PHILIPPINE ENGINEERING JOURNAL
PEJ 2026; Vol. 47, No. 1:1-18

*Received: 22 November 2025*
*Received in Revised Form: 28 January 2026*
*Accepted: 13 April 2026*

# Human-Referenced Evaluation of Large Language Models on Course-Embedded Mining Engineering Multiple-Choice Questions

***Christian G. Arranz***
*Department of Mining, Metallurgical and Materials Engineering,*
*University of the Philippines Diliman, Quezon City, Metro Manila, Philippines*
*Corresponding author: cgarranz@up.edu.ph*

***Abstract*** *– Generative AI models have shown varying levels of success in answering academic and professional exam questions across different fields, such as medicine, but less so in engineering. Previous studies highlight their strengths in handling straightforward, fact-based questions while also pointing out difficulties with more nuanced and complex queries. This study evaluates eight different versions of large language models (LLMs): DeepSeek v3.2, Claude Sonnet 4.5, Google Gemini (1.0, 1.5, and 2.5), and ChatGPT (4.0, 4o, and 5), compared to 100 undergraduate students answering 100 multiple-choice questions in mining engineering. Questions were classified by Bloom's Taxonomy into objective types (Remembering and Understanding) and numerical problem-solving types (Applying and Analyzing). Results show a performance hierarchy, with the latest models- ChatGPT 5, Gemini 2.5, and Claude Sonnet 4.5- matching or surpassing high-performing students, reaching the 80th to 84th percentile. ChatGPT 5 outperformed earlier versions and other models in accuracy and consistency. Statistical analysis using McNemar's and chi-square tests revealed no significant association between model performance and question type for all LLMs except ChatGPT 4.0 (p=0.0346), indicating consistent accuracy across categories. Despite comparable performance on certain questions, these AI models exhibit significant limitations in understanding complex, context-dependent problems that require higher-order thinking skills, leading to inconsistencies in generating accurate responses. The study also discusses the implications of AI performance in academic exams for future mining engineers, emphasizing the need to evolve assessment strategies in engineering education and advocating for the integration of AI as a supplementary instructional tool while maintaining academic integrity.*

***Keywords****: large language models, generative artificial intelligence, mining engineering education, engineering education*

## I. INTRODUCTION

The rapid development and ubiquity of Generative Artificial Intelligence (GenAI), specifically Large Language models (LLMs), have caused an unprecedented and significant shift in the world, particularly in work and education [1-3]. Advanced LLMs such as GPT-5 [4], Anthropic Claude 4 [5], DeepSeek [6] and Google Gemini 2.5 [7], which are built on massive datasets and transformer-based architectures that utilize deep learning and natural language processing, now demonstrate advanced abilities such as generating coherent human-like text and code, summarizing information, solving problems, understanding context, and answering questions. They even outperform human participants in specific scenarios,

particularly in multiple-choice questions (MCQs) [2, 8-10]. This development has numerous implications, especially in higher education and professional assessments and certifications where MCQs remain widely used across numerous domains, including medical education and licensing examinations, due to their ease of administration, standardized format, and broad curricular coverage [11].

Recent benchmarking studies show that newer LLMs can achieve accuracy rates comparable to, and in some cases surpassing, average student performance in various standardized exams, although performance still varies across different models [2, 3, 12-16]. However, studies also show that performance varies across subject domains and question types [17,18]; while LLMs perform well on straightforward factual questions, they may still struggle with complex mathematical concepts, nuanced or scenario-based questions, and highly specialized, context-dependent problem-solving in engineering [19-21].

To evaluate cognitive thresholds, the revised Bloom's Taxonomy is often used to categorize the cognitive complexity of assessments and to analyze LLMs' abilities across different levels [22]. It comprises a knowledge dimension and a cognitive process dimension, which consists of six levels: remember, understand, apply, analyze, evaluate, and create [22, 23]. Studies have also shown that AI models demonstrate proficiency with objective-type questions and numerical problem-solving, but may struggle with questions that require interpreting visual information, such as diagrams or graphs [24-26].

Although evidence is accumulating across engineering and adjacent professional fields, course-embedded, discipline-specific studies remain sparse, especially in Philippine engineering contexts that use syllabus-aligned items and cognitive-level analyses [27]. Because of the rapid evolution of GenAI LLMs and the inconsistent performance across models, domains, and tasks, continuous benchmarking and performance monitoring across multiple LLM platforms and successive versions remain essential [2, 3, 11, 18, 28].

While AI's abilities have been extensively studied in fields like medicine, computer science, accounting and other standardized tests worldwide, there remains a significant knowledge gap in understanding its applicability and effectiveness in specialized engineering fields, such as mining engineering. Engineering disciplines differ fundamentally due to their reliance on applied science, complex problem-solving skills, and mathematical concepts [19], [21, 29, 30]. Accordingly, this study examines whether the observed difficulty with numerical problems and applied technical reasoning persists in mining engineering, a specialized engineering discipline concerned with the extraction of valuable minerals from the earth across all phases, with due consideration for economic, social, and environmental sustainability [31]. From tertiary-level assessments to licensure examinations in Mining Engineering, MCQs are commonly used to evaluate fundamental concepts and students' understanding of mining engineering principles because they can be easily marked, efficiently cover a wide range of topics, and provide a standardized, objective assessment format. At the same time, empirical evidence suggests that LLMs perform better on lower-order cognitive tasks (Remembering and Understanding), and may be less effective in evaluating higher-order thinking or problem-solving skills, which are crucial in engineering [8, 32, 33]. Mining engineering provides a useful case because it combines domain-specific concepts and integrative yet specialized

applied technical reasoning requirements, such as handling geological and related data, understanding specialized concepts, and applying engineering, scientific, and mathematical principles within the field, that have not yet been rigorously assessed in previous research.

