Clinical Anatomy

WILEY

CLINICAL ANATOMY

COMMUNICATION OPEN ACCESS

Generative Artificial Intelligence Performance on University-Level Human Anatomy Examinations: A Structured Narrative Review and Proposed MATRIX-Anatomy Framework

Juan A. Sanchis-Gimeno[1] | Juan José Valenzuela-Fuenzalida[2,3] | Alejandro Bruna-Mejías[2,4] | Mathias Orellana-Donoso[5] | Glen J. Paton[6] | Shahed Nalla[7]

[1]GIAVAL Research Group, Faculty of Medicine, University of Valencia, Valencia, Spain | [2]Departamento de Morfología, Facultad de Medicina, Universidad Andrés Bello, Viña del Mar, Chile | [3]Departamento de Ciencias Químicas y Biológicas, Facultad de Ciencias de la Salud, Universidad Bernardo O'Higgins, Santiago, Chile | [4]Departamento de Ciencias y Geografía, Facultad de Ciencias Naturales y Exactas, Universidad de Playa Ancha, Valparaíso, Chile | [5]Escuela de Medicina, Universidad Finis Terrae, Santiago, Chile | [6]Department of Chiropractic, Faculty of Health Sciences, University of Johannesburg, Johannesburg, South Africa | [7]Department of Human Anatomy and Physiology, Faculty of Health Sciences, University of Johannesburg, Johannesburg, South Africa

Correspondence: Shahed Nalla (shahedn@uj.ac.za)

Received: 14 July 2026 | **Revised:** 4 August 2026 | **Accepted:** 20 August 2026

Keywords: anatomy education | assessment validity | clinical anatomy | generative artificial intelligence | large language models | university examinations

ABSTRACT

Generative artificial intelligence (GenAI) can perform strongly on written anatomy examinations, but whether such scores represent anatomical competence remains uncertain because results vary with the model, assessment, protocol, modality, scoring, and comparator. We synthesized studies evaluating GenAI as the examinee in university-level human anatomy assessments and proposed MATRIX-Anatomy, a reporting and interpretive framework not yet externally validated. PubMed, Scopus, and Web of Science Core Collection were searched for records published from January 2022 to 11 July 2026. Eligible studies used university examinations, course item banks, or curriculum-aligned undergraduate benchmarks and reported quantitative performance. Three reviewers completed study selection, data extraction, and narrative synthesis by consensus. Of 115 records, 51 duplicates were removed, 64 were screened, and 15 studies were included. Leading systems scored 76% to 98% on text-based multiple-choice assessments. On a fixed 120-item set, accuracy increased from 45.8% with ChatGPT-3.5 to 86.7% with ChatGPT-5. Human comparisons were mixed. Visuospatial performance was weaker: ChatGPT-4o identified 22.26% of cadaveric structures after up to three attempts; ChatGPT-4.0 achieved 17.3% end-to-end accuracy on image-based anatomy; and ChatGPT-5.1 reached 74.4% on a surgical-anatomy subset. Repeated runs revealed volatility and consistently incorrect responses. The findings support supervised formative use with authoritative verification and retention of secure supervised, oral, constructed-response, visuospatial, and practical assessments. MATRIX-Anatomy specifies six domains (Model, Assessment, Testing protocol, Reference standard, Input, and eXternal validity) for reproducible reporting and defensible interpretation, but requires formal external validation.

1 | Introduction

Human anatomy is a core component of undergraduate curricula in medicine, dentistry, physiotherapy, nursing, and related health professions (Connolly et al. 2018; Matthan et al. 2020; Smith et al. 2016; Woodley et al. 2023). Anatomical competence is assessed through heterogeneous tasks that sample different constructs. Written multiple-choice and single-best-answer examinations can assess terminology, structural relations, function, and clinical application, whereas practical examinations may require recognition and localization on cadavers or prosections, models, radiographs, cross-sectional images, and other visual materials (Choudhury and Freemont 2017; Rowland et al. 2011; Smith et al. 2016). A score is therefore interpretable only in relation to the task sampled and the educational decision for which it is used; performance on a text-only test cannot, by itself, establish the broader visuospatial and practical competence intended by an anatomy curriculum (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME] 2014; Kane 2013).

Generative artificial intelligence (GenAI) systems, particularly large language models (LLMs), have rapidly become examinees in anatomy research. Early work showed that a freely available ChatGPT version could pass a head-and-neck anatomy test, and later studies reported near-ceiling performance by contemporary models on selected-response university assessments (Gürses and Ceylan 2026; Keskin and Aygün 2025; Mehrabanian and Zariat 2023; Nahir et al. 2026). However, high benchmark accuracy is not equivalent to reliable or transferable anatomical competence. Models have selected keyed options while providing incorrect explanations, reproduced stable incorrect answers across independent administrations, and achieved low accuracy on cadaveric and standardized image-based tasks (Gürses and Ceylan 2026; Melczewski et al. 2026; Paslı and Günenç Beşer 2026; Sarantopoulos et al. 2026). The relevant educational question is therefore not simply whether a model can "pass anatomy," but what construct was tested, under which conditions, and what inference the score can support.

Cross-study comparison is difficult because the technological object and the assessment procedure both vary. The umbrella label "ChatGPT" has encompassed GPT-3.5, GPT-4/4o, GPT-5, and reasoning variants evaluated at different access dates and under different interfaces, conversation structures, and tool conditions (Bolgova and Mavrych 2025; Gürses and Ceylan 2026; Paslı and Günenç Beşer 2026). Studies have also differed in prompt design, number and spacing of runs, feedback policy, language, modality, item format, scoring rules, pass standards, human comparators, and assessment provenance. Fixed-item comparisons document substantial generational gains, whereas repeated administrations reveal session-to-session variability, stable errors, and possible order or context effects (Bolgova and Mavrych 2025; Gürses and Ceylan 2026; Mavrych et al. 2025). Consistent with current LLM reporting guidance, reproducible evaluation requires the provider, exact model and variant, access date, interface or application programming interface, browsing or retrieval status, prompts, session structure, relevant generation settings, and number of independent runs (Gallifant et al. 2025). Claims about performance on a university examination must also be proportional to the curricular alignment, security, and authenticity of the item source (AERA, APA, and NCME 2014).

A recent systematic review synthesized ChatGPT across anatomy education broadly and documented substantial heterogeneity in model versions, topics, and performance (Erbek et al. 2026). A focused synthesis is still needed for university-level human anatomy examinations because this domain combines written knowledge testing with visuospatial, cadaveric, and practical assessment and therefore requires an interpretive structure that preserves distinctions among model, assessment, protocol, comparator, input, and educational context.

The primary objective of this review was to synthesize empirical studies evaluating GenAI systems as examinees on university-level human anatomy examinations. The secondary objective was to propose, rather than validate, MATRIX-Anatomy as a six-domain framework for reproducible reporting and defensible interpretation of anatomy examination benchmarks.

2 | Materials and Methods

2.1 | Design and Reporting Approach

This study was designed as a structured narrative review with reproducible database searching. The narrative format was selected because the eligible reports differed substantially in assessment source, item count and format, model family, access date, prompt design, number of runs, language, scoring rules, pass standards, and human comparators. Pooling these heterogeneous proportions into one accuracy estimate would imply interchangeability that the evidence does not support. The manuscript was developed with reference to SANRA for narrative review quality and PRISMA-S for transparent reporting of literature searches (Baethge et al. 2019; Rethlefsen et al. 2021).

2.2 | Information Sources and Search Strategy

PubMed, Scopus, and Web of Science Core Collection were searched on 11 July 2026 for records published from 1 January 2022 onward. No database-level language restriction was applied. Search strategies combined four mandatory concepts: (1) GenAI, large language models, multimodal language models, or named systems; (2) anatomy or neuroanatomy near examination-related terms; (3) a university, undergraduate, medical-school, or student context; and (4) an answering, scoring, accuracy, passing, or benchmarking outcome. Full database-specific strategies are reproduced in Appendix S1. PubMed was searched without the MEDLINE subset filter so that recent records not yet fully indexed in MEDLINE were retained.

