BMC Medical Education

RESEARCH Open Access

Consistency over accuracy: run-to-run stability of contemporary large language models on Turkish curriculum-aligned theoretical anatomy multiple-choice questions

Ömer Alperen Gürses[1*] and İsmail Ceylan[1]

Abstract

Background Stability across repeated administrations is essential for educational use of large language models (LLMs), yet it is rarely quantified in non-English, curriculum-aligned anatomy contexts.

Methods Eleven contemporary LLMs answered 100 Turkish, faculty-authored, curriculum-aligned anatomy multiple-choice questions from AYDEP, targeting the undergraduate Physiotherapy and Rehabilitation anatomy curriculum in three independent runs (≥ 12-hour intervals). Testing used developers' web interfaces in Turkey with browsing disabled and default generation settings (August–September 2025). Performance was summarized with a stability-aware 0–3 item score (number of correct responses across three runs) and predefined response-consistency classes.

Results A subset of models achieved near-ceiling totals with high run-to-run stability, whereas others showed greater session-to-session variability (i.e., changes in the selected option across independent runs initiated as separate sessions under identical inputs). Several nominal differences among higher-performing systems did not remain significant after multiplicity control. Within-family updates produced selective, not universal, gains. Many models exhibited medians of 3 (IQR 3–3) on the 0–3 scale (ceiling effects), and lower means were accompanied by larger dispersion. Consistency profiles provided information beyond mean accuracy by distinguishing reliably correct from volatile behavior. In addition, we observed "consistent & wrong" patterns on a subset of items, where the same incorrect option was repeatedly selected across runs.

Conclusion In Turkish, curriculum-aligned anatomy items, contemporary LLMs can be both accurate and stable, but single-trial accuracy can mask volatility and stable systematic errors. Adoption decisions should prioritize stability-aware appraisal (including consistent-correct and consistent-wrong rates), with local validation on institutional item banks and periodic re-evaluation as models evolve. Extending this framework to multimodal anatomy and constructed-response tasks will further inform trustworthy, learner-facing use.

Keywords Large language models, ChatGPT, Gemini, DeepSeek, Grok, Claude, Anatomy examination

*Correspondence:
Ömer Alperen Gürses
omeralperengurses@gmail.com
[1]School of Physical Therapy and Rehabilitation, Department of Physiotherapy and Rehabilitation, Kırşehir Ahi Evran University, 40100 Merkez, Kırşehir, Türkiye

Background

Rapid advances in artificial intelligence—especially in machine learning and natural language processing—have enabled large language models (LLMs) that interpret and generate text with high proficiency across domains [1, 2]. In medical education, early reports that ChatGPT could approach the passing threshold on United States Medical Licensing Examination (USMLE) items sparked broad interest in LLM-supported learning and assessment [3, 4]. Yet single-shot correctness is an insufficient basis for educational adoption; long-standing standards emphasize the need for reliability, validity, and clearly articulated limitations [5].

In this paper, we treat run-to-run "consistency" as a stability construct that is related to—but distinct from—accuracy, reliability, and robustness. Accuracy concerns whether an answer is correct in a single attempt. By contrast, run-to-run consistency captures whether the model produces the same answer when the identical item is re-administered in separate runs under comparable conditions, acknowledging that LLM outputs can remain non-identical across repeated inference due to the inherent stochasticity of large language models [6]. This notion overlaps with reliability in the broad sense of repeatability, yet it is not a psychometric reliability estimate of a test score; rather, it is an item-level stability indicator for generative systems. Consistency also differs from robustness, which typically refers to maintaining performance under systematic input perturbations (e.g., paraphrasing or noise) [7]. Finally, reproducibility concerns whether independent researchers can obtain similar results using transparent procedures and reporting practices, whereas consistency is a property of the model's repeated outputs within a given setup [8].

These requirements are particularly salient for anatomy, where dense terminology, spatial reasoning, and tight alignment to course objectives challenge generalization from generic benchmarks [9]. Recent gross-anatomy work in English settings shows promise but also variability across repeated trials, cautioning against viewing LLMs as instructor replacements [10]. Recent anatomy education evaluations have likewise reported strong single-run performance for contemporary LLMs, reinforcing the need to complement accuracy-focused evidence with stability-oriented evaluation when considering educational deployment [11, 12].

Notably, an anatomy-focused benchmark using USMLE-style gross-anatomy MCQs administered multiple attempts per model reported high overall accuracy for newer systems but persistent topic-level heterogeneity, illustrating that single-score summaries can obscure reliability-relevant variation [11]. Similarly, a cross-platform embryology MCQ evaluation reported both accuracy and test–retest reliability (e.g., ICC) across repeated attempts, reinforcing that repeatability can differ meaningfully across model families even when average accuracy is high [12].

Complementing these reports, a recent anatomy study comparing Gemini 2.5 Flash, Gemini 2.0, DeepSeek V3, and ChatGPT-4o on 55 MCQs found near-ceiling overall accuracy (~96%) with no significant pairwise differences, while identifying 'General Anatomy' as comparatively more challenging—yet the evaluation relied on single attempts and text-only items [13]. As the model landscape evolves rapidly across families (e.g., ChatGPT, Gemini, Claude, Grok, DeepSeek), reasoning-focused systems have shown strong performance on board-style questions in specialty contexts (e.g., ophthalmology) while exhibiting sensitivity to task modality and cognitive complexity [14]. Nevertheless, direct evidence of within-family, version-to-version gains remains limited—particularly in non-English, curriculum-aligned settings. Outside English, national licensing studies in Japanese and Polish contexts report within-family, version-to-version improvements (e.g., GPT-4 over GPT-3.5) yet underscore that gains are context-contingent and require local appraisal [15, 16]. Importantly, because our analysis is limited to Turkish, curriculum-aligned, text-only theoretical anatomy MCQs, the observed run-to-run stability patterns should not be assumed to generalize directly to more reasoning-intensive domains (e.g., physiology, pathology, or clinical reasoning) or to multimodal anatomy assessments.