Thus, this study contributes course-embedded evidence on contemporary LLM performance in a specialized engineering assessment context, particularly mining engineering MCQs. By comparing eight contemporary LLMs to human undergraduate students in answering mining engineering MCQs across Bloom's Taxonomy cognitive levels, the main contributions of this study are as follows: First, it provides a course-embedded, domain-specific empirical evaluation of contemporary LLM performance in mining engineering using a curated 100-item syllabus-aligned MCQ set, interpreted relative to human undergraduate student performance. Secondly, it contributes comparative item-level statistical evidence on version-to-version and cross-model LLM performance differences. And third, it provides evidence on whether LLM performance varies across Bloom's Taxonomy cognitive levels in mining engineering MCQs. By focusing on mining engineering as a specialized technical field, the study addresses a critical gap in the evidence on AI performance in discipline-specific assessment evaluations in a higher education context.

## II. METHODOLOGY

This comparative study evaluated eight publicly available LLM versions on a 100-item mining engineering MCQ set that includes model selection, item classification, response collection and item- and level-based statistical comparisons.

### *2.1 Study Design, LLMs, and Human Reference Group*

The study evaluated eight (8) publicly available and stable LLM versions accessible during the 2025 data collection period. Table 1 below summarizes the providers and model versions used, including the reported release dates. Recent data-driven traffic studies of AI chatbots identify these models among the most-used chatbot platforms worldwide, with ChatGPT leading by annual web visits to maximize the real-world relevance and cross-provider coverage [34, 35].

**Table 1.** Summary of LLMs evaluated (stable releases available in 2025)

| Provider | Model family | Version used | Reported release (month/year) | Remarks on inclusion |
|---|---|---|---|---|
| DeepSeek | DeepSeek | V3.2 | Sep 2025 | Major stable release available during the study window |
| Anthropic | Claude | Sonnet 4.5 | Sep 2025 | Major stable Sonnet variant available during the study window |
| Google | Gemini | 1.0 | Dec 2023 | Baseline Gemini generation used for cross-version comparison |
| Google | Gemini | 1.5 | Feb 2024 | Successor Gemini generation used for cross-version comparison |
| Google | Gemini | 2.5 | Mar 2025 | Latest Gemini generation during the study window |
| OpenAI | ChatGPT / GPT | GPT-4 | Mar 2023 | Baseline GPT generation used for cross-version comparison |
| OpenAI | ChatGPT / GPT | GPT-4o | May 2024 | Multimodal "omni" generation used for cross-version comparison |
| OpenAI | ChatGPT / GPT | GPT-5 | Aug 2025 | Latest GPT generation during the study window |

By incorporating various versions, the study aims to capture the evolutionary trajectory of the performance of these models. Analyzing responses across different iterations allows for an examination of improvements in accuracy, complexity, and handling of domain-specific queries.

To establish a human performance comparison, de-identified assessment records from one hundred (N=100) undergraduate students from Mining Engineering, Metallurgical Engineering (service-course takers), and Geology (elective takers) who completed the same Mining Engineering course assessment were analyzed. In this study, 'representativeness' is operationalized at the course level (i.e., a course-embedded reference context), not as population-level representativeness of all mining engineering students. Specifically, the cohort shares a common assessment context: the same course learning outcomes and syllabus alignment, the same instructor-authored answer key, and comparable online administration within the institutional learning management system. Student records were pooled across multiple sections and two academic-year cohorts (Academic Year 2022–2023 and Academic Year 2023–2024) to provide a more stable course-level human baseline [2].

The study utilized pre-existing, retrospective, anonymized academic performance data and did not involve direct student interaction or the use of personal identifiers. All student identities were fully anonymized prior to analysis. No personally identifiable information was collected, and the data were strictly processed to ensure that no personal identifiers remained in the final dataset. Data were compiled from institutional records using an anonymized scoring system. This strategy of using de-identified educational data is consistent with large-scale comparative LLM studies [2, 36, 37].

*2.2 Question Set and Classification*

The assessment material consisted of one hundred (100) MCQs drawn from the archived quiz and major-exam item banks of EM 10 (Introductory Mining Engineering) at the University of the Philippines Diliman. The archived items were authored by the course instructors and administered in prior offerings; items were not publicly released or posted online. The final 100-item MCQs set was selected from the pooled archive and reviewed to retain coverage across the main course modules spanning the entire mining life cycle and fundamental of mining engineering across several topics as shown in Table 2 below:

**Table 2.** Topic distribution of the 100-item mining engineering MCQ set

| **Topic Area** | **Number of MCQ Items** |
|---|---|
| Prospecting and Exploration | 15 |
| Economic Geology and Mining Geology | 15 |
| Mine Development | 15 |
| Underground Mining Methods and Design | 15 |
| Mining Laws and Recent Developments | 14 |
| Surface Mining Methods and Design | 13 |
| Rehabilitation and Environmental Protection | 13 |
| **Total** | 100 |

The assessment material consisted of 100 non-public and instructor-authored MCQs as determined by power analysis to ensure a comprehensive evaluation and reduce potential biases caused by selective sampling, using a significance level (α) of 0.05, a statistical power (1-β) of 0.80, and an assumed effect size (d) of 0.5 [38]. To ensure representative and broad coverage of the entire course curriculum topics, questions were drawn from a variety of course-embedded quizzes and exams [39, 40]. Items were classified into two revised Bloom's Taxonomy-based groupings: (1) Remembering and Understanding (Lower-Order Cognitive Skills) items that evaluated both the students' and AI models' ability to recall and comprehend foundational mining engineering concepts, such as definitions, basic principles, and essential processes; and (2) Applying and Analyzing (Higher-Order Cognitive Skills) items that focused on applying knowledge to solve numerical problems, analyze given scenarios, and make informed decisions. The categorization into Bloom's taxonomy levels was crucial in evaluating whether AI models could move beyond mere factual recall to more complex analytical problem-solving.

