporting the broader observation that LLM classification behaviour evolves over time. Together, these findings suggest that difficulty with higher-order cognitive content is a consistent LLM limitation independent of whether the task involves answering, generating, or classifying items.

Gemini was the only model to achieve consistent per-level F1 across all three SOLO levels (Uni-structural: 0.681; Multi-structural: 0.660; Relational: 0.620). DeepSeek achieved the highest SOLO accuracy (66.5%) but primarily through Multi-structural recall at the expense of Relational items. Uniform coverage across levels is more defensible for assessment auditing than level-specific dominance.

Confusion matrix analysis revealed distinct directional biases. ChatGPT overclassified 50 of 95 Bloom Level 2 items as Level 4; Kimi misclassified 55 of the same items as Level 1. Gemini's errors were distributed across adjacent levels without a consistent directional bias - a property that makes it more suitable for taxonomy-based item auditing. Gemini uniquely assigned Miller Level 3 (Shows How) to eight items, all of which the expert panel had classified at Level 2 (Knows How) with high SOLO-Relational demand. This finding may reflect the interpretive tension at the Knows How/Shows How boundary discussed in the Miller framework literature, where higher levels involve elements of observable performance and entrustment beyond declarative knowledge [15]. Although the expert panel classified Shows How as a performance-oriented level typically assessed through simulations or observed clinical tasks rather than MCQs, clinically integrative items may nevertheless invite this interpretation. Such discrepancies may still be informative during expert panel deliberations, even when they fall outside the final consensus classification.

ChatGPT ranked last for Bloom classification accuracy (37.5%) and last for SOLO in all pairwise comparisons yet achieved the highest Miller Pyramid classification accuracy (69.5%) and the only fair-agreement κ_w for that framework (0.372). The reversal was statistically significant in all three pairwise comparisons (vs. Kimi, DeepSeek, and Gemini; all $p \leq 0.003$ after Holm–Bonferroni correction). A mechanistic explanation lies in ChatGPT's confusion matrix pattern. ChatGPT systematically over-assigned Bloom Level 4 (Analyse) to items classified as Level 2 (Understand) by the expert panel - an upward inflation of perceived cognitive complexity. This same tendency may favour Miller accuracy: An item interpreted as requiring analytical integration is more likely to be classified as Knows How rather than Knows, which is precisely the distinction the expert panel made for 66% of items. Both the Bloom over-estimation and the Miller advantage may thus reflect a single underlying disposition to contextualize knowledge within applied reasoning, expressed differently across frameworks with different anchoring principles.

This finding demonstrates that Bloom, SOLO, and Miller capture different educational dimensions and therefore cannot be used interchangeably as proxies for cognitive demand. ten Cate et al. [15] underscored this distinction by explaining that Miller's framework concerns performance-based assessment, not cognitive complexity itself. An LLM that excels at complexity classification (Gemini, DeepSeek) may perform poorly on clinical applicability stratification, and vice versa. Taxonomy-specific LLM selection is therefore appropriate depending on the classification purpose.

Gemini and DeepSeek led both Bloom and SOLO classifications with significant margins over ChatGPT and Kimi. This convergence is consistent with the theoretical overlap between the two frameworks [13]. LLMs with stronger capacity for cognitive complexity discrimination perform better on both frameworks precisely because Bloom and SOLO measure the same underlying construct from different structural angles.

Miller taxonomy results showed a pattern that was completely independent from the Bloom and SOLO frameworks. Models that performed best in Bloom and SOLO did not perform best in Miller, while models with weaker Bloom and SOLO performance did not necessarily rank poorly in Miller. This divergence is theoretically plausible because Miller's distinction between Knows and Knows How reflects a shift from declarative knowledge to procedural and clinically applied knowledge, which does not correspond directly to Bloom or SOLO levels. For example, an item classified as SOLO Level 3 or Bloom Level 2 may still require clinical application consistent with Knows How, whereas a Bloom Level 4 item may remain within the Knows domain if the analysis is primarily theoretical. As a result, LLMs that are more sensitive to this distinction may perform differently across the Miller framework compared with the Bloom and SOLO frameworks.

A significant positive correlation was observed between examination year and expert-assigned Bloom level (Spearman $\rho = 0.301$, $p < 0.001$), indicating that DUS endodontic questions became progressively more cognitively demanding between 2012 and 2025. This trend is consistent with the broader adoption of taxonomy-based assessment frameworks designed to promote higher-order cognitive evaluation [9].