The present study addresses these needs by benchmarking contemporary LLM families on a Turkish-language, multiple-choice theoretical anatomy examination aligned with a competency-based digital platform used in routine instruction (AYDEP) and authored by faculty to reflect actual curricular objectives [17]. Beyond overall performance, we explicitly evaluate run-to-run stability and within-family generational change in an authentic, non-English educational setting, aligning appraisal with established educational-measurement principles [5]. In doing so, we address limitations noted in recent anatomy evaluations by moving beyond single-trial accuracy through independent repeated runs that quantify variance across administrations, and by employing a larger, curriculum-aligned item bank [13, 14].

We therefore ask: (i) Do models differ in accuracy and run-to-run stability on Turkish theoretical anatomy items? (ii) Do newer releases within a family deliver reliable improvements? (iii) Are the effects statistically robust after multiple-comparison control? By answering these, we provide decision-ready evidence for formative assessment, feedback, model selection, and item development in medical education. In learner-facing use, output instability can introduce "feedback noise" that confuses

self-testing, whereas stable outputs support dependable feedback and item-bank vetting [18].

Methods

Study setting and data source

This study was conducted in the Faculty of Physiotherapy and Rehabilitation at Kırşehir Ahi Evran University and used multiple-choice theoretical anatomy items administered through the university's competency-based digital platform "AYDEP" (Ahi Yeterlilik Tabanlı Dijital Eğitim Platformu, Ahi Competency-Based Digital Education Platform) between 2022 and 2024. AYDEP is the institution's official online environment for teaching and assessment; it links course learning objectives with item content and aggregates student response statistics that are used operationally to classify item difficulty. The items analysed here were authored and reviewed by faculty responsible for the anatomy curriculum [17].

Item selection and difficulty banding

A pool of 320 AYDEP theoretical anatomy items (2022–2024) was screened. In routine practice, AYDEP labels items into four difficulty bands using historical student correct-response statistics ("very difficult", "acceptable", "recommendable", "very easy"); however, fixed numerical cutoffs for these labels were not available for extraction in our dataset/materials, so we retained AYDEP's operational labels as provided. Within the screened pool, ($n = 94$) items were labeled "acceptable" and ($n = 107$) were labeled "recommendable"; the remaining items fell into the "very difficult" or "very easy" bands and were not considered further. For this study, we used stratified random sampling to select 100 items from the "acceptable" and "recommendable" bands with an equal allocation (50 "acceptable" and 50 "recommendable") to balance challenge while avoiding ceiling/floor effects. The "very easy" and "very difficult" bands were excluded a priori because extreme difficulty levels can inflate ceiling/floor effects and reduce discrimination when quantifying run-to-run stability on an ordinal 0–3 framework. All items were theoretical, written in Turkish, and required no interpretation of images or figures. This restriction was applied to standardize the evaluation across models and interfaces and to isolate text-based question answering without confounding effects from multimodal (image) capabilities. The sampled items covered a broad range of gross anatomy domains, including upper and lower limb musculoskeletal anatomy, neuroanatomy and cranial nerves, thoracic and abdominal organs, pelvis and urogenital structures, lymphoid organs, and basic topographical and anatomical terminology. Sample size of 100 items was selected to balance computational feasibility with statistical power for detecting meaningful differences in model performance while ensuring stable estimates across repeated runs, consistent with recommendations for comparative educational assessments [19].

Table 1 Models and access/version details

Model	Developer	Version date
ChatGPT-4o	OpenAI	May 2024
ChatGPT-5	OpenAI	Aug 2025
ChatGPT-5 Thinking	OpenAI	Aug 2025
Gemini 2.5 Flash	Google	June 2025
Gemini 2.5 Pro	Google	June 2025
DeepSeek-R1	DeepSeek	May 2025
DeepSeek-V3.1	DeepSeek	Aug 2025
Grok-3	xAI	Feb 2025
Grok-4	xAI	July 2025
Claude Sonnet 4	Anthropic	May 2025
Claude Opus 4.1	Anthropic	Aug 2025

Expert validation of cross-model failure items

To examine whether items that multiple high-performing LLMs answered incorrectly reflected flawed answer keys or genuinely challenging content, we conducted a targeted expert review based on cross-model failure. For each item, we inspected first-run responses across all 11 models and flagged questions for audit if at least 5 models answered them incorrectly. This threshold (≥ 5 models incorrect) was used to enrich the expert-audit set for potentially ambiguous or high-risk items while keeping faculty workload feasible; it was intended as a targeted quality check rather than a representative estimate of overall item quality. This criterion yielded 7 items from the 100-item pool, which are labelled CF1–CF7 in Supplementary Table S1. Two board-certified anatomy faculty members who were not involved in the original item writing independently answered this subset under exam-like conditions, blinded to the official keys and to all LLM outputs. Experts were not provided with any model-level results (i.e., which models missed each item) or performance summaries. Disagreements were not forced to consensus; instead, inter-rater agreement was quantified (Cohen's κ) and both experts' responses were retained for audit. For each item, we recorded whether each expert selected the keyed correct option and calculated the proportion of items answered correctly by both experts, as well as Cohen's κ with 95% confidence intervals to quantify inter-rater agreement.