2.3 | Eligibility Criteria

Original empirical studies were eligible when: (a) a GenAI or large language model acted as the examinee or respondent; (b) the primary subject was human anatomy or an anatomy component was explicitly identifiable; (c) questions came from a university examination, past course examination, institutional item bank, faculty-authored curriculum-aligned assessment, or a clearly undergraduate-level anatomy benchmark; and (d) quantitative performance was reported, including accuracy, score, pass/fail status, consistency, or comparison with learners or experts. Medicine, dentistry, physiotherapy, and comparable health-professions curricula were eligible.

Studies were excluded when they evaluated national licensing, postgraduate board or specialty, residency, or university-entrance examinations; focused on question generation, automated grading, tutoring, or student learning outcomes rather than AI-as-examinee performance; addressed veterinary or nonhuman anatomy; were non-empirical reviews, editorials, comments, or correspondence without eligible quantitative data; used mixed basic-science examinations without a separable anatomy outcome; or assessed general question answering or image interpretation without an examination or assessment benchmark. Constructed benchmarks were retained only when their content and format were explicitly aligned with undergraduate anatomy and were interpreted as lower-authenticity evidence. Mixed-domain image studies were used only when the anatomy component was identifiable; pooled results were not treated as anatomy-specific estimates.

2.4 | Record Management and Study Selection

The database exports contained 115 records: 41 from PubMed, 42 from Scopus, and 32 from Web of Science Core Collection. Record management and deduplication were undertaken by three reviewers. Records were deduplicated first by DOI and then by normalized title. Fifty-one duplicates were removed, leaving 64 unique references. The three reviewers screened all titles and abstracts against the predefined eligibility criteria and assessed full texts or publisher records for potentially eligible studies. Differences in eligibility judgments were discussed until consensus was reached. Forty-nine records were excluded, and 15 studies were included.

2.5 | Data Extraction and Synthesis

The following variables were extracted when reported: publication year; educational program; assessment source and authenticity; number and format of items; anatomical domain; language; model provider and exact version; access conditions; prompt and session structure; number of runs; use of images; scoring method; pass threshold; student or expert comparator; and principal quantitative findings. Data extraction was conducted by three reviewers using a shared extraction framework. Extracted values and study-level classifications were reviewed collectively, and any discrepancies were resolved through discussion and consensus among all three reviewers.

Data were extracted from full texts when available. For records for which full text could not be retrieved, only information explicitly reported in abstracts or publisher metadata was used; no unreported values were inferred. The three reviewers jointly cross-checked bibliographic details and DOIs against PubMed, publisher records, and DOI resolution and organized findings into text-based theoretical examinations, model-generation effects, human comparisons, image-based or practical examinations, and repeatability or protocol effects. Interpretive differences were resolved through discussion and consensus.

2.6 | Methodological Appraisal

Given the heterogeneous benchmark designs and the absence of an established appraisal instrument tailored to GenAI examination studies, no numerical risk-of-bias score was assigned. Instead, the three reviewers appraised model specification, assessment provenance, protocol transparency, comparator adequacy, input modality, and external validity descriptively. These recurring methodological features informed development of MATRIX-Anatomy.

2.7 | Development of the Proposed MATRIX-Anatomy Framework

During narrative synthesis, the three reviewers inductively coded recurring sources of interpretive uncertainty and consolidated them through discussion and consensus into six domains: Model identity and temporal lock; Assessment provenance and curricular authenticity; Testing protocol and repeatability; Reference standard and human comparator; Input modality, item format, and cognitive demand; and eXternal validity and educational interpretation. The domains were aligned with assessment-validity principles that require explicit links between observed scores, score interpretations, and intended uses (AERA, APA, and NCME 2014; Kane 2013). MATRIX-Anatomy was developed as a conceptual reporting and interpretive proposal, not as a validated measurement instrument. Formal validation was not undertaken in this review and would require independent external experts, prespecified consensus criteria, pilot application, and reliability and content-validity testing. Accordingly, MATRIX-Anatomy should not be used as a risk-of-bias instrument or converted into a total score.

3 | Results

3.1 | Study Selection and Evidence-Based Characteristics

Database searches completed on 11 July 2026 identified 115 records: 41 from PubMed, 42 from Scopus, and 32 from Web of Science Core Collection. After 51 duplicate records were removed, 64 unique references were screened. Forty-nine records were excluded, leaving 15 studies for inclusion in the narrative synthesis (Table 1). Table 1 also states the

TABLE 1 | Study-selection dispositions and operational scope boundaries.

Disposition	Exclusion or inclusion category	Operational scope boundary	Records, *n*
Excluded	AI did not act as the examinee	Studies of question generation, grading, tutoring, or learner interventions were outside the AI-as-examinee objective.	21
Excluded	Not a university-level human anatomy examination or anatomy not primary/separable	Assessments had to be explicitly aligned with undergraduate human anatomy and report a separable anatomy outcome.	13
Excluded	Non-empirical publication without eligible data	Reviews, commentaries, correspondence, and methodological papers without quantitative GenAI-as-examinee performance were excluded.	8
Excluded	Postgraduate specialty or professional examination	Residency, board, specialty, and other postgraduate examinations were outside the university-course scope.	4
Excluded	Clinical scenario response study	General clinical response generation without an examination or exam-style item set was ineligible.	1
Excluded	Mixed basic-science examination without an anatomy-specific outcome	Mixed anatomy, physiology, or biochemistry scores were retained only when an anatomy result was separable.	1
Excluded	National medical licensing examination	National licensing examinations were excluded to preserve the university-level course construct.	1
Included	Eligible GenAI-as-examinee study	University examination, curriculum-linked item bank, or explicitly undergraduate-aligned human anatomy benchmark with quantitative performance.	15

Note: "University-level" denotes undergraduate learners in medicine, dentistry, physiotherapy, or a comparable health-professions curriculum. Constructed benchmarks were eligible only when explicitly aligned with undergraduate anatomy.

operational scope boundary applied to each exclusion or inclusion category.

The included studies were published between 2023 and 2026: one in 2023, six in 2025, and eight in 2026. Eleven studies primarily evaluated text-based selected-response questions, whereas four examined mixed, image-based, or practical tasks. Assessment sizes ranged from 23 prompts to 555 theoretical questions and 265 labeled cadaveric structures. OpenAI models were evaluated most frequently, although the evidence base also included Gemini, Claude, Copilot, DeepSeek, Grok, Bard/PaLM, and Polish-language models. Assessment sources ranged from past university examinations and institutional course banks to faculty-written or textbook-derived undergraduate benchmarks. Study-level characteristics are summarized in Tables 2 and 3.

3.2 | Performance on Text-Based Theoretical Examinations

The most consistent finding was strong performance by recent systems on text-based multiple-choice question (MCQ) examinations. Copilot, ChatGPT, and Gemini scored 97.7%, 94.4%, and 86.5%, respectively, on 286 questions from first- and second-year medical examinations (Keskin and Aygün 2025). Four systems scored between 91.4% and 95.7% on 70 first- and second-year anatomy questions, and four contemporary platforms achieved an aggregate accuracy of 95.9% on a 55-item multi-topic set (Nahir et al. 2026; Singal and Goyal 2025). On 100 Turkish curriculum-aligned physiotherapy anatomy questions, the leading reasoning model was consistently correct across all three runs on 98 items (Gürses and Ceylan 2026).