*2.3 Data Collection and Testing Procedure*

A standardized testing protocol was implemented to ensure comparability between AI models and human participants; it kept phrasing, terminology, and format constant to eliminate potential biases arising from different interpretations or language complexity, ensuring that all participants were tested under consistent conditions. In administering questions, the following was done:

For the human undergraduate student scores, the 100-item MCQ assessment was administered via the University Virtual Learning Environment (UVLe; Moodle) from students of the Introductory Mining Engineering course. Student raw scores (0–100) were obtained

from the UVLe gradebook, which automatically grades MCQs using the instructor-defined answer key and generates a total score upon submission. The exported gradebook totals were used as the student score distribution for descriptive statistics and percentile comparisons.

Each of the eight (8) versions of the LLM/AI chatbots: DeepSeek v3.2, Claude Sonnet 4.5, Google Gemini (1.0, 1.5, and 2.5), and ChatGPT (4.0, 4o, and 5), and the set of human students were asked the same set of 100 MCQs. All models received the same minimalistic or zero-shot prompt: "Give the letter of the answer to all of the following questions." No specific persona or role-based prompting, delimiters, scaffolding, or examples were utilized in conjunction with this minimal prompt [10, 41-43].

For each LLM platform, three (3) trials/replicates were conducted in separate chat sessions ('New Chat') such that each replicate contained no prior conversation context. This procedure was used to reduce within-session carryover and to sample the non-deterministic response variability of LLM outputs. Because commercial chat interfaces do not provide externally audited guarantees that sessions are fully independent beyond the supplied prompt/context, replicates were treated as independent at the prompt level and acknowledge residual cross-session dependence as a limitation [44-46]. The answers of the LLMs were then tabulated, and the mean averages of the 3 trials were determined for the direct comparison. Out of the 100 questions, 94 questions were in a purely text-based format, while the final 6 questions (items 95 to 100) had diagrams and were inputted as images, for which the image upload features of each LLM platform were used to enable the LLMs to answer.

Figure 1 summarizes the methodology workflow and separates the human reference pathway from the LLM testing pathway prior to convergence in the statistical analyses. The two pathways were then synthesized through descriptive comparison and item- and level-based inferential tests as detailed in Section 2.4.

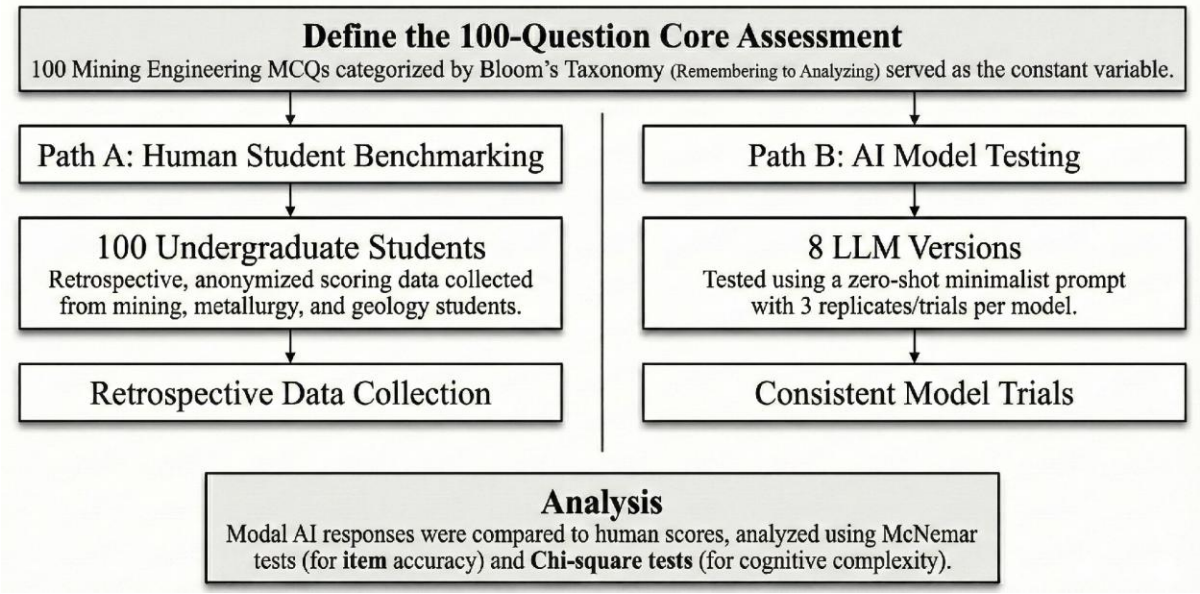



**Figure 1.** Methodology flow diagram

*2.4 Evaluation Metrics, Outcome Measures, and Statistical Analysis*

A series of metrics and statistical analyses were employed:

Accuracy was defined as the number of correctly answered items out of 100 (raw score, 0–100). For human students, raw scores were taken from the UVLe (Moodle) gradebook totals produced by automatic MCQ grading using the instructor-defined answer key. Question While the scores for each of the three individual trials and the mean scores for each AI LLM model were recorded for the descriptive statistical analysis, the definitive answer for each question per LLM version was determined by the mode of the three responses, which reflects the most common answer choice across the three attempts. This modal response score for each question was used for subsequent statistical analysis as described below. The human cohort completed the assessment once under standard test conditions; thus, human variability is captured across individuals (n=100) rather than across repeated attempts by the same individual. In contrast, repeated LLM trials sample within-model stochasticity under identical prompts. To avoid interpreting replication as 'multiple chances,' per-trial scores and the mean for descriptive benchmarking were reported, but the per-item 'definitive' LLM response was defined using the modal (majority-vote) answer across trials for item-level inferential tests.