Language models and testing protocol

We evaluated eleven models spanning five families (OpenAI ChatGPT, Google Gemini, DeepSeek, xAI Grok, Anthropic Claude). Full model version details are provided in Table 1. Each model answered the complete 100-item set three times in independent sessions separated by at least 12 h to assess both accuracy and run-to-run stability. Within each administration, the entire 100-item

list was provided as a single exam-like batch rather than one-by-one, mirroring how students access AYDEP-based assessments. To minimize context carry-over and ensure independence: (1) a new conversation was initiated for each 100-item administration; (2) web browsing and search features were disabled; (3) default sampling parameters were maintained; (4) no system-level prompts or role instructions were provided; (5) items were presented with stem and options only; (6) no feedback on correctness was given during testing; (7) persistent memory features were disabled where applicable. Sampling parameters (e.g., temperature/top-p) were not user-configurable in the developer managed web interfaces; therefore, all runs used platform defaults. We intentionally did not apply prompt engineering or ensembling (e.g., persona prompting, output-format constraints, or majority voting) to benchmark baseline, exam-like performance under comparable conditions. All model administrations were conducted between August 2025 and September 2025 via the developers' official web interfaces in Turkey, with browsing disabled and default sampling.

Scoring of model performance (stability-aware 0–3 scale)

The primary outcome was the mean 0–3 item score per model across the 100 items and three runs. Secondary outcomes were (i) total accuracy (% correct) and (ii) run-to-run response consistency, summarised via the four pattern classes described below. For each item, a model's three responses were compared with the answer key and assigned an ordinal stability-aware score: 3 if all three runs were correct; 2 if exactly two runs were correct; 1 if exactly one run was correct; 0 if none were correct. Conceptually, this 0–3 score is simply the count of correct responses across three independent administrations (a transparent ordinal summary of repeated correctness) rather than a psychometrically validated scale; its purpose is to jointly reflect correctness and run-to-run stability at the item level. Per-model totals therefore ranged from 0 to 300, and means (± SD) were computed across the 100 items. This metric summarizes repeated correctness; repeatability (including consistently wrong outputs) is captured by the predefined response-pattern classes.

Run-to-run response consistency (secondary outcome)

Independent of the 0–3 score, response patterns across runs were classified into four mutually exclusive categories. These four categories were predefined a priori to provide an interpretable taxonomy of stability profiles under repeated administrations. Consistent & correct (all three runs identical and correct), Consistent & wrong (all three identical but wrong), Partially consistent (exactly two runs identical; the third differs), and Inconsistent (three different options). Counts per model were summarised to characterise reliability profiles.

Statistical analysis

Analyses were conducted in SPSS (v25; IBM). Because item scores (0–3) are ordinal, all between-model inference used non-parametric methods. Item-level scores were summarised descriptively as mean ± SD and total (0–300). Overall differences were tested with Kruskal–Wallis H; upon significance, two-sided Mann–Whitney U tests were used for pairwise contrasts with Bonferroni family-wise error control ($\alpha = 0.05$). Bonferroni-adjusted significant contrasts are reported in Table 3; Holm-adjusted p-values were also computed for completeness. Effect sizes were reported as epsilon-squared (ε^2) for Kruskal–Wallis tests and as r for Mann–Whitney U tests ($r = |Z|/\sqrt{N}$). Consistency-class comparisons were evaluated with χ^2 tests of independence (or Fisher's exact test when any expected cell count < 5).

Quality assurance and controls

To reduce information leakage and prompt bias, we used uniform prompts, reset sessions between complete administrations, disabled memory features where available, and withheld correctness feedback during testing. All items were Turkish text without images. The final dataset, scoring procedure, and analysis specifications are available upon reasonable request.

Results

Across 100 items, fully correct counts (3/3) were: ChatGPT-5 Thinking 98/100, Gemini 2.5 Pro 97/100, Claude Opus 4.1 93/100, Gemini 2.5 Flash 93/100, DeepSeek-V3.1 92/100, Claude Sonnet 4 91/100, Grok-3 89/100, ChatGPT-4o 85/100, DeepSeek-R1 82/100, Grok-4 68/100, and ChatGPT-5 30/100 (Fig. 1). Corresponding stability-aware 0–3 scores (mean ± SD, total 0–300) are reported in Table 2.

Descriptive statistics for the stability-aware 0–3 item scores (mean ± SD and totals on a 0–300 scale) are summarised for each model. The highest averages were observed for ChatGPT-5 Thinking and Gemini 2.5 Pro, with Claude Opus 4.1 and Gemini 2.5 Flash close behind, whereas ChatGPT-5 had the lowest overall performance. Many models yielded a median of 3 (IQR 3–3) on the stability-aware 0–3 scale, indicating ceiling effects that compress apparent differences among top performers once multiplicity is controlled (shown in Table 2).