Strong performance was not universal. On 50 upper-limb questions from a gross-anatomy course database, GPT-4 achieved 60.5%, while five other systems ranged from 33.5% to 42.0% (Mavrych et al. 2025). The early head-and-neck benchmark yielded 73.33%, and ChatGPT scored 60% on an introductory dental anatomy examination (Mehrabanian and Zariat 2023; Ullah et al. 2026). These differences show why "ChatGPT performance" is not a stable outcome: model generation, item bank, curriculum, prompt, and scoring protocol can move the same broad product category from marginal to near-ceiling performance.

Item format and intended cognitive demand also mattered. In 555 past examination questions, multiple-answer items

TABLE 2 | Characteristics of text-based university-level human anatomy examination studies.

Study	Educational context and assessment source	Item format and anatomical scope	Models and testing protocol	Reference standard or human comparator	Principal findings
Mehrabanian and Zariat (2023)	Dental learners; textbook-derived constructed benchmark.	75 single-answer text items in head-and-neck anatomy; no images.	Free ChatGPT (GPT-3.5); written administration.	Reported pass criterion; no human comparator.	73.33% correct and reported as passing.
Ganapathy and Kaushal (2025)	Curriculum-manual-derived undergraduate anatomy questions.	Text questions classified as knowledge, comprehension, or application.	ChatGPT-4o mini and Gemini.	Expert-prepared answer key; no learner comparator reported.	Overall 76.15% vs. 72.84%, favoring ChatGPT. Gemini was higher on comprehension; both produced factual or contextual errors.
Mavrych et al. (2025)	Expert-reviewed random sample from a gross-anatomy course examination database.	50 USMLE-style upper-limb MCQs.	GPT-4, GPT-3.5 variants, Copilot, PaLM 2, Bard, and Gemini; five successive attempts.	Expert-reviewed course items; cross-model comparison; no learner comparator.	GPT-4: 60.5% ± 1.9%; Copilot: 42.0%; GPT-3.5: 41.0%; other systems were lower. Repetition exposed variability.
Keskin and Aygün (2025)	First- and second-year medical midterm, board, final, and make-up examinations; item analysis available for 222 questions.	286 text-based anatomy questions.	Copilot, ChatGPT, and Gemini under direct comparative testing.	Student cohort performance and item-difficulty data.	97.7%, 94.4%, and 86.5%; all exceeded students overall, but the advantage narrowed on very difficult items.
Singal and Goyal (2025)	Undergraduate anatomy context; item provenance not reported in the available abstract.	55 text MCQs across multiple anatomy subtopics.	Gemini 2.5 Flash, Gemini 2.0 Flash, DeepSeek V3, and ChatGPT-4o; further protocol details not reported in the abstract.	No human comparator reported.	Aggregate accuracy 95.9%; Gemini 2.5 ranked first. Between-model differences were not statistically significant.
Bolgova and Mavrych (2025)	Gross-anatomy course examination database.	325 USMLE-style MCQs across seven gross-anatomy regions.	GPT-4o, Claude 3.5 Sonnet, Copilot, and Gemini 1.5 Flash; three attempts; historical GPT-3.5 comparator.	Historical model and cross-model comparison; no learner comparator reported.	Current-model mean 76.8% ± 12.2% vs. GPT-3.5 44.4% ± 8.5%. GPT-4o scored 92.9%; 2.5% of items were never correct.

(Continues)

TABLE 2 | (Continued)

Study	Educational context and assessment source	Item format and anatomical scope	Models and testing protocol	Reference standard or human comparator	Principal findings
Nahir et al. (2026)	First- and second-year medical examination questions.	70 MCQs across seven anatomical systems.	GPT-4.1, Copilot, DeepSeek, and Gemini under common conditions.	Item-difficulty analysis; no learner comparator reported.	95.7%, 94.3%, 92.9%, and 91.4%; no significant relation between accuracy and item difficulty.
Ullah et al. (2026)	Undergraduate introductory dental-anatomy examination.	Text-based MCQ examination.	ChatGPT; expert grading of explanations.	Direct comparison with dental students.	Students: 74.28%; ChatGPT: 60%. ChatGPT passed but was less accurate and less reliable than the cohort.
Gürses and Ceylan (2026)	Turkish faculty-authored, curriculum-aligned undergraduate physiotherapy assessment.	100 anatomy MCQs.	Eleven contemporary models; three independent runs at least 12 h apart; browsing disabled.	Faculty answer key; repeatability classification; no learner comparator.	Consistently correct: ChatGPT-5 Thinking 98/100 and Gemini 2.5 Pro 97/100. Stable wrong answers and model-specific volatility persisted.
Paslı and Günenç Beşer (2026)	Same faculty-written item set administered over approximately two years.	120 five-option text MCQs.	Free ChatGPT-3.5, ChatGPT-4o, and ChatGPT-5; fixed-item longitudinal comparison.	Fixed answer key and longitudinal model comparator; no human cohort.	Accuracy rose from 45.8% to 73.3% and 86.7% ($p < 0.001$). Informational errors remained dominant.
Chaba-Karnaś et al. (2026)	Past medical-program examinations: 150 Polish and 405 English questions.	555 single- and multiple-answer anatomy MCQs.	Seven models; one run per item.	Past-examination answer keys; cross-language and cross-model comparison; no human comparator reported.	ChatGPT-4o 83.1%, Copilot 79.6%, Gemini 78.8%, DeepSeek 76.9%; multiple-answer items were hardest (37.6% overall).

Abbreviations: LLM, large language model; MCQ, multiple-choice question; USMLE, United States Medical Licensing Examination. "Not reported" indicates that the information was unavailable in the source used for this review.

TABLE 3 | Characteristics of mixed, image-based, and practical human anatomy examination studies.

Study	Educational context and assessment source	Item format and anatomical scope	Models and testing protocol	Reference standard or human comparator	Principal findings
Al-Khater (2025)	Constructed mixed-modality undergraduate benchmark using USMLE Step 1-type material.	23 prompts: 10 knowledge, 10 analysis-based questions, and 3 radiographs.	ChatGPT, Gemini, and Claude.	No learner comparator; responses were evaluated for accuracy and qualitative comprehensiveness.	ChatGPT was reported as 100% accurate; Gemini 60%; Claude had the strongest qualitative comprehensiveness. Three radiographs were insufficient for a robust visual inference.
Melczewski et al. (2026)	Cadaveric photographs designed to mimic a practical anatomy examination.	Open-ended identification of 265 marked structures.	Free ChatGPT-4o; up to three attempts with standardized feedback after errors.	Anatomical target labels and feedback-assisted cumulative scoring; no human comparator.	Overall accuracy 22.26%; osteology 64.71%; isolated thoracic organs 8.82%. Non-existent anatomical terms occurred.
Sarantopoulos et al. (2026)	Standardized image-based examinations in anatomy, pathology, and pediatrics.	Image-based anatomy items evaluated with recognition-gated and end-to-end scoring.	Vision-capable ChatGPT-4.0; deterministic and stochastic protocol.	Direct medical-student comparison.	In anatomy, mean difference vs. students −0.387 ($p < 0.00001$; $d = 2.10$). End-to-end anatomy accuracy was 17.3%.
Lu et al. (2026)	Image-only MD-program assessment with separable surgical-anatomy results.	208 questions; 43 (20.7%) surgical anatomy, remainder anatomical pathology or radiology; selected and constructed responses.	ChatGPT-5.1, Gemini-2.5, Claude Sonnet-4.5, and Grok-4.	Distinct surgical-anatomy subgroup; no learner comparator reported.	Surgical-anatomy accuracy: 74.4%, 58.1%, 51.2%, and 34.9%, respectively. Selected-response and recognition-only items were easier across the full mixed-domain set.

Note: Mixed-domain headline scores were not interpreted as anatomy-specific unless a distinct anatomy subgroup was reported.
Abbreviation: USMLE, United States Medical Licensing Examination.

produced the weakest aggregate performance (37.6%), whereas questions concerning organ function were easier (Chaba-Karnaś et al. 2026). In a curriculum-linked cognitive-domain comparison, ChatGPT-4o mini was stronger overall and on application items, whereas Gemini was stronger on comprehension items (Ganapathy and Kaushal 2025). Total accuracy should therefore be accompanied by item format and cognitive-demand analyses. Table 2 summarizes the corresponding study-level characteristics.