Pairwise differences in item-by-item correctness between LLMs were evaluated using McNemar's test for paired nominal data. For each model pair, the test statistic ($X^2$) was computed from the discordant counts as shown in Equation 1:

$$X^2 = \frac{(b-c)^2}{b+c} \quad (1)$$

Where b is the number of items answered correctly by Model A but incorrectly by Model B, and c is the number of items answered correctly by Model B but incorrectly by Model A [47,48].

Associations between LLM performance (correct/incorrect) and Bloom's Taxonomy level were tested separately for each model using a chi-square test of independence:

$$X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (2)$$

Where $O_{ij}$ is the observed frequency of correct/incorrect answers for cognitive level j, and $E_{ij}$ is the expected frequency under the assumption of independence.

## III. RESULTS AND DISCUSSION

### *3.1 Establishing a Human-contextualized Performance Baseline*

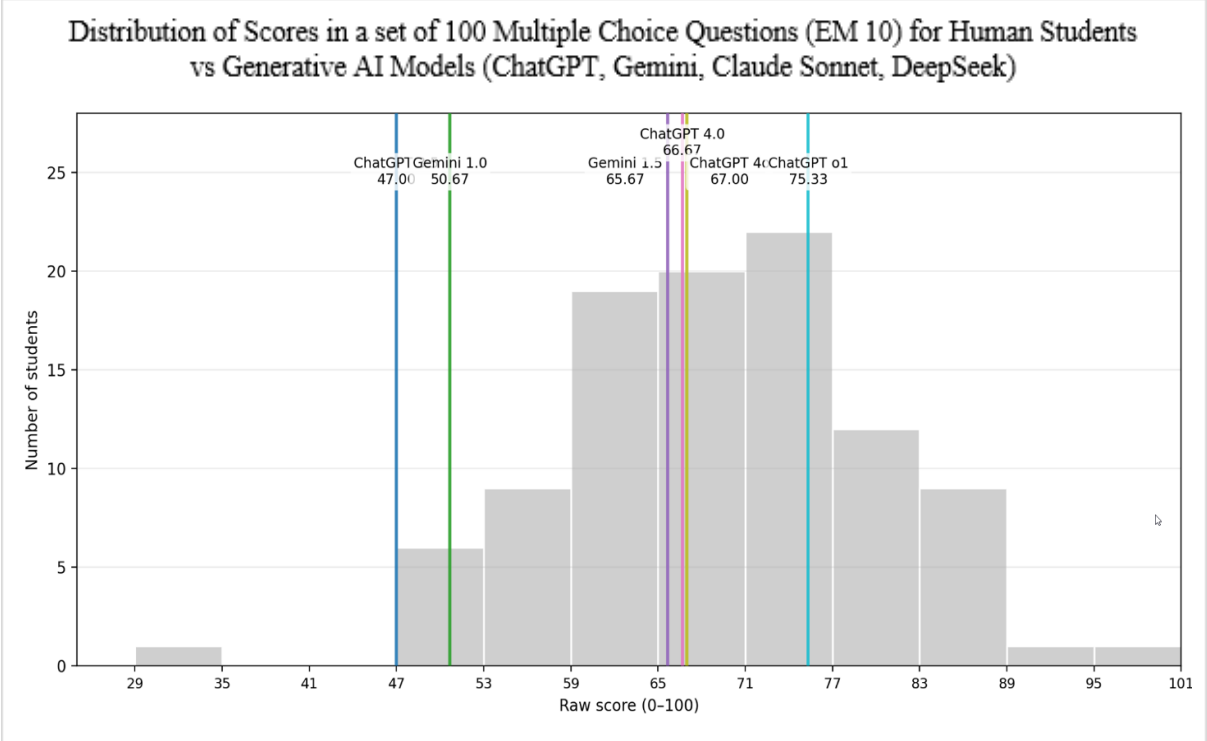



**Figure 1.** Histogram Showing the Distribution of Raw Scores of Human Students and the Mean Scores of Generative AI Models (Google Gemini and ChatGPT Versions)