Additional support comes from studies questioning the validity of DUS as a competency assessment because of high LLM answer accuracy rates, as well as findings showing that endodontics was among the specialties with the lowest LLM performance in DUS examinations, which may reflect the greater cognitive complexity of questions in this field [20, 21]. This trend has a direct implication for future studies. Research sampling from earlier DUS items will overestimate LLM classification fidelity. Higher-order items (Bloom Levels 3 to 5) comprised only 21.0% of the current dataset but are increasing over time. LLM suitability for taxonomy classification should therefore be re-evaluated as examination content evolves.

Çekiç and Tavşan [7] evaluated eight LLMs using 151 Turkish DUS endodontic MCQs and identified ChatGPT-4o as the highest-performing model, with an answer accuracy of 81.5%. The present study extends this line of research in two important ways. First, the focus shifted from answer correctness to taxonomy classification. Although ChatGPT demonstrated high answer accuracy in the previous study, it showed the lowest Bloom classification fidelity among the four models evaluated here, indicating that successful item answering and accurate recognition of cognitive demand represent distinct capabilities. Second, the simultaneous application of Bloom, SOLO, and Miller frameworks revealed a complete reversal of model rankings within the Miller taxonomy, a finding that would not have been detectable using a single-framework or accuracy-only design. A related pilot study analysing DUS paediatric dentistry questions used Bloom's revised taxonomy to evaluate the cognitive structure of examination items and reported that questions were concentrated primarily at the Understand and Apply levels, with no items classified at the Evaluate or Create levels [22]. The study also found no clear temporal progression toward higher-order cognition and emphasized a potential mismatch between examination structure and advanced postgraduate competency expectations. Our present study extends these findings by moving beyond descriptive taxonomy mapping to examine LLM-based taxonomy classification performance across multiple educational frameworks. In contrast to the paediatric dentistry study, which relied exclusively on expert human classification within Bloom's taxonomy, the current study compared inter-model agreement patterns across Bloom, SOLO, and Miller frameworks and demonstrated framework-dependent reversals in model performance. Methodologically, the earlier study focused on cognitive distribution and curriculum alignment within a single specialty domain, whereas the present study evaluated comparative taxonomy recognition capabilities of multiple LLMs within endodontics and incorporated statistical analyses of inter-model variation and classification fidelity.

The 200-item dataset spanned 18 topic domains. Basic Sciences (15.0%), Instrumentation (9.5%), Pulpal and Periapical Diseases (9.0%), and Complications (9.0%) were the most represented. The predominance of Basic Sciences items - which characteristically target factual recall of microbiology, histology, pharmacology, and pathology - partly explains the 79.0% proportion of Bloom Levels 1 and 2 items in the gold standard. Digital Technologies (n = 2) and Biochemistry (n = 1) were too sparsely represented for subgroup analysis; future item banks may allow this.

There were several limitations in our study. First, the present study employed a single-session design rather than within-window replication, in which each annual batch would have been submitted twice to each LLM during the original February 2026 data collection period, permitting direct estimation of single-session reliability. This was not undertaken in the present study, which was designed as an exploratory comparative evaluation. To assess whether post-hoc replication could provide a substitute reliability estimate, we re-submitted the 2012 batch (20 items) to all four LLMs in a separate session and compared the classifications against the original outputs. Per-framework stability ranged from 60% to 85%, but the directions of change differed systematically across models; ChatGPT and Gemini drifting predominantly toward lower cognitive levels, DeepSeek toward higher levels, and Kimi showing a framework-specific downward collapse on the Miller scale. No model exhibited random or balanced drift. This pattern is incompatible with random session-level stochasticity and is consistent with independent vendor-side model updates between data collection windows. Post-hoc replication in consumer-interface LLM evaluation therefore confounds session variability with version drift and cannot serve as a substitute for planned within-window replication. The present findings should be interpreted as a snapshot of model behaviour during the February 2026 collection window (GPT-5.2 Instant, Gemini 3.1 Pro, Kimi K2.5, DeepSeek V3.2), and reliability estimation in this research context warrants explicit attention to within-window design at the time of data collection. Subsequent studies in this research programme will incorporate within-window replication as a planned design element.