3.3 | Model-Generation Effects and Temporal Drift

Two studies provided direct evidence of generational change while holding the item set relatively stable. Paslı and Günenç Beşer (2026) reused 120 faculty-written questions and observed an increase from 45.8% with ChatGPT-3.5 to 73.3% with ChatGPT-4o and 86.7% with ChatGPT-5. Bolgova and Mavrych (2025) found that four current models averaged 76.8%, compared with 44.4% for historical GPT-3.5 on the same 325-question bank; GPT-4o reached 92.9%. These within-dataset comparisons are more informative than comparisons across unrelated studies because the assessment remains relatively constant.

Newer branding did not guarantee uniformly better behavior. In the three-run Turkish study, ChatGPT-5 Thinking was consistently correct on 98 of 100 items, whereas the base ChatGPT-5 interface was consistently correct on only 30; the authors also observed late-batch deterioration for selected models (Gürses and Ceylan 2026). Exact model variant, interface, access date, conversation structure, and tool settings are therefore essential metadata rather than optional technical detail.

3.4 | Human Comparators

Only a minority of studies compared AI and learners on the same assessment, and the direction of effect was mixed. All three systems exceeded aggregate student accuracy on institutional medical examinations, although the advantage narrowed on very difficult items (Keskin and Aygün 2025). By contrast, dental students averaged 74.28% while ChatGPT scored 60% on an introductory dental anatomy examination (Ullah et al. 2026). On standardized image-based anatomy items, ChatGPT-4.0 substantially underperformed students, with a mean difference of −0.387 and a large standardized effect (Sarantopoulos et al. 2026). A pass mark, cohort mean, percentile, and item-level difficulty index answer different questions and should not be treated as interchangeable definitions of "human-level" performance.

3.5 | Image-Based and Practical Anatomy Performance

The sharpest contrast with text-based results occurred in authentic visuospatial tasks. Free ChatGPT-4o identified only 22.26% of 265 labeled cadaveric structures after up to three attempts; first-attempt success occurred for 33 structures, and performance was lowest for isolated thoracic organs (8.82%) (Melczewski et al. 2026). In standardized image-based examinations, end-to-end ChatGPT-4.0 accuracy was 17.3% for anatomy when image-recognition failures were scored as incorrect, and the model remained substantially below students after recognition-gated comparison (Sarantopoulos et al. 2026).

Other multimodal findings require cautious interpretation. A 23-prompt constructed benchmark reported 100% accuracy for ChatGPT but contained only three radiographs, an insufficient visual sample for a strong practical inference (Al-Khater 2025). In a larger 208-item image-only study, 43 questions were surgical anatomy. Surgical-anatomy accuracy was 74.4% for ChatGPT-5.1, 58.1% for Gemini-2.5, 51.2% for Claude Sonnet-4.5, and 34.9% for Grok-4; selected-response and recognition-only items were easier across the full mixed-domain set (Lu et al. 2026). These findings indicate improving multimodal capability, but performance remained model- and task-sensitive and did not erase the poor results observed on unfamiliar cadaveric material. Table 3 summarizes the corresponding study-level characteristics for mixed, image-based, and practical studies.

3.6 | Repeatability, Stable Errors, and Protocol Effects

Repeated administrations showed that single-run accuracy is an incomplete measure of reliability. Mavrych et al. (2025) used five attempts, Bolgova and Mavrych (2025) used three, and Gürses and Ceylan (2026) classified responses as consistently correct, consistently wrong, partially consistent, or inconsistent. The last study demonstrated that near-ceiling means could coexist with a smaller set of stable systematic errors. In an educational context, a repeated false answer may be particularly hazardous because consistency can be mistaken for trustworthiness.

Administration conditions can alter the response process and, consequently, the construct represented by the resulting score (AERA, APA, and NCME 2014). In one exploratory anatomy study, a 100-item single-batch administration showed marked late-sequence declines for ChatGPT-5 and Grok-4, a pattern consistent with possible order or context-window effects, although the effect was small for most other models (Gürses and Ceylan 2026). Starting a new conversation for each administration can minimize within-conversation carryover by design, while repeated independent runs under otherwise identical inputs can reveal session-to-session variation and consistently incorrect responses (Gürses and Ceylan 2026). Other repeated-attempt anatomy studies likewise documented nonzero between-run variation (Bolgova and Mavrych 2025; Mavrych et al. 2025). In the cadaveric study, the model was allowed up to three attempts with standardized feedback after incorrect responses; cumulative accuracy therefore represented feedback-assisted correction rather than first-attempt examination performance (Melczewski et al. 2026). First-attempt accuracy, repeated-run stability, and feedback-assisted performance should consequently be reported as distinct outcomes (AERA, APA, and NCME 2014; Gallifant et al. 2025).

3.7 | Proposed MATRIX-Anatomy Synthesis

The proposed six-domain MATRIX-Anatomy framework is presented in Figure 1 and operationalized in Table 4. Its purpose is to prevent a common interpretive error: treating one accuracy percentage as a timeless property of a product rather than the result of a defined model, assessment, protocol, reference standard, input, and educational context. The framework specifies the minimum information needed to determine whether a benchmark can be reproduced and what inference its score can reasonably support. It is a conceptual reporting and interpretive framework, not a validated checklist or risk-of-bias instrument, and no total score is recommended.

Table 4 translates the proposed domains into minimum reporting items, core appraisal questions, interpretive purposes, and representative evidence patterns.

4 | Discussion

4.1 | Principal Findings

This review supports a clear but bounded conclusion. Contemporary GenAI systems can be strong examinees on text-based university-level human anatomy assessments, and several current models reached the 90% to 98% range on selected-response item banks (Gürses and Ceylan 2026; Keskin and Aygün 2025; Nahir et al. 2026). Fixed-item comparisons also show large gains across model generations (Bolgova and Mavrych 2025; Paslı and Günenç Beşer 2026). These results are educationally consequential because non-invigilated text-only examinations can no longer be assumed to reflect unaided student knowledge when GenAI access is uncontrolled (AERA, APA, and NCME 2014; Lodge et al. 2023).

The evidence also establishes important boundaries. Textual success did not reliably transfer to cadaveric identification or image-based anatomy, direct human comparisons were inconsistent, and repeated runs revealed volatility and stable errors (Melczewski et al. 2026; Sarantopoulos et al. 2026; Ullah et al. 2026). "Passing anatomy" is therefore an assessment-specific statement, not evidence of comprehensive anatomical competence.