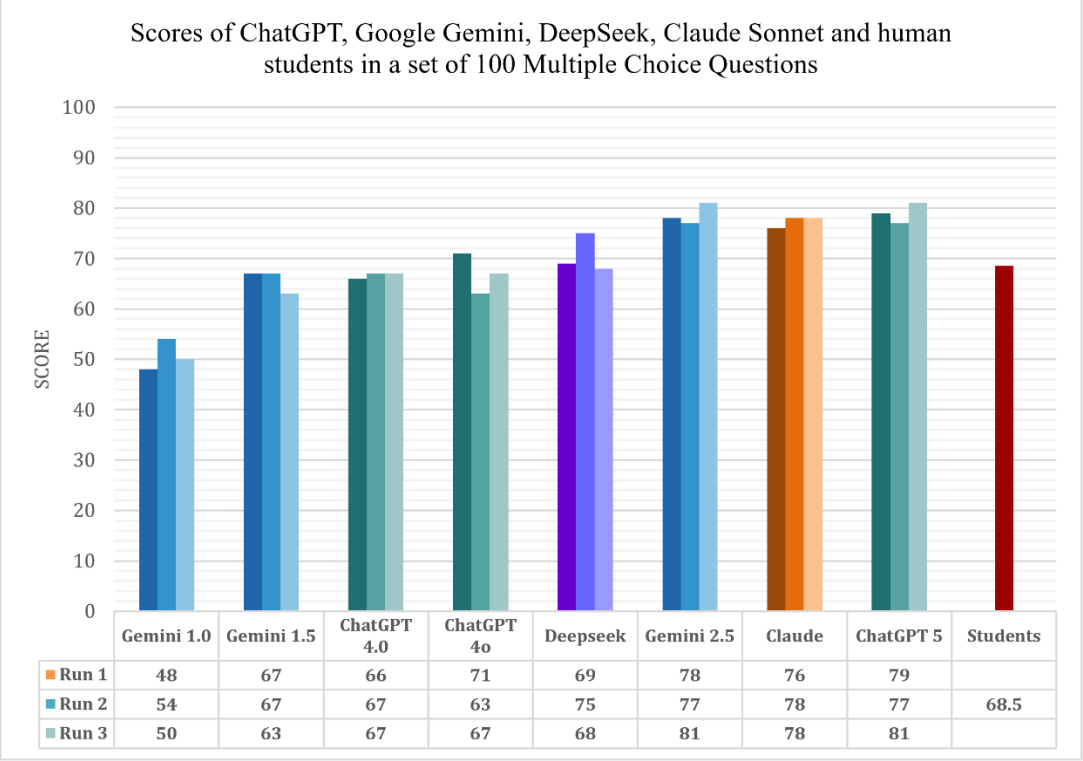
Figure 1 illustrates the distribution of human students (scores) compared to the eight Generative AI models on a 100-item MCQ exam in mining engineering, presented as a frequency distribution of the students' raw scores, while the mean scores (μ) of the AI models are plotted across these score ranges for comparison. This format emphasizes the AI models' mean scores, enabling a straightforward comparison with the full range of student performance. Because the item bank was drawn from a single course (EM 10) at a single institution (UP Diliman), the findings should be interpreted as course-embedded evidence rather than a comprehensive representation of all mining engineering curricula. The mean scores covered a broad range, from a minimum of 50.67 (Gemini 1.0) to a maximum of 79.00 (ChatGPT 5), while human scores ranged from 29 to 95, highlighting significant variability in individual student performance. Table 1 similarly presents the percentile rankings of the mean scores of multiple trials for each Generative AI model in comparison to human student scores from a set of 100 MCQs.

Gemini 1.0 (μ = 50.67) appears near the lower student deciles, placing it in the 5th percentile and making it comparable only to the lowest-performing students. This marks the lower boundary of model performance. A second group includes DeepSeek v3.2 (μ = 70.67), ChatGPT 4o (μ = 67.00), ChatGPT 4.0 (μ = 66.67), and Gemini 1.5 (μ = 65.67), placing them in the middle of student scores, performing near or slightly above the median student level.

These are similar to "average-to-above-average" students, with mean scores generally in the mid-60s to low-70s. Three models cluster at the top: ChatGPT 5 (μ = 79.00), Gemini 2.5 (μ = 78.67), and Claude Sonnet 4.5 (μ = 77.33). All sit in the upper tail of the student score distribution. These latest iterations cluster in the top quintile (20%) of the human scores, demonstrating proficiency similar to high-achieving human learners. ChatGPT 5 ranked at the 84$^{th}$ percentile, exceeding the mean score of human students of 68.5 as shown in Figure 2. Within this dataset, it was among the highest performing models evaluated.

The data clearly illustrate a rapid evolutionary gain within model families. Google's models jumped 78-percentile-points from Gemini 1.0 (5th percentile) to Gemini 2.5 (83rd percentile) in approximately 15 months. Similarly, OpenAI's models advanced from the 43rd percentile (ChatGPT 4.0) to the 84th percentile (ChatGPT 5). In practical terms, the newest systems perform like upper-quartile students, whereas early-generation models perform like lower-quartile students. That ranking relationship aligns with discipline-related benchmarking, where later model generations outperform earlier ones.

Across three trials, within-model variability was relatively low, with per-model trial standard deviations ranging from 0.6 to 4.0 points out of 100. This indicates that the ranking patterns reported in the benchmarking and McNemar comparisons are stable/mostly stable with respect to stochastic response variation, although replication remains an approximation of independence rather than a guarantee.




| | Gemini 1.0 | Gemini 1.5 | ChatGPT 4.0 | ChatGPT 4o | Deepseek | Gemini 2.5 | Claude | ChatGPT 5 | Students |
|---|---|---|---|---|---|---|---|---|---|
| Run 1 | 48 | 67 | 66 | 71 | 69 | 78 | 76 | 79 | |
| Run 2 | 54 | 67 | 67 | 63 | 75 | 77 | 78 | 77 | 68.5 |
| Run 3 | 50 | 63 | 67 | 67 | 68 | 81 | 78 | 81 | |

**Figure 2.** Comparative Scores of Generative AI Models
(DeepSeek v3.2, Claude Sonnet 4.5, Google Gemini 1.0, Gemini 1.5, Gemini 2.5, ChatGPT 4.0, ChatGPT 4o, and ChatGPT 5), and the mean score of Human Students in a set of 100 MCQs

*3.2 Comparative Statistical Analysis of AI LLM Performance*

The results of the McNemar test and p-values that support the performance stratification are shown in Table 3. Entries indicate whether the two models make systematically different errors (discordant pairs), not which one is higher.