Overall differences were significant (Kruskal–Wallis $H(10) = 274.35$, $p < 0.001$, $\varepsilon^2 = 0.243$). Bonferroni-adjusted, two-sided Mann–Whitney U results for significant contrasts only are listed in Table 3; Holm-adjusted p-values were also computed for completeness. Effect sizes suggested that the most practically meaningful separations involved ChatGPT-5 versus all other models (large r values) and Grok-4 versus the leading tier (moderate-to-large r values). In contrast, several statistically significant

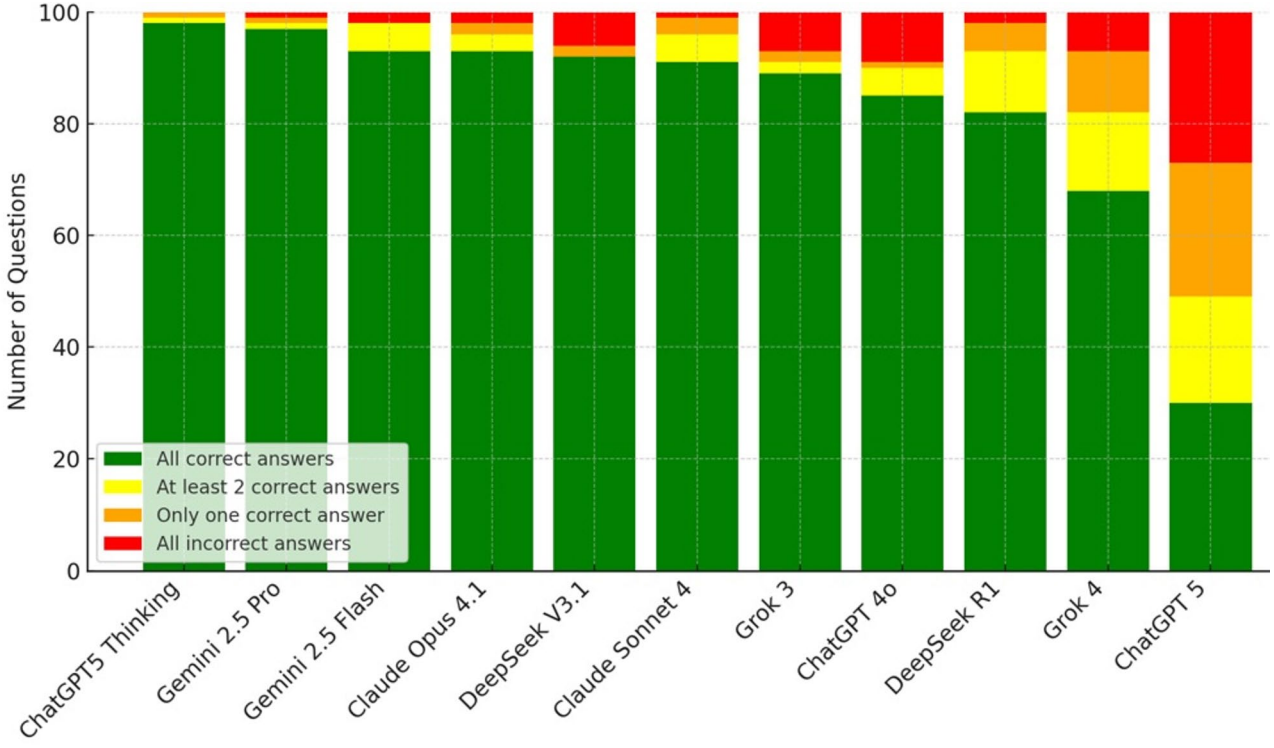

Fig. 1 Distribution of response accuracy by language model

Table 2 Descriptive statistics for stability-aware item scores (0–3) by model

Model	Mean ± SD	Median (IQR)	*N*	Total Score (0–300) for each model
ChatGPT-5 Thinking	2.97 ± 0.22	3 (3–3)	100	297
Gemini 2.5 Pro	2.94 ± 0.37	3 (3–3)	100	294
Gemini 2.5 Flash	2.89 ± 0.47	3 (3–3)	100	289
Claude Opus 4.1	2.87 ± 0.53	3 (3–3)	100	287
Claude Sonnet 4	2.84 ± 0.51	3 (3–3)	100	286
DeepSeek-V3.1	2.78 ± 0.76	3 (3–3)	100	278
DeepSeek-R1	2.73 ± 0.66	3 (3–3)	100	273
Grok-3	2.73 ± 0.81	3 (3–3)	100	273
ChatGPT-4o	2.57 ± 1.00	3 (3–3)	100	266
Grok-4	2.43 ± 0.95	3 (2–3)	100	243
ChatGPT-5	1.49 ± 1.18	1 (0–3)	100	152

Note: Unit of analysis = item ($n = 100$ per model across three runs). Mean ± SD and Total (0–300) summarise the ordinal 0–3 scores descriptively; between-model inference uses Kruskal–Wallis and Bonferroni-adjusted two-sided Mann–Whitney U tests.

contrasts among otherwise high-performing models showed small effects (e.g., $r \approx 0.22$–0.26), indicating limited practical educational relevance despite statistical significance.

Run-to-run response consistency mirrored the accuracy ranking. ChatGPT-5 Thinking and Gemini 2.5 Pro concentrated most items in the "Consistent & correct" class, indicating high accuracy with high reliability.