4.2 | Why Reported Performance Varies

The observed range, from 22.26% cumulative accuracy after up to three attempts on cadaveric photographs to 97.7% accuracy on an institutional text-based MCQ set, with another study reporting 98 of 100 items consistently correct across three independent runs, is not paradoxical when the MATRIX-Anatomy domains are made explicit (Gürses and Ceylan 2026; Keskin and Aygün 2025; Melczewski et al. 2026). Across studies, the evaluated systems differed in model generation and variant, access date, interface, prompt and conversation structure, language, assessment provenance, item format, cognitive demand, and repetition protocol (Bolgova and Mavrych 2025; Chaba-Karnaś et al. 2026; Gürses and Ceylan 2026; Paslı and Günenç Beşer 2026). For example, the cadaveric study required open-ended identification of marked structures, whereas

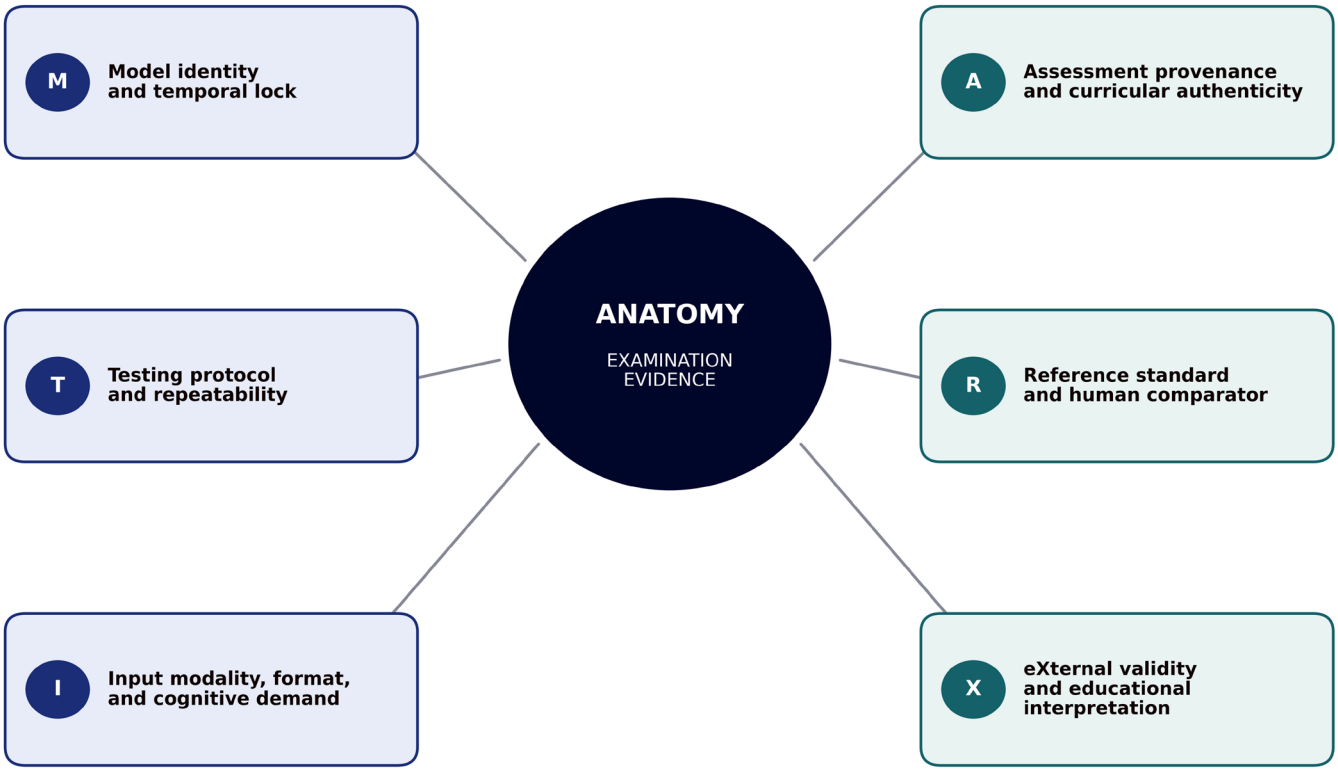

FIGURE 1 | Proposed MATRIX-Anatomy reporting and interpretive framework. The six domains identify the minimum information required for reproducible reporting and defensible interpretation of generative artificial intelligence performance on university-level human anatomy examinations. The framework has not yet been externally validated, is not a risk-of-bias instrument, and should not be scored.

TABLE 4 | Proposed MATRIX-Anatomy domains, reporting requirements, core appraisal questions, interpretive purposes, and representative evidence patterns.

Domain	Minimum reporting items	Core appraisal question	Interpretive purpose	Representative evidence pattern
M—Model identity and temporal lock	Provider; exact model and variant; interface; free or paid access; date and time; browsing, retrieval, tools, multimodal capability; relevant settings.	Which exact technological system was tested, and can it be recreated at that date?	Identifies the technological object tested and whether it can be reproduced.	Performance differed substantially across model generations and variants, including within the same product family (Bolgova and Mavrych 2025; Gürses and Ceylan 2026; Paslı and Günenç Beşer 2026).
A—Assessment provenance and curricular authenticity	Institution and program; course and year; operational, archival, banked, or constructed items; item security; source validation; pass standard.	How closely does the item set represent a secure, authentic university anatomy assessment?	Determines whether the item set supports an inference about a real university examination.	The evidence included authentic institutional examinations and constructed or textbook-derived benchmarks with different levels of curricular authenticity and potential exposure risk (Al-Khater 2025; Keskin and Aygün 2025; Mehrabanian and Zariat 2023).
T—Testing protocol and repeatability	Full prompts; system instructions; language and translation; one item versus batch; new-session policy; number and spacing of runs; feedback; scoring rules.	Could prompts, session structure, feedback, or repeated runs change the observed score?	Shows whether procedural choices could alter the result and whether performance is stable.	Repeated administrations revealed response volatility, stable errors, and context or order effects that single-run designs could not detect (Bolgova and Mavrych 2025; Gürses and Ceylan 2026; Mavrych et al. 2025).
R—Reference standard and human comparator	Answer-key provenance; expert adjudication; pass mark; learner cohort and year; item difficulty; inter-rater agreement.	Against which answer standard, pass criterion, or human group is performance judged?	Clarifies relative to whom or what performance is judged "good."	AI exceeded students on some theoretical examinations but underperformed them in dental and image-based anatomy assessments (Keskin and Aygün 2025; Sarantopoulos et al. 2026; Ullah et al. 2026).
I—Input modality, item format, and cognitive demand	Text, image, cadaver, radiograph, model, or mixed input; selected versus constructed response; single versus multiple answer; cognitive level; anatomical region.	Which component of anatomical competence did the input and response format sample?	Defines which component of anatomical competence was sampled.	Text-based selected-response items generally produced higher scores than multiple-answer, constructed-response, cadaveric, or image-recognition tasks (Chaba-Karnaś et al. 2026; Lu et al. 2026; Melczewski et al. 2026; Sarantopoulos et al. 2026).

(Continues)

TABLE 4 | (Continued)

Domain	Minimum reporting items	Core appraisal question	Interpretive purpose	Representative evidence pattern
X—eXternal validity and educational interpretation	Language; country; curriculum; local terminology; student availability; contamination risk; temporal reproducibility; intended educational use.	To which curricula, languages, institutions, dates, and educational uses can the findings reasonably generalize?	Limits generalization beyond the test, date, institution, and intended use.	Transferability across languages, curricula, model access dates, and institutions remained uncertain, and training-data exposure was rarely examined directly (Chaba-Karnaś et al. 2026; Gürses and Ceylan 2026; Paslı and Günenç Beşer 2026).

Note: MATRIX-Anatomy is a proposed reporting and interpretive framework. It has not been externally validated, is not a risk-of-bias instrument, and should not be converted into a total score.

other studies used five-option MCQs; some studies queried each item only once, whereas others repeated the complete item set across several independent administrations (Bolgova and Mavrych 2025; Chaba-Karnaś et al. 2026; Melczewski et al. 2026; Paslı and Günenç Beşer 2026). Pooling these non-equivalent outcomes into a single accuracy estimate would obscure educationally meaningful heterogeneity and create a false impression of precision. A structured, domain-stratified narrative synthesis was therefore used (AERA, APA, and NCME 2014; Campbell et al. 2020).

Assessment provenance is particularly important. Publicly available or textbook-derived items may have entered model pretraining or fine-tuning data, but this possibility was generally not investigated directly in the included anatomy studies and should therefore be described as a potential contamination risk rather than assumed prior exposure (Li et al. 2024; Mehrabanian and Zariat 2023). Secure, previously unpublished local questions may reduce but cannot guarantee the absence of prior exposure; they may also reflect institution-specific terminology, curricular sequencing, and pass standards that limit external generalizability (AERA, APA, and NCME 2014; Gürses and Ceylan 2026; Keskin and Aygün 2025). Neither source is inherently superior for every evaluative purpose: public benchmarks facilitate replication and cross-model comparison, whereas locked local examinations provide stronger evidence of curricular authenticity and assessment security. Future studies should ideally combine a secure local assessment with an external replication set and prespecify the interpretation supported by each component (AERA, APA, and NCME 2014; Gallifant et al. 2025).