**Table 3.** Pairwise comparisons of Large Language Models for all questions

| | Gemini 1.0 | Gemini 1.5 | ChatGPT 4.0 | ChatGPT 4o | ChatGPT 5 | Gemini 2.5 | DeepSeek v3.2 | Claude 4.5 |
|---|---|---|---|---|---|---|---|---|
| Gemini 1.0 | - | **0.008** | **0.018** | **0.012** | **0.000** | **0.000** | **0.000** | **0.000** |
| Gemini 1.5 | 0.008 | - | 0.803 | 1.000 | **0.019** | **0.034** | 0.230 | **0.044** |
| ChatGPT 4.0 | 0.018 | 0.803 | - | 1.000 | **0.008** | **0.031** | 0.095 | **0.014** |
| ChatGPT 4o | 0.012 | 1.000 | 1.000 | - | **0.016** | **0.046** | 0.170 | **0.022** |
| ChatGPT 5 | 0.000 | 0.019 | 0.008 | 0.016 | - | 0.803 | 0.383 | 0.803 |
| Gemini 2.5 | 0.000 | 0.034 | 0.031 | 0.046 | 0.803 | - | 0.606 | 1.000 |
| DeepSeek v3.2 | 0.000 | 0.230 | 0.095 | 0.170 | 0.383 | 0.606 | - | 0.606 |
| Claude 4.5 | 0.000 | 0.044 | 0.014 | 0.022 | 0.803 | 1.000 | 0.606 | - |

*Note: p-values < 0.05 (bolded) indicate a statistically significant difference in the proportion of correct answers between the paired models.*

Gemini 1.0 is statistically confirmed as an outlier, with all pairwise comparisons involving Gemini 1.0 yielding statistically significant p-values (all $p \leq 0.018$; several $< 0.001$), indicating a distinct error pattern on this item set. Its performance was notably worse than that of all other tested models, aligning with its 5th-percentile descriptive ranking. Gemini 1.5, ChatGPT 4.0, and ChatGPT 4o are statistically indistinguishable from one another in their item-by-item accuracy. For example, comparisons within this group produced high p-values (e.g., Gemini 1.5 vs. ChatGPT 4.0, $p = 0.803$). However, this group is statistically different than the top tier, showing small but significant p-values when compared with models like ChatGPT 5 ($p = 0.019$, 0.008, and 0.016, respectively) and Gemini 2.5 ($p = 0.034$, 0.031, and 0.046, respectively).

A key finding in evaluating modern LLMs is that the top four models, ChatGPT 5, Gemini 2.5, Claude 4.5, and DeepSeek v3.2, are also statistically indistinguishable on this specialized test set. Pairwise comparisons within this group consistently show non-significant p-values (e.g., Gemini 2.5 vs. Claude 4.5, $p = 1.000$; ChatGPT 5 vs. DeepSeek v3.2, $p = 0.383$). This suggests that, for this domain specific MCQ task, current leading LLMs appear to have reached similar performance capabilities, where differences in accuracy are minor and statistically unreliable.

These results generally reflect the rapid growth in LLM capabilities, where newer or architecturally different models often outperform older or less advanced ones. For example, the significant differences observed between Gemini 1.0 and most other listed models align with the overall trend seen in studies: older, base-level LLMs tend to perform significantly worse than current-generation models, especially in specialized fields.

*3.3 Evaluation of AI LLM Performance by Cognitive Complexity*

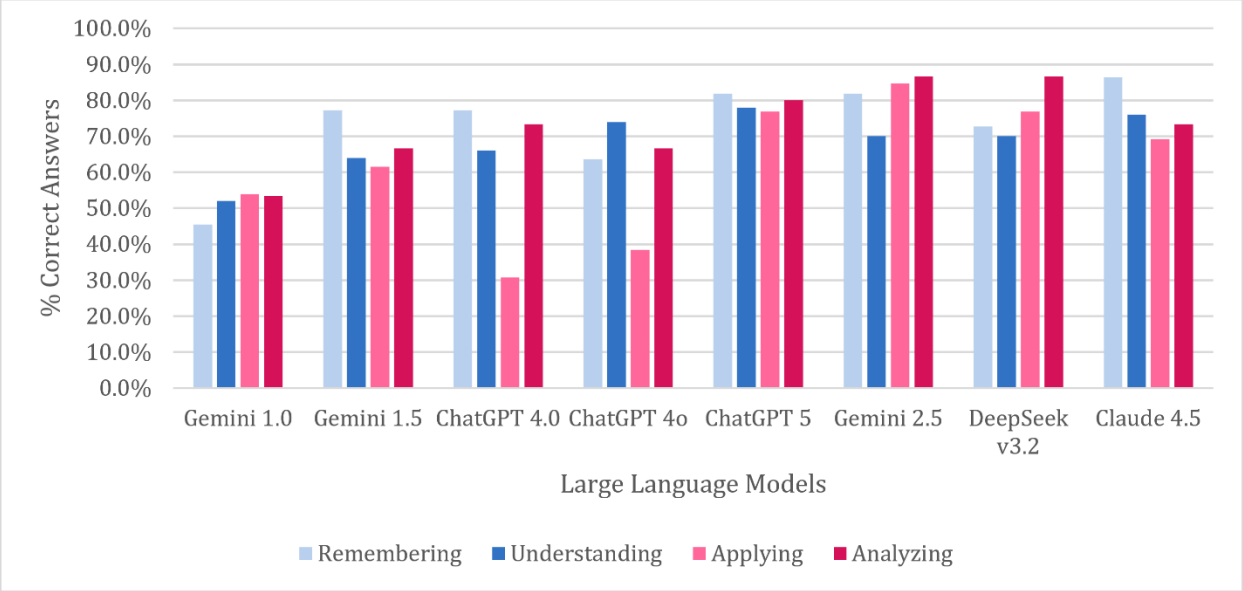



**Figure 3.** Large Language Models' accuracy by Bloom's Taxonomy levels on 100 MCQs

The results of the bivariate Chi-Squared ($\chi^2$) Test for Independence are presented in Table 4:

**Table 4.** Chi-Squared ($\chi$2) Test of Independence Results for LLM Performance vs. Bloom's Taxonomy Cognitive Level (Note: p-values < 0.05 (bolded) indicate a rejection of the null hypothesis and a significant association)

| LLM vs Bloom's Taxonomy | Chi-Squared Test p-value | Modal response score |
|---|---|---|
| Gemini 1.0 | 0.9473 | 51 |
| Gemini 1.5 | 0.6986 | 67 |
| ChatGPT 4.0 | **0.0346** | 65 |
| ChatGPT 4o | 0.1177 | 66 |
| ChatGPT 5 | 0.981 | 79 |
| Gemini 2.5 | 0.4091 | 77 |
| DeepSeek v3.2 | 0.6274 | 74 |
| Claude 4.5 | 0.6427 | 77 |

For seven of the eight models, performance was found to be statistically independent of cognitive complexity ($p > 0.05$), with ChatGPT 4.0 as the sole exception. The models' accuracy, whether generally low (Gemini 1.0 at $p = 0.9473$) or high (ChatGPT 5 at $p = 0.981$), remained consistent across all tested Bloom's Taxonomy levels (e.g., Remembering, Understanding, Applying, and Analyzing). The results suggest that, for these models,

performance did not vary strongly by nominal Bloom category under the present design; however, non-significance should not be interpreted as equivalence across levels. Furthermore, this challenges the assumption that modern LLMs are primarily proficient only in factual recall (LOCS).

The sole exception was ChatGPT 4.0, which exhibited a significant association between performance and cognitive complexity ($p = 0.0346$). This indicates that ChatGPT 4.0's performance was significantly correlated with the question's cognitive level, suggesting an uneven ability profile across the taxonomy. This dependency was notably absent in its successors (ChatGPT 4o, $p = 0.1177$; ChatGPT 5, $p = 0.981$), strongly indicating it was an architectural limitation addressed in later versions. However, without analyzing standardized residuals, the specific level(s) (e.g., applying or analyzing) causing this association cannot be identified. Additionally, caution is advised because the significance of $p = 0.0346$ depends on whether it would remain significant after adjusting for multiple tests.

*3.4 Interpretations and Discussions*

Taken together, the generational improvement as shown by percentile shifts, the lack of significant statistical differences among top models, and the largely non-significant Bloom-level associations, show performance comparability across the different models. There is also an observed consistency with emerging findings from other professional and technical assessment studies, where later model versions outperform earlier ones [49]. In this respect, this study extends the literature by providing a transparent, comparable performance baseline for contemporary LLMs in a highly specialized, non-public field. While LLMs have been widely benchmarked in areas like medicine, computer science, and general assessments, discipline-specific, course-embedded studies are still rare, particularly in technical fields that demand applied science and mathematical concepts [28, 50, 51]. The study provides instructors with solid guidance on which model versions are interchangeable in accuracy (the top cluster), which are clearly outdated (Gemini 1.0), and whether the pedagogical approach of increasing the cognitive level (HOCS) effectively challenges LLMs, which is generally not the case, as shown by the Chi-squared tests. The finding that advanced LLMs now perform proficiently across all cognitive levels may suggest a significant improvement in their reasoning capabilities, extending beyond superficial knowledge retrieval. However, this proficiency likely reflects an architectural alignment between the closed-ended format of MCQs and the probabilistic retrieval mechanisms of LLMs. Consequently, there may be inherent limitations of the MCQ format itself that may be exploited by the increasing sophistication of LLMs, suggesting that MCQs may not reliably measure the true capabilities being assessed [52]. Furthermore, their high performance is not absolute, and while even the best models often achieve high accuracy rates, there still exists variability in their answers per individual run [53]. This further validates the stochastic nature of LLMs, which can yield inconsistent responses even under identical conditions. LLM output must be carefully vetted for both content accuracy and structural integrity, as underlying question characteristics have been shown to significantly influence response-explanation consistency [54].

The results suggest that closed-ended MCQ assessments, particularly in unsupervised or remote settings, may be increasingly vulnerable to the use of high-performing LLMs. In this dataset, increasing the nominal Bloom cognitive level alone did not consistently reduce model

accuracy, suggesting that shifting classification from lower- to higher-order categories may not, by itself, be sufficient to distinguish between unassisted human performance and LLM-assisted responses. These results support further evaluation of assessment formats that require constructed responses, applied judgement, or supervised performance, while broader claims regarding instructional or curriculum redesign remain beyond the scope of the present study.

*3.5. Implications for Assessment Practice*

Due to the results, educators are strongly encouraged to modify assessment methods by incorporating authentic, human-centered evaluations that are difficult to automate, such as project-based work and in-person oral exams. There is an increasing recognition of the potential of AI to enhance learning, particularly through personalized feedback and support, which could be valuable in challenging fields such as engineering [55]. Consequently, the proven high performance of Tier 1 models may be utilized as supplementary tools by integrating them into the curriculum as "AI assistants" or "co-pilots," focusing on the ethical and responsible use of these tools for complex engineering problem-solving.