Table 3 Pairwise differences between Language models on stability-aware 0–3 scores (Bonferroni-adjusted p-values; significant contrasts only)

Comparison	Bonferroni-adjusted *p*	Effect size (*r*)
ChatGPT-5 vs. Gemini 2.5 Pro	< 0.001	0.66
ChatGPT-5 vs. ChatGPT-5 Thinking	< 0.001	0.61
ChatGPT-5 vs. Gemini 2.5 Flash	< 0.001	0.62
ChatGPT-5 vs. Claude Opus 4.1	< 0.001	0.67
ChatGPT-5 vs. Claude Sonnet 4	< 0.001	0.65
ChatGPT-5 vs. DeepSeek-V3.1	< 0.001	0.64
ChatGPT-5 vs. Grok-3	< 0.001	0.71
ChatGPT-5 vs. DeepSeek-R1	< 0.001	0.57
ChatGPT-5 vs. ChatGPT-4o	< 0.001	0.56
ChatGPT-5 vs. Grok-4	< 0.001	0.71
Gemini 2.5 Pro vs. Grok-4	< 0.001	0.68
ChatGPT-5 Thinking vs. Grok-4	< 0.001	0.67
Gemini 2.5 Flash vs. Grok-4	< 0.001	0.62
Grok-4 vs. Claude Opus 4.1	< 0.001	0.65
DeepSeek-V3.1 vs. Grok-4	0.004	0.38
Grok-4 vs. Claude Sonnet 4	0.008	0.35
ChatGPT-4o vs. Gemini 2.5 Pro	0.020	0.26
Grok-4 vs. Grok-3	0.035	0.25
Gemini 2.5 Pro vs. DeepSeek-R1	0.043	0.23
ChatGPT-4o vs. ChatGPT-5 Thinking	0.043	0.22

Note: Only contrasts that remain significant after Bonferroni adjustment are shown

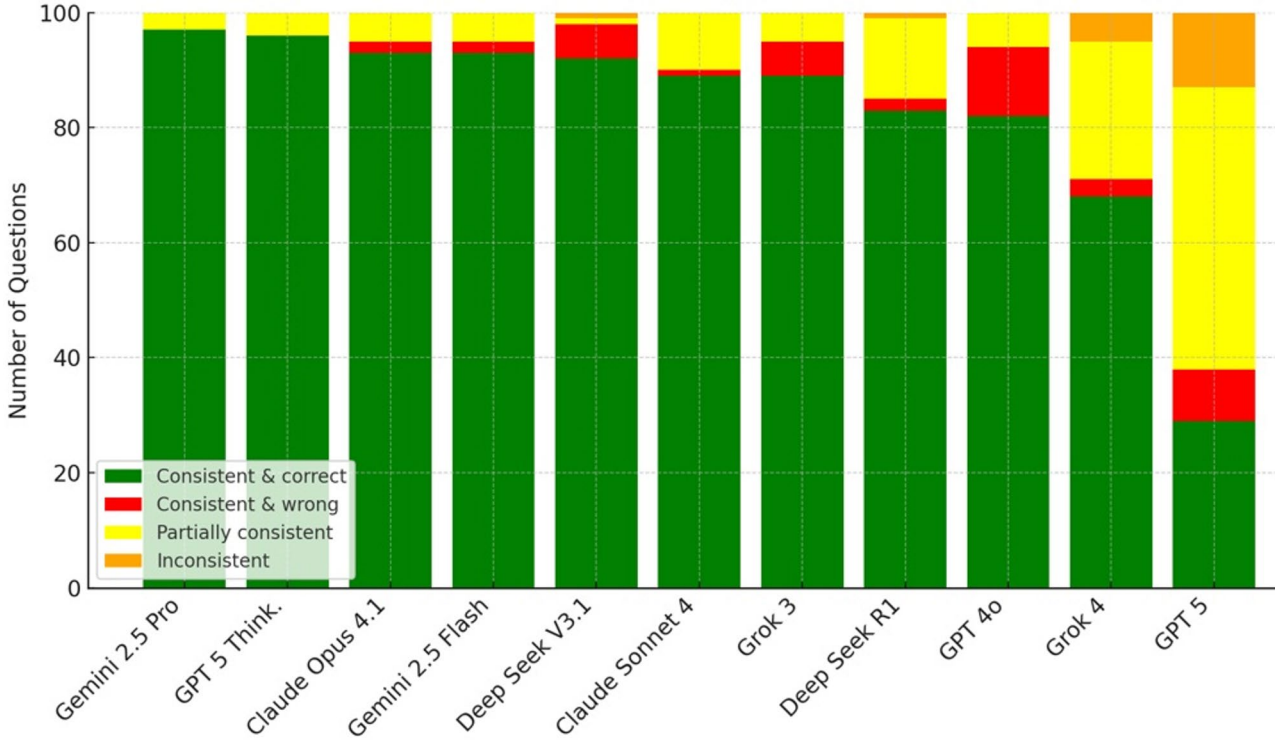

Fig. 2 Run-to-run response consistency by language model

Table 4 Summary of response-consistency classes by model

Model	Consistent & correct	Consistent & wrong	Partially consistent	Inconsistent
ChatGPT-5 Thinking	98	0	2	0
Gemini 2.5 Pro	97	0	3	0
Gemini 2.5 Flash	93	2	5	0
Claude Opus 4.1	93	2	5	0
DeepSeek-V3.1	92	6	1	1
Grok-3	89	6	5	0
Claude Sonnet 4	91	1	8	0
DeepSeek-R1	83	2	14	1
ChatGPT-4o	85	9	6	0
Grok-4	68	3	24	5
ChatGPT-5	29	9	49	13

Claude Opus 4.1 and Claude Sonnet 4 showed similar profiles with a small tail of "Consistent & wrong" items, compatible with rare systematic mis-selections. DeepSeek-V3.1, DeepSeek-R1, ChatGPT-4o, and Grok-3 displayed more "Partially consistent" responses, reflecting some session-to-session variability (i.e., changing selected options across runs initiated as separate sessions). Grok-4 and especially ChatGPT-5 combined lower accuracy with higher rates of inconsistent or consistently wrong responses (shown in Fig. 2; Table 4).