4.3 | Contribution of This Review

This review makes four distinct contributions. First, it restricts its eligibility criteria and conclusions to university-level human anatomy examinations; national licensing, residency, entrance, and specialty examinations were excluded. Second, it integrates written, cadaveric, radiographic, and multimodal assessments across multiple model families while preserving distinctions among model versions and access conditions (Al-Khater 2025; Lu et al. 2026; Melczewski et al. 2026; Sarantopoulos et al. 2026). Third, it distinguishes model-generation effects from those attributable to assessment design and administration protocol (Bolgova and Mavrych 2025; Gürses and Ceylan 2026; Paslı and Günenç Beşer 2026). Fourth, it translates these sources of heterogeneity into the proposed MATRIX-Anatomy framework, a domain-specific interpretive and reporting framework grounded in assessment-validity principles (AERA, APA, and NCME 2014; Kane 2013). Relative to the broader systematic review of ChatGPT as a supportive tool across anatomy education by Erbek et al. (2026), the present review narrows the unit of inference to university-level examination performance and extends the synthesis beyond a single overall claim about accuracy. It provides an interpretive structure for specifying what was tested, how it was tested, against which reference standard, and which conclusions the resulting evidence can defensibly support.

4.4 | Implications for University Assessment

For take-home or otherwise non-invigilated assessment, conventional text MCQs now provide weaker evidence of unaided individual performance when GenAI access is uncontrolled, because substantial GenAI use is difficult to preclude or detect and contemporary systems can achieve high scores on university-level anatomy MCQs (Bolgova and Mavrych 2025; Keskin and Aygün 2025; Lodge et al. 2023). Written assessment remains useful, but administration conditions, permitted tool use, and alignment among the task, intended construct, and score interpretation should be stated explicitly (AERA, APA, and NCME 2014; Lodge et al. 2023; Perkins et al. 2024). At key progression points, a more defensible assessment program may triangulate evidence through secure supervised testing and interactive oral follow-up (Eachempati et al. 2025; Lodge et al. 2023). Under controlled conditions, additional formats may include progressive-disclosure tasks, brief constructed responses requiring a rationale, and supervised anatomy practicals using cadavers, prosections, models, radiographs, or cross-sectional images (Choudhury and Freemont 2017; Rowland et al. 2011; Runyon et al. 2024; Vuma and Sa 2015). Rules governing permitted, restricted, and prohibited GenAI use should be communicated prospectively and transparently (Lodge et al. 2023; Perkins et al. 2024).

Simply increasing item difficulty is unlikely to provide a reliable safeguard. In one institutional dataset, Copilot remained highly accurate across difficulty categories, whereas ChatGPT and Gemini approached student performance on the most difficult items; in another, item difficulty was not significantly associated with chatbot accuracy (Keskin and Aygün 2025; Nahir et al. 2026). Assessment programs should instead triangulate evidence from processes that are less readily outsourced without revealing the student's own understanding, such as live explanation, follow-up questioning, comparison of plausible alternatives, and justification or defense of reasoning (Eachempati et al. 2025; Lodge et al. 2023). In anatomy, this evidence may be complemented by localization and interpretation on unfamiliar cadaveric, radiographic, cross-sectional, or other visual material, including transfer across views and integration of structure with function (Choudhury and Freemont 2017; Melczewski et al. 2026; Rowland et al. 2011; Sarantopoulos et al. 2026). These formats must nevertheless be supported by evidence of validity, reliability, fairness, accessibility, and feasibility; novelty alone does not establish sound assessment design (AERA, APA, and NCME 2014; Kane 2013; van der Vleuten and Schuwirth 2005).

4.5 | Implications for Teaching and Learning

Strong performance on text-based anatomy examinations creates plausible opportunities for formative use, although benchmark performance alone does not demonstrate educational effectiveness. Large language models can generate personalized practice materials, alternative explanations, and rapid feedback, but their educational use should be embedded in a human-supervised verification workflow that emphasizes critical evaluation and fact-checking (Kasneci et al. 2023; Miao and Holmes 2023). Students should compare generated outputs with authoritative anatomical atlases, textbooks, and faculty guidance, verify anatomical terminology, and distinguish linguistic fluency from anatomical correctness. This caution is supported by the included evidence: informational errors and "guessed-correct" responses persisted across ChatGPT generations (Paslı and Günenç Beşer 2026), hallucinated anatomical terms occurred during cadaveric-image identification (Melczewski et al. 2026), and some models repeatedly selected the same incorrect option across independent runs (Gürses and Ceylan 2026).

A defensible educational model is supervised augmentation rather than uncritical substitution. Faculty could use model outputs as objects of critique by asking learners to identify factual and terminological errors, annotate flawed explanations, compare textual claims with cadaveric or imaging evidence, and justify why an apparently plausible distractor is incorrect. In the present review, these activities should be presented as proposed pedagogical applications rather than as interventions with demonstrated anatomy-learning effects. Their rationale is to combine disciplinary reasoning with critical AI literacy while preserving human agency and faculty oversight (Kasneci et al. 2023; Miao and Holmes 2023).

4.6 | Research Priorities and Framework Validation

Future benchmarking studies should preregister or prospectively publish the assessment and analysis plan, identify and time-lock the exact model and access date, archive complete prompts and raw outputs, and prespecify whether browsing, retrieval, code execution, or other external tools are permitted and how their use will be documented (Gallifant et al. 2025; Nosek et al. 2018). Because single administrations can conceal response volatility and consistently incorrect answers, studies should use repeated independent runs and report item-level stability in addition to aggregate accuracy (Bolgova and Mavrych 2025; Gürses and Ceylan 2026; Mavrych et al. 2025). To strengthen both curricular inference and reproducibility, benchmarking designs should ideally combine a secure, authentic local examination with a prespecified external replication set, define human comparator groups and pass standards in advance, and report item-level difficulty and expert-adjudication procedures (AERA, APA, and NCME 2014; Gallifant et al. 2025; Kane 2013). For image-based assessments, recognition failure should be separated from downstream reasoning failure, and first-attempt accuracy should be distinguished from performance after feedback or repeated correction (Lu et al. 2026; Melczewski et al. 2026; Sarantopoulos et al. 2026).

Outcome reporting should extend beyond total accuracy to include run-to-run consistency, proportion of consistently incorrect responses, item format, anatomical region, cognitive demand, response completeness, hallucinated terminology, and language sensitivity (Chaba-Karnaś et al. 2026; Gürses and Ceylan 2026; Melczewski et al. 2026; Paslı and Günenç Beşer 2026). When model confidence is elicited, calibration should be evaluated rather than assuming that verbal confidence corresponds to correctness (Xiong et al. 2024). Formal

external validation would substantially strengthen the impact and usability of MATRIX-Anatomy. Because the present framework was derived and refined by the three review authors, the next stage should include broader external expert consensus with an a priori definition of consensus, pilot application by reviewers independent of the development team, assessment of inter-rater agreement using a statistic appropriate to the rating scale, evaluation of content validity, and prospective testing of whether its use improves reporting completeness and interpretive consistency (Diamond et al. 2014; McHugh 2012; Moher et al. 2010; Polit and Beck 2006). Until those steps are completed, MATRIX-Anatomy should be treated as a proposed interpretive aid rather than a validated appraisal instrument.