Second, the present study employed a deliberative consensus design for gold standard establishment rather than an independent parallel rating design; the latter would have permitted estimation of inter-rater reliability in human taxonomy classification, a distinct methodological question that future work could address directly.

Third, generation parameters including temperature were platform-managed and could not be standardised. API-based access with fixed settings would improve reproducibility in future work.

Fourth, the study was conducted exclusively in Turkish, which was central to its ecological validity but limits generalisability. Whether LLM classification fidelity differs across languages cannot be assessed without a parallel English-language evaluation.

Finally, uneven topic representation precluded subgroup analyses for several clinically relevant domains. Future studies with larger, balanced item banks would allow systematic investigation of whether classification fidelity varies by content domain.

Conclusion

This study showed that LLMs can classify the cognitive complexity of DUS endodontic MCQs with meaningful fidelity for Bloom’s and SOLO taxonomies, particularly for Gemini and DeepSeek, but performance declined consistently at higher cognitive levels across all models. In contrast, agreement for Miller's Pyramid did not exceed fair for any model, and model rankings were completely reversed, with ChatGPT performing best on Miller despite ranking lowest on Bloom and SOLO. These findings

indicate that Bloom, SOLO, and Miller assess distinct constructs and should not be treated as interchangeable frameworks. Accordingly, LLM selection should be tailored to the specific taxonomy and assessment purpose.

A significant increase in item cognitive complexity from 2012 to 2025 suggests that DUS questions are becoming progressively more demanding, meaning that studies based on older item banks may overestimate current LLM classification performance.

Finally, the discrepancy between answer accuracy and taxonomy classification fidelity demonstrates that these represent separate capabilities. High performance in answering MCQs does not necessarily indicate that an LLM can accurately identify the cognitive demand of those items. Consequently, taxonomy classification fidelity should be evaluated independently when LLMs are used for assessment quality assurance or curriculum mapping.

Declarations

Ethics approval and consent to participate

This study analysed multiple-choice examination items and did not involve human participants, human tissue, or identifiable personal data. Ethics committee approval and informed consent were therefore not required, in accordance with the Declaration of Helsinki and the national regulations of Türkiye governing non-interventional research. The 200 endodontic items were obtained from DUSDATA under a written permission agreement for academic research use.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Use of large language models

Claude (Anthropic) was used solely for language editing and improvement of clarity in the manuscript text. The model was not used for the design of the study, data analysis, interpretation of results, generation of scientific content, or drafting of original scientific arguments. All content produced with language-editing assistance was critically reviewed, verified, and edited by the authors, who take full responsibility for the integrity, accuracy, and originality of the work.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

[BHŞ] and [OA] conceived and designed the study. [AY] and [JO] independently classified the 200 MCQs according to the topic coding scheme and resolved discordant items through structured panel discussion. [AY] and [DKA] also performed the expert classification of the items across the three taxonomic frameworks (Revised Bloom's Taxonomy, SOLO Taxonomy, and Miller's Pyramid) and established the gold standard through structured consensus deliberation. [BHŞ] administered the standardised classification prompt to the four large language models and compiled the resulting classification dataset. [BHŞ] performed the statistical analysis and interpreted the results. [BHŞ], [JO], and [OA] drafted the manuscript. [BHŞ], [AY], and [DKA] critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Acknowledgement

The authors gratefully thank DUSDATA (https://www.dusdata.com) for granting permission to access the archived Turkish Dentistry Specialty Examination items used in this study. The authors also thank Assoc. Prof. Aysel Başer (Department of Medical Education, İzmir Demokrasi University, İzmir, Türkiye) for delivering the calibration training, conducting the procedural audit, and serving as an independent procedural endorser during the consensus classification process, and Dr. Bilgecan Şen (Research Scientist, Department of Forest and Wildlife Ecology, University of Wisconsin–Madison, USA) for his guidance on the statistical analysis.The authors used PubMed, Consensus (https://consensus.app), and Elicit (https://elicit.com) as search and screening tools during the literature review phase of this study. All references included in the manuscript were independently verified and assessed by the authors.

Data Availability

The 200 endodontic multiple-choice questions analysed in this study were obtained from DUSDATA (https://www.dusdata.com), which holds copyright over the Turkish Dentistry Specialty Examination (DUS) item archive, under a written permission agreement granted to the authors for academic research use. The original questions cannot be redistributed by the authors due to these copyright restrictions; researchers wishing to access the items should contact DUSDATA directly. The expert gold standard classifications, the complete LLM classification outputs for ChatGPT, Gemini, Kimi, and DeepSeek, and the Python analysis scripts generated during the current study are available from the corresponding author on reasonable request. The standardised classification prompt applied to all four LLMs is provided in full in Supplementary Appendix I.