Item-level analysis of "Consistent & wrong" responses: Across the 100 items, 23/100 (23%) were answered consistently incorrectly by at least one model (i.e., the same wrong option in all three runs). Of these, 12 items were shared by ≥2 models, while 11 items were unique to a single model. The largest cross-model clusters were Item 6 (5 models), Item 98 (4 models), Item 7 (4 models), and Item 5 (3 models). All seven cross-model failure items (≥5 models incorrect in Run 1; Items 5, 6, 7, 40, 89, 97, and 98) were included among these shared, consistently wrong patterns. The full item-level listing (item number, answer key, and the models/options exhibiting "Consistent & wrong" behaviour) is provided in Supplementary Table S2.

Expert validation of cross-model failure items

Among the 100 items, 7 met the cross-model failure criterion defined in the Methods section. Two independent anatomy faculty selected the keyed option for all 7 items (100%), suggesting no evidence of miskeying within this targeted subset (Supplementary Table S1). Inter-rater agreement for this subset was perfect ($\kappa = 1.00$; 95% CI 1.00–1.00). This targeted audit was not intended as a representative validation of overall item quality across the full bank.

Exploratory order-effect check

To assess potential order/context-window effects from the single-prompt administration, we compared the first 25 versus last 25 items. Early–late differences were small

for most models (median |Δ accuracy| = 2.7% points), whereas ChatGPT-5 (84.0% → 29.3%) and Grok-4 (89.3% → 54.7%) showed marked declines with reduced stability toward the end of the prompt.

Discussion

Our principal findings are fourfold. First, on Turkish, curriculum-aligned theoretical anatomy MCQs, a small subset of systems—most notably ChatGPT-5 Thinking and Gemini 2.5 Pro—achieved near-ceiling totals with very high run-to-run reproducibility, whereas several mid- and lower-tier models (e.g., ChatGPT-4o, DeepSeek-R1, Grok-4, and the base ChatGPT-5) showed pronounced session-to-session variability. Second, several nominal pairwise differences among leading systems did not remain significant after family-wise error control, indicating broadly comparable performance once multiplicity is addressed. Third, within-family (older→newer) contrasts yielded selective, not universal, improvements—underscoring that recency alone does not guarantee educationally meaningful gains. Fourth, response-pattern profiles (e.g., consistent-correct vs. inconsistent) provided actionable information beyond mean accuracy by distinguishing reliably correct from noisily variable performers. Mid-tier systems with acceptable means but wide dispersion illustrate that mean accuracy can mask run-to-run volatility, reinforcing the value of a stability-aware metric for learner-facing use.

These results align with and extend prior work in medical education. Early classroom and domain reports emphasized that LLMs can assist instruction but should not be viewed as instructor replacements, in part because of trial-to-trial variability and domain limits [10, 20]. More recently, a small, text-only comparison in anatomy (55 MCQs) reported near-ceiling averages among leading models (e.g., ChatGPT-4o, DeepSeek V3, Gemini 2.0/2.5 Flash) with few or no significant pairwise differences and noted topic-level heterogeneity (e.g., "General Anatomy" harder), yet relied on single-attempt administration and excluded visual content—design choices that can mask reliability differences [13]. By adopting repeated, independent runs on a larger, faculty-authored, curriculum-aligned bank, our study makes those stability gaps explicit in a non-English context. A similar multi-run design has been recommended and applied in specialty board-review settings, where three-pass evaluations exposed volatility that single trials missed [21].

Specialty evaluations of reasoning-oriented systems likewise suggest rapid gains alongside task sensitivity. In ophthalmology, frontier models—including DeepSeek-R1, Grok-3/4, Gemini, and OpenAI's reasoning variants—have achieved strong performance on board-style problems, but spreads widen with increased cognitive complexity and modality shifts (text→image) [14, 22]. This echoes our observation that apparent ties at high average accuracy can separate on stability once evaluation emphasizes reproducibility rather than single-shot correctness. In basic sciences closest to our task, recent histology work found uniformly high accuracies (> 90%) with limited separation among systems, reinforcing the possibility of ceiling effects in terminology-dense, text-only settings and the value of stability-aware appraisal to reveal residual differences of practical importance [10]. When accuracy approaches ceiling, the 0–3 stability-aware score can show reduced discriminative power because fewer items remain sensitive to differences in both correctness and run-to-run variability. Future evaluations could mitigate this by broadening item difficulty and cognitive demand (and/or increasing repetitions) to better resolve stability differences among high-performing systems.

Evidence from national licensing examinations outside English further supports selective—rather than universal—version gains. In Japan and Poland, GPT-4 outperformed GPT-3.5 on medical licensing content, yet effect sizes and pass-rate margins varied by exam design and language, underscoring context dependence [15, 16]. Broader syntheses across countries similarly document that performance depends on model class, item type, and local curricular alignment [23, 24]. Against that backdrop, our within-family analyses confirm that gains are selective: educators should verify improvements on local, curriculum-matched items rather than assume benefit from recency alone.