4.7 | Strengths and Limitations

This review has several methodological and conceptual strengths. It addresses a tightly defined educational question and applies reproducible search strategies across three major and complementary databases: PubMed, Scopus, and Web of Science Core Collection. The eligibility criteria clearly distinguished university-level human anatomy assessments from national licensing, residency, entrance, and postgraduate specialty examinations, thereby reducing construct mixing and strengthening alignment between the evidence base and the review objective. The synthesis incorporated text-based, image-based, and practical assessment modalities and preserved clinically and educationally meaningful differences in model version, access date, assessment provenance, administration protocol, item format, and human comparator rather than reducing heterogeneous results to a single overall accuracy estimate.

Record management, deduplication, title and abstract screening, full-text eligibility assessment, data extraction, bibliographic and DOI verification, narrative synthesis, and framework development were undertaken by three reviewers, with all differences resolved through discussion and consensus. The exact database strategies, search date, eligibility criteria, and exclusion categories are reported transparently. In addition, MATRIX-Anatomy provides an anatomy-specific, evidence-informed structure for interpreting study reproducibility, assessment validity, and educational relevance. Together, these features support a more focused and methodologically coherent interpretation than would be possible in a broader synthesis combining anatomy with unrelated health-professions or professional licensing examinations.

Several considerations define the scope of the conclusions. The search covered three large databases with complementary biomedical and multidisciplinary coverage; nevertheless, reports indexed exclusively in education-specific, regional, or grey-literature sources may not have been captured. The substantial heterogeneity of model generations, assessment sources, item formats, modalities, prompts, scoring procedures, and comparator groups made a meta-analysis methodologically inappropriate. Narrative synthesis was therefore selected to preserve these distinctions and to avoid a potentially misleading pooled estimate.

Some included studies used investigator-constructed or textbook-derived benchmarks rather than operational university examinations, and some image-based studies combined anatomy with related domains such as pathology or radiology. These findings were retained only when they were relevant to the predefined review question and were interpreted conservatively, with mixed-domain headline scores not presented as anatomy-specific estimates. Prior exposure of publicly available items to model-training data, as well as possible overlap between question banks, could not be fully excluded because these risks were rarely examined directly in the primary studies.

The deliberate exclusion of national licensing, residency, entrance, and postgraduate specialty examinations limits generalization beyond university-level anatomy curricula, but it also strengthens the conceptual coherence of the review and the specificity of its educational conclusions. Rapid changes in model architecture, interfaces, and multimodal capabilities are an inherent feature of this research field; the synthesis is therefore anchored to the search date of 11 July 2026 and should be interpreted as an assessment of the evidence available at that time. Most importantly, MATRIX-Anatomy is an original proposal developed through consensus among the three review authors and has not been externally validated. Its content, usability, and interpretive consistency may therefore reflect the judgments of the development team. Independent expert consensus, pilot application by external reviewers, assessment of inter-rater agreement, formal content validity evaluation, and prospective utility testing are required before it can be considered a standardized appraisal instrument. These considerations delimit, rather than undermine, the review's principal conclusions.

5 | Conclusion

Contemporary generative AI systems can achieve high, and in some studies near-ceiling, accuracy on text-based, selected-response university-level human anatomy examinations. However, performance remains strongly dependent on the exact model and access date, assessment provenance, administration protocol, item format, input modality, and comparator. Direct comparisons with students were context-specific, while cadaveric and image-based assessments revealed substantially weaker or more variable performance. Run-to-run instability, consistently incorrect responses, and correct options accompanied by flawed explanations further indicate that success on a written examination should not be interpreted as evidence of comprehensive, transferable anatomical competence.

For universities, the appropriate response is not to abandon written assessment, but to protect high-stakes administration and triangulate evidence through supervised, oral, constructed-response, visuospatial, and practical tasks. Generative AI may also support formative anatomy learning when its outputs are critically evaluated against authoritative anatomical resources and faculty guidance rather than accepted as an unquestioned source.

Future benchmarking should report the following six domains of the proposed MATRIX-Anatomy framework: Model identity and temporal lock; Assessment provenance and curricular

authenticity; Testing protocol and repeatability; Reference standard and human comparator; Input modality, item format, and cognitive demand; and eXternal validity and educational interpretation. Studies should also distinguish first-attempt accuracy, repeated-run stability, and feedback-assisted performance. MATRIX-Anatomy provides a transparent structure for comparing and interpreting this rapidly evolving evidence base, but formal external validation is required before it is used as a standardized appraisal instrument. Accordingly, each reported score should be understood as evidence about a defined system tested under defined conditions at a specific point in time, not as a timeless measure of "AI competence."

Acknowledgments

The authors have nothing to report.

Funding

The authors have nothing to report.

Ethics Statement

The authors have nothing to report.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Al-Khater, K. M. K. 2025. "Comparative Assessment of Three AI Platforms in Answering USMLE Step 1 Anatomy Questions or Identifying Anatomical Structures on Radiographs." *Clinical Anatomy* 38, no. 2: 186–199. https://doi.org/10.1002/ca.24243.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association.

Baethge, C., S. Goldbeck-Wood, and S. Mertens. 2019. "SANRA: A Scale for the Quality Assessment of Narrative Review Articles." *Research Integrity and Peer Review* 4: 5. https://doi.org/10.1186/s41073-019-0064-8.

Bolgova, O., and V. Mavrych. 2025. "Evolution of AI in Anatomy Education Study Based on Comparison of Current Large Language Models Against Historical ChatGPT Performance." *Scientific Reports* 15: 37545. https://doi.org/10.1038/s41598-025-22437-w.

Campbell, M., J. E. McKenzie, A. Sowden, et al. 2020. "Synthesis Without Meta-Analysis (SWiM) in Systematic Reviews: Reporting Guideline." *BMJ* 368: l6890. https://doi.org/10.1136/bmj.l6890.

Chaba-Karnaś, W., N. Kozioł, N. Kalita, et al. 2026. "Performance of Seven Large Language Models on Anatomy Examination Questions." *Clinical Anatomy*. https://doi.org/10.1002/ca.70076.

Choudhury, B., and A. Freemont. 2017. "Assessment of Anatomical Knowledge: Approaches Taken by Higher Education Institutions." *Clinical Anatomy* 30, no. 3: 290–299. https://doi.org/10.1002/ca.22835.

Connolly, S. A., T. H. Gillingwater, C. Chandler, et al. 2018. "The Anatomical Society's Core Anatomy Syllabus for Undergraduate Nursing." *Journal of Anatomy* 232, no. 5: 721–728. https://doi.org/10.1111/joa.12782.

Diamond, I. R., R. C. Grant, B. M. Feldman, et al. 2014. "Defining Consensus: A Systematic Review Recommends Methodologic Criteria for Reporting of Delphi Studies." *Journal of Clinical Epidemiology* 67, no. 4: 401–409. https://doi.org/10.1016/j.jclinepi.2013.12.002.

Eachempati, P., R. Komattil, and A. Arakala. 2025. "Should Oral Examination Be Reimagined in the Era of AI?" *Advances in Physiology Education* 49, no. 1: 208–209. https://doi.org/10.1152/advan.00191.2024.

Erbek, E., S. Çalışkan Pala, G. Bolatlı, and F. Çavuş. 2026. "Artificial Intelligence in Anatomy Education: A Systematic Review of ChatGPT's Effectiveness as a Learning Tool." *Baylor University Medical Center Proceedings* 39, no. 2: 315–323. https://doi.org/10.1080/08998280.2026.2612683.

Gallifant, J., M. Afshar, S. Ameen, et al. 2025. "The TRIPOD-LLM Reporting Guideline for Studies Using Large Language Models." *Nature Medicine* 31, no. 1: 60–69. https://doi.org/10.1038/s41591-024-03425-5.

Ganapathy, A., and P. Kaushal. 2025. "Cognitive Domain Assessment of Artificial Intelligence Chatbots: A Comparative Study Between ChatGPT and Gemini's Understanding of Anatomy Education." *Medical Science Educator* 35, no. 3: 1295–1304. https://doi.org/10.1007/s40670-025-02303-0.