References

1. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT, in medical education: a systematic scoping review. J Educ Eval Health Prof. 2024;21:6. 10.3352/jeehp.2024.21.6.
2. Liu M, Okuhara T, Huang W, Ogihara A, Nagao HS, Okada H, et al. Large language models in dental licensing examinations: systematic review and meta-analysis. Int Dent J. 2025;75:213–22. 10.1016/j.identj.2024.10.014.
3. Chau RCW, Thu KM, Yu OY, Hsung RT, Lo ECM, Lam WYH. Performance of generative artificial intelligence in dental licensing examinations. Int Dent J. 2024;74:616–21. 10.1016/j.identj.2023.12.007.
4. de Araujo LP, Moreno LB, Chaves de Araujo BC, Chaves ET, Botero TM, Romero VHD. From evidence-based endodontics to generative AI: a comparative study of 11 large language models. J Endod. 2026. 10.1016/j.joen.2026.01.009.
5. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. BMC Oral Health. 2025;25:573. 10.1186/s12903-025-05926-2.
6. Kandemir M, Sarıbaş EE. Comparative performance of large language models in answering periodontology questions from the Turkish Dental Specialty Examination. BMC Oral Health. 2025;25:1804. 10.1186/s12903-025-07241-2.
7. Çekiç EC, Tavşan O. Evaluating large language models using national endodontic specialty examination questions: are they ready for real-world dentistry? BMC Med Educ. 2025;25:1308. 10.1186/s12909-025-07896-z.
8. Anderson LW, Krathwohl DR, editors. A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York: Longman; 2001.
9. 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;17:102444. 10.1016/j.cptl.2025.102444.
10. Herrmann-Werner A, Festl-Wietek T, Holderried F, Herschbach L, Griewatz J, Masters K, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. J Med Internet Res. 2024;26:e52113. 10.2196/52113.
11. Law AK, So J, Lui CT, Choi YF, Cheung KH, Kei-Ching Hung K, et al. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med Educ. 2025;25:208. 10.1186/s12909-025-06796-6.
12. Biggs JB, Collis KF. Evaluating the quality of learning: the SOLO taxonomy (Structure of the Observed Learning Outcome). New York: Academic; 1982. 10.1016/C2013-0-10375-3.
13. Sirri D, Harper A, Brown GTL. Revised Bloom's Taxonomy and the Structure of the Observed Learning Outcome taxonomy for classifying multiple-choice questions: a study using the bookmark

procedure. Asia Pac Sci Educ. 2025;11:266–94. 10.1163/23641177-bja10101.
14. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65(9 Suppl):S63–7. 10.1097/00001888-199009000-00045.
15. ten Cate O, Carraccio C, Damodaran A, Gofton W, Hamstra SJ, Hart DE, et al. Entrustment decision making: extending Miller's pyramid. Acad Med. 2021;96:199–204. 10.1097/ACM.0000000000003800.
16. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74. 10.2307/2529310.
17. Crichton R, Liu B, Hothi S. Large language model performance in clinical cardiology multiple choice questions: has reasoning improved performance? Eur Heart J Digit Health. 2026;7(Suppl1). ztaf143.011.
18. 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:e0346800. 10.1371/journal.pone.0346800.
19. Bajčetić M, Mirčić A, Rakočević J, Đoković D, Milutinović K, Zaletel I. Comparing the performance of artificial intelligence learning models to medical students in solving histology and embryology multiple choice questions. Ann Anat. 2024;254:152261. 10.1016/j.aanat.2024.152261.
20. Kinikoglu I. Evaluating ChatGPT and Google Gemini performance and implications in Turkish dental education. Cureus. 2025;17:e77292. 10.7759/cureus.77292.
21. Şişmanoğlu S, Şirinoğlu Capan B. 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;25:214. 10.1186/s12909-024-06389-9.
22. Çerçi Akçay H, Üstün N. Cognitive alignment and assessment validity in a high-stakes dental specialty examination: a Bloom's taxonomy–based analysis. BMC Med Educ. 2026. 10.1186/s12909-026-09338-w.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryAppendix1.docx