Educationally, treating stability as a first-order criterion matters for how these tools are used. High "consistent-correct" rates reduce feedback noise in self-testing, support dependable item vetting, and make learning analytics more trustworthy, whereas models with volatile response patterns (e.g., the base ChatGPT-5 or some Grok configurations) risk confusing learners despite attractive mean accuracies. Equally important, the "Consistent & wrong" class is pedagogically critical: stable yet incorrect behaviour can degrade feedback more than random inconsistency, warranting targeted item review and cautious deployment. To support practical mitigation, we provide item-level "Consistent & wrong" clusters (item number, key, and the specific models/options) in Supplementary Table S2; the cross-model failure subset spans multiple anatomical systems (Supplementary Table S1). In practice, this risk can be mitigated by (i) running a small number of repeat queries for high-stakes learner-facing outputs and flagging items that remain consistently incorrect, (ii) cross-checking answers against authoritative resources (course notes, textbooks, or guideline-style references) and requiring citations, and (iii) using human-in-the-loop review for items intended for assessment or feedback. Institutions can also deploy

stability-aware dashboards that highlight "consistent & wrong" items for rapid faculty audit and item-bank correction. Concretely, this framework supports low-stakes self-study tools and low-stakes quizzes (where consistent-correct behaviour reduces confusing feedback), as well as faculty-led item-bank screening and explanation drafting before classroom use. For summative or other high-stakes assessment contexts, outputs should remain assistive only and be used under strict human oversight rather than as an autonomous answer source. This "adjunct rather than replacement" stance—already urged in early anatomy reports—thus accords with emerging proposals to embed LLMs as scaffolded assistants (e.g., domain-tuned tools such as AnatomyGPT) while maintaining expert oversight [20, 25]. This emphasis on reproducibility and qualified use is consistent with long-standing guidance that reliability is a prerequisite for validity in educational measurement [18]. In multilingual programs, our Turkish results provide a local appraisal complementing global exam studies: high proficiency is attainable where terminology is conserved and items are tightly aligned to course objectives, but model-specific vetting remains essential [23, 24].

In sum, leading systems such as ChatGPT-5 Thinking and Gemini 2.5 Pro now deliver not only high accuracy but—crucially for education—high stability on Turkish theoretical anatomy MCQs. Because several competitive models (e.g., ChatGPT-4o, DeepSeek-R1, Grok-4) can mask inconsistency behind good averages, stability-aware appraisal offers a more practical basis for adoption decisions. As curricula, platforms, and models evolve, coupling local benchmarking with reproducibility metrics is likely to remain the clearest path to trustworthy, learner-facing use in medical education.

Limitations and future work

This study has several limitations. The evaluation was restricted to a specific set of Turkish-language, curriculum-aligned text-only theoretical anatomy multiple-choice questions across the tested LLM platforms, which limits the generalizability of findings to other linguistic contexts and question formats, particularly image-based or clinical reasoning assessments. Generalizability may also be constrained by language- and curriculum-specific factors (e.g., Turkish terminology usage, local item-writing conventions, and AYDEP-aligned learning objectives) and by the knowledge-oriented nature of theoretical anatomy MCQs. Replication across other languages, curricula, and more reasoning-intensive domains will be necessary to determine whether the observed stability profiles persist beyond this setting. The reported run-to-run consistency figures are inherently subject to change due to the dynamic nature of LLMs, as the models are continually updated by developers. Accordingly, the reported stability profiles should be interpreted as snapshots tied to the specific access dates and deployed versions, and institutions should re-evaluate models periodically as version drift occurs. Because runs were conducted via developer-managed web interfaces using default settings, LLM generation is inherently stochastic in these interfaces, and run-to-run variation and unannounced backend changes can affect reproducibility; thus, our results should be interpreted as time-stamped, platform-specific estimates of stability. Despite these constraints, the study highlights the necessity of using stability-aware scoring metrics over mere accuracy. Future research should apply this consistency framework to international curricula and broader item types, including visual questions, to establish a global benchmark for LLM trustworthiness in medical education, while also exploring the effect of prompt engineering on reducing output volatility.

Conclusion

Contemporary LLMs can achieve both high accuracy and high run-to-run stability on Turkish, curriculum-aligned theoretical anatomy MCQs. Because acceptable mean accuracy can conceal response volatility, adoption decisions should prioritize stability-aware scoring and consistent-correct rates rather than single-trial accuracy alone. We view stability as complementary to accuracy rather than a replacement for it, and we do not propose a universal cutoff; thresholds should depend on educational risk. As an operational guide, for low-stakes self-study tools, programs may require high Consistent & correct rates with very low Consistent & wrong rates (e.g., ≥ 85% and ≤ 5%, respectively), whereas for higher-stakes assessment support, stricter criteria are advisable (e.g., ≥ 90% Consistent & correct with ≤ 1% Consistent & wrong), with outputs remaining assistive under human oversight. Version updates within a family may yield selective—not guaranteed—gains; institutions should therefore validate locally on their own item banks and re-evaluate periodically as models evolve. For anatomy quizzes and item vetting, prefer models with high consistent-correct rates over those with similar means but greater run-to-run spread. Looking ahead, extending this stability-aware framework to multimodal anatomy items and to constructed-response tasks that assess explanation quality will provide a more complete foundation for trustworthy, learner-facing use in medical education.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-026-08656-3.

Supplementary Material 1

Supplementary Material 2

Acknowledgements
The authors would like to thank Kırşehir Ahi Evran University for providing access to the AYDEP platform and supporting the digital infrastructure necessary for conducting this study.

Authors' contributions
Ömer Alperen GÜRSES: Methodology; statistical analysis; supervision; writing; data interpretation; original draft preparation.İsmail CEYLAN: Conceptualization; data collection; writing– review and editing.All authors read and approved the final manuscript.

Funding statement
There is no grant funding to disclose.