Gürses, Ö. A., and İ. Ceylan. 2026. "Consistency Over Accuracy: Run-To-Run Stability of Contemporary Large Language Models on Turkish Curriculum-Aligned Theoretical Anatomy Multiple-Choice Questions." *BMC Medical Education* 26: 285. https://doi.org/10.1186/s12909-026-08656-3.

Kane, M. T. 2013. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50, no. 1: 1–73. https://doi.org/10.1111/jedm.12000.

Kasneci, E., K. Seßler, S. Küchemann, et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." *Learning and Individual Differences* 103: 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Keskin, A., and T. Aygün. 2025. "Determining the Reliability of Artificial Intelligence Programs in Medical Faculty Board Exams." *Pakistan Journal of Medical Sciences* 41, no. 12: 3354–3358. https://doi.org/10.12669/pjms.41.12.12855.

Li, Y., Y. Guo, F. Guerin, and C. Lin. 2024. "An Open-Source Data Contamination Report for Large Language Models." In *Findings of the Association for Computational Linguistics: EMNLP 2024*, edited by Y. Al-Onaizan, M. Bansal, and Y. N. Chen, 528–541. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.30.

Lodge, J. M., S. Howard, M. Bearman, and P. Dawson, & Associates. 2023. *Assessment Reform for the Age of Artificial Intelligence*. Tertiary Education Quality and Standards Agency.

Lu, M., J. Cheng, and V. Gopalan. 2026. "Performance of Multimodal Large Language Models on Image-Based Surgical Anatomy, Anatomical Pathology, and Radiology Questions." *Anatomical Sciences Education* 19: 1288–1299. https://doi.org/10.1002/ase.70256.

Matthan, J., M. Cobb, S. McHanwell, B. J. Moxham, and G. M. Finn. 2020. "The Anatomical Society's Core Anatomy Syllabus for Dental Undergraduates." *Journal of Anatomy* 236, no. 4: 737–751. https://doi.org/10.1111/joa.13133.

Mavrych, V., P. Ganguly, and O. Bolgova. 2025. "Using Large Language Models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in Gross Anatomy Course: Comparative Analysis." *Clinical Anatomy* 38, no. 2: 200–210. https://doi.org/10.1002/ca.24244.

McHugh, M. L. 2012. "Interrater Reliability: The Kappa Statistic." *Biochemia Medica* 22, no. 3: 276–282. https://doi.org/10.11613/BM.2012.031.

Mehrabanian, M., and Y. Zariat. 2023. "ChatGPT Passes Anatomy Exam." *British Dental Journal* 235, no. 5: 295. https://doi.org/10.1038/s41415-023-6286-7.

Melczewski, P., M. Chowaniec, W. Larysz, et al. 2026. "Accuracy of ChatGPT-4o in Identifying Anatomical Structures on Cadaveric Images: A Practical Anatomy Examination Study." *Clinical Anatomy*. https://doi.org/10.1002/ca.70115.

Miao, F., and W. Holmes. 2023. *Guidance for Generative AI in Education and Research*. UNESCO. https://doi.org/10.54675/EWZM9535.

Moher, D., K. F. Schulz, I. Simera, and D. G. Altman. 2010. "Guidance for Developers of Health Research Reporting Guidelines." *PLoS Medicine* 7, no. 2: e1000217. https://doi.org/10.1371/journal.pmed.1000217.

Nahir, M., A. Kasap, and B. Şahin. 2026. "System-Based Comparison of the Knowledge Level of Popular AI Chatbots on Human Anatomy: A Multiple-Choice Exam Analysis of GPT-4.1, DeepSeek, Co-Pilot, and Gemini Models." *Surgical and Radiologic Anatomy* 48: 1. https://doi.org/10.1007/s00276-025-03769-8.

Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences of the United States of America* 115, no. 11: 2600–2606. https://doi.org/10.1073/pnas.1708274114.

Paslı, B., and C. Günenç Beşer. 2026. "Is Artificial Intelligence Getting Better at Anatomy? A Two-Year Review of ChatGPT's Free Public Versions." *Anatomical Sciences Education* 19: 1279–1287. https://doi.org/10.1002/ase.70255.

Perkins, M., L. Furze, J. Roe, and J. MacVaugh. 2024. "The Artificial Intelligence Assessment Scale (AIAS): A Framework for Ethical Integration of Generative AI in Educational Assessment." *Journal of University Teaching and Learning Practice* 21, no. 6: 6. https://doi.org/10.53761/q3azde36.

Polit, D. F., and C. T. Beck. 2006. "The Content Validity Index: Are You Sure You Know What's Being Reported? Critique and Recommendations." *Research in Nursing & Health* 29, no. 5: 489–497. https://doi.org/10.1002/nur.20147.

Rethlefsen, M. L., S. Kirtley, S. Waffenschmidt, et al. 2021. "PRISMA-S: An Extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews." *Systematic Reviews* 10: 39. https://doi.org/10.1186/s13643-020-01542-z.

Rowland, S., K. Ahmed, D. C. Davies, et al. 2011. "Assessment of Anatomical Knowledge for Clinical Practice: Perceptions of Clinicians and Students." *Surgical and Radiologic Anatomy* 33, no. 3: 263–269. https://doi.org/10.1007/s00276-010-0748-8.

Runyon, C. R., M. A. Paniagua, F. A. Rosenthal, et al. 2024. "SHARP (SHort Answer, Rationale Provision): A New Item Format to Assess Clinical Reasoning." *Academic Medicine* 99, no. 9: 976–980. https://doi.org/10.1097/ACM.0000000000005769.

Sarantopoulos, A., Z. D. Pana, A. Larentzakis, et al. 2026. "ChatGPT-4.0 and Medical Students: A Recognition-Gated Comparative Evaluation on Image-Based Medical Examinations." *Journal of Medical Education and Curricular Development* 13: 23821205261449387. https://doi.org/10.1177/23821205261449387.

Singal, A., and S. Goyal. 2025. "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." *Surgical and Radiologic Anatomy* 47: 193. https://doi.org/10.1007/s00276-025-03707-8.

Smith, C. F., G. M. Finn, J. Stewart, et al. 2016. "The Anatomical Society Core Regional Anatomy Syllabus for Undergraduate Medicine." *Journal of Anatomy* 228, no. 1: 15–23. https://doi.org/10.1111/joa.12405.

Ullah, R., M. S. Shaikh, N. Shahani, M. A. Lone, M. A. Fareed, and M. S. Zafar. 2026. "Comparing ChatGPT and Dental Students' Performance in an Introduction to Dental Anatomy Examination: A Cross-Sectional Study." *European Journal of Dentistry* 20, no. 1: 287–294. https://doi.org/10.1055/s-0045-1808254.

van der Vleuten, C. P. M., and L. W. T. Schuwirth. 2005. "Assessing Professional Competence: From Methods to Programmes." *Medical Education* 39, no. 3: 309–317. https://doi.org/10.1111/j.1365-2929.2005.02094.x.

Vuma, S., and B. Sa. 2015. "Evaluation of the Effectiveness of Progressive Disclosure Questions as an Assessment Tool for Knowledge and Skills in a Problem-Based Learning Setting Among Third-Year Medical Students at the University of the West Indies, Trinidad and Tobago." *BMC Research Notes* 8: 673. https://doi.org/10.1186/s13104-015-1603-0.

Woodley, S. J., R. A. Green, and A. L. Webb. 2023. "A Core Musculoskeletal Anatomy Syllabus for Undergraduate Physical Therapy Student Education." *Clinical Anatomy* 36, no. 2: 190–223. https://doi.org/10.1002/ca.23953.

Xiong, M., Z. Hu, X. Lu, et al. 2024. "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs." *International Conference on Learning Representations* 2024: 23650–23678.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Appendix S1:** Supporting Information.