Data availability
The datasets generated 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. This study did not involve human participants, human data, or human tissue.
This study is not a clinical trial and therefore does not require registration.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 24 November 2025 / Accepted: 19 January 2026
Published online: 23 January 2026

References

1. Contreras Kallens P, Kristensen-McLachlan RD, Christiansen MH. Large Language models demonstrate the potential of statistical learning in Language. Cogn Sci. 2023;47(3):e13256. https://doi.org/10.1111/cogs.13256.
2. Masters K, Benjamin J, Agrawal A, MacNeill H, Pillow MT, Mehta N. Twelve tips on creating and using custom GPTs to enhance health professions education. Med Teach. 2024;46(6):752–6. https://doi.org/10.1080/0142159X.2024.2305365.
3. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. Correction: How does ChatGPT perform on the united States medical licensing examination (USMLE)? The implications of large Language models for medical education and knowledge assessment. JMIR Med Educ. 2024;10:e57594. https://doi.org/10.2196/45312.
4. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large Language models. PLoS Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198.
5. Zahner D, Steedle JT, Soland J, Welch C, Qin Q, Thompson K, Phelps R. Results from NCME Survey on Revisions to the Standards for Educational and Psychological Testing. EdWorkingPaper No. 25-1124. *Annenberg Institute for School Reform at Brown University* 2025. https://doi.org/10.26300/7zzf-9193
6. Ribeiro MT, Wu T, Guestrin C, Singh S. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:200504118* 2020. https://doi.org/10.48550/arXiv.2005.04118
7. Dodge J, Gururangan S, Card D, Schwartz R, Smith NA. Show your work: improved reporting of experimental results. ArXiv Preprint arXiv. 2019;190903004. https://doi.org/10.48550/arXiv.1909.03004.
8. Atil B, Aykent S, Chittams A, Fu L, Passonneau RJ, Radcliffe E, Rajagopal GR, Sloan A, Tudrej T, Ture F. Non-determinism of deterministic Llm settings. ArXiv Preprint arXiv. 2024;240804667. https://doi.org/10.48550/arXiv.2408.04667.
9. Lee H. The rise of chatgpt: exploring its potential in medical education. Anat Sci Educ. 2024;17(5):926–31. https://doi.org/10.1002/ase.2270.
10. Mavrych V, Ganguly P, Bolgova O. Using large Language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in gross anatomy course: comparative analysis. Clin Anat. 2025;38(2):200–10. https://doi.org/10.1002/ca.24244.
11. Bolgova O, Mavrych V. Evolution of AI in anatomy education study based on comparison of current large Language models against historical ChatGPT performance. Sci Rep. 2025;15(1):37545. https://doi.org/10.1038/s41598-025-22437-w.
12. Bolgova O, Ganguly P, Mavrych V. Comparative analysis of LLMs performance in medical embryology: A cross-platform study of ChatGPT, Claude, Gemini, and copilot. Anat Sci Educ. 2025;18(7):718–26. https://doi.org/10.1002/ase.70044.
13. Singal A, Goyal S. Comparative evaluation of AI platforms Google gemini 2.5 Flash, Google gemini 2.0 Flash, deepseek V3 and ChatGPT 4o in solving multiple-choice questions from different subtopics of anatomy. Surg Radiol Anat. 2025;47(1):1–8. https://doi.org/10.1007/s00276-025-03707-8.
14. Shean R, Shah T, Pandiarajan A, Tang A, Bolo K, Nguyen V, Xu B. A comparative analysis of deepseek R1, deepseek-R1-Lite, openai o1 Pro, and Grok 3 performance on ophthalmology board-style questions. Sci Rep. 2025;15(1):23101. https://doi.org/10.1038/s41598-025-08601-2.
15. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ. 2023;9(1):e48002. https://doi.org/10.2196/48002.
16. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K, Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish medical final examination. Sci Rep. 2023;13(1):20512. https://doi.org/10.1038/s41598-023-46995-z.
17. Şimşek H, Çavdar L. Bir öğrenme yönetim sistemi Olarak aydep'in sistem Kabul edilebilirlik modeline göre incelenmesi. Ahi Evran Üniversitesi Sosyal Bilimler Enstitüsü Dergisi. 2021;7(2):698–717. https://doi.org/10.31592/aeusbed.842764.
18. American Educational Research Association; American Psychological Association; National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association. 2014. ISBN: 978-0-935302-35-6.
19. Norman G, Monteiro S, Salama S. Sample size calculations: should the emperor's clothes be off the Peg or made to measure? BMJ. 2012;345. https://doi.org/10.1136/bmj.e5278.
20. Mogali SR. Initial impressions of ChatGPT for anatomy education. Anat Sci Educ. 2024;17(2):444–7. https://doi.org/10.1002/ase.2261.
21. Bitterman J, D'Angelo A, Holachek A, Eubanks JE. Advancements in large language model accuracy for answering physical medicine and rehabilitation board review questions. *PM&R* 2025. https://doi.org/10.1002/pmrj.13386
22. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci. 2023;3(4):100324. https://doi.org/10.1016/j.xops.2023.100324.
23. Zong H, Wu R, Cha J, Wang J, Wu E, Li J, Zhou Y, Zhang C, Feng W, Shen B. Large Language models in worldwide medical exams: platform development and comprehensive analysis. J Med Internet Res. 2024;26:e66114. https://doi.org/10.2196/66114.
24. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, Rios CJG, Godoy CAR, Quezada KLTP, Gutierrez-Arratia JD. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National licensing medical examination: a cross-sectional study. J Educational Evaluation Health Professions. 2023;20. https://doi.org/10.3352/jeehp.2023.20.30.
25. Collins BR, Black EW, Rarey KE. Introducing anatomygpt: A customized artificial intelligence application for anatomical sciences education. Clin Anat. 2024;37(6):661–9. https://doi.org/10.1002/ca.24178.

Publisher's note