Based on the strong generational signal and the achievement of cognitive stability in specialized domains, the field of engineering education is expected to experience a shift over the next years, where top systems will be largely interchangeable for recall and single-step applied tasks. Differentiation will shift toward complex task scaffolding, constrained reasoning, and item-bank craft, requiring programs to regularly conduct model and version audits and to design assessments that ensure performance parity among leading tools. Future teaching must focus on integrated human-AI workflows, evaluating students on their ability to critically assess, synthesize, and apply AI-generated outputs to new and complex engineering challenges, rather than just answering questions that an AI can already solve.

*3.6 Limitations*

Several limitations of the present LLM performance evaluation and study design should be recognized: First, this study was limited exclusively to MCQs, which were chosen for their objectivity, standardization, and efficiency. However, MCQs do not capture the full range of engineering competencies, particularly higher-order thinking skills of creating [56, 57]. In addition, only six test items (number 95 to 100) contained visual or non-textual components, such that the findings primarily reflect text-dominant items. Incorporating open-ended and visual/image based questions may provide a broader assessment of AI problem-solving performance.

Second, generalizability of findings is limited by the single-institution, course-embedded design and use of one Mining Engineering assessment context. Although the 100-item test provides a useful reference context for this course, results may not transfer to other institutions, cohorts, instructors, or mining engineering curricula.

Third, although commercial LLM chatbots do not retain or learn from individual user interactions directly for their permanent, static knowledge base, thereby supporting user privacy and data confidentiality [58], the literature does not provide formal, peer-reviewed evidence that commercial LLM chat sessions are fully independent beyond the prompt/context

provided [59-61]. To reduce possible short-term memory retention bias, a new chat session was initiated for each question, consistent with prior methodological practice [45, 60, 62].

Fourth, the study utilized a uniform, minimalist, zero-shot prompt (i.e., "Give the letter of the answer to all of the following questions") across all platforms and model versions to ensure consistency and rely solely on the models' core pre-trained knowledge. However, LLM performance is known to be sensitive to prompt phrasing and structure, including role-based prompting (e.g., instructing the LLM platform to "Act as an expert mining engineer"). Accordingly, the reported accuracies should not be interpreted as ceiling performance, as alternative prompting strategies may produce different results, especially for technical, or multi-step reasoning tasks [63-65].

Finally, the exploratory scope of the study forewent a full psychometric analysis based on Classical Test Theory (CTT) including the Item Difficulty Index (DIF I), the Item Discrimination Index (DI) since these are derived from human performance data and would require dedicated resources and broader field testing beyond the scope of the present comparative analysis. Emphasis of this study was placed on comparative model performance metrics rather than full item validation consistent with previous studies [66, 67]. A fuller psychometric evaluation would require a subsequent validation phase.

## IV. CONCLUSION AND RECOMMENDATION

Addressing the limited evidence on LLM performance in specialized technical assessments, this study provides a course-embedded and human-referenced evaluation of eight (8) LLMs' performance on a curated 100-item mining engineering MCQ assessment, showing how contemporary LLMs perform relative to undergraduate student performance at the overall, item-comparison and Bloom's taxonomy levels . Among the evaluated models, ChatGPT 5, Gemini 2.5, and Claude Sonnet 4.5 reach the upper segment of the student score distribution, matching the scores of high-performing students at approximately the 80th to 84th percentile. Item-level comparisons further showed that the top-performing models were statistically indistinguishable on the shared question set.

Additionally, for seven of the eight models, correctness did not differ significantly across the Bloom's Taxonomy category groupings used in this study. ChatGPT 4.0 was the only model that exhibited a significant association between correctness and question type. Taken together, these findings may suggest that more recent models can perform at levels comparable to high-performing students on this MCQ set, and have shown notable improvements, particularly in handling higher-order questions that involve numerical problem-solving. These findings should be interpreted within the limits of the study design: a single-institution, course-embedded, 100-item MCQ context from one mining engineering course, and require further validation for larger, multi-course, and multi-institution assessment sets. Accordingly, these results should not be generalized to other courses, open-ended problem formats, practical judgment tasks, or broader mining engineering competence without further validation. Additionally, the use of 28 pairwise McNemar comparisons without family-wise adjustment limits the precision of specific cross-tier comparisons. Finally, because development is rapid,

the present results should also be interpreted as a time-bound comparative evidence rather than an endpoint and will need further updates.

Future studies should extend the evidence base to a larger and more diverse student cohorts, multi-course and multi-institutional item banks, as well as other assessment types beyond MCQs, including open-ended, image-based, and applied complex problem-solving tasks. Further research could also examine the sensitivity of results to prompting strategies, session isolation conditions, repeated-trial scoring, and alternative statistical procedures for multiple comparisons to strengthen the precision and comparability of cross-model inferences.

**References:**


[1] Nikolic S, Sandison C, Haque R, Daniel S, Grundy S, Belkina M, Lyden S, Hassan G, Neal P, 2024. ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering. Australasian Journal of Engineering Education. 29(2):126–153. doi: 10.1080/22054952.2024.2372154.

[2] Sabri S, Saleh M, Hazrati P, Marchant K, Misch J, Kumar P, Wang H, Barootchi S. 2025. Performance of three artificial intelligence (AI)-based large language models in standardized testing; implications for AI-assisted dental education. J. Periodontal Res. 60(2):121–133, Feb. 2025, doi: 10.1111/jre.13323.

[3] Sallam M, Irshaid A, Snygg J, Albadri R, Sallam M. 2025. Rapid evolution of large language models in medical education: comparative performance of ChatGPT-3.5, ChatGPT-5, and DeepSeek on medical microbiology MCQs. Contemporary Education and Teaching Research. 6(8):295–309. doi: 10.61360/BONICETR252018770801

[4] OpenAI. 2025. Retrieved from https://openai.com/index/introducing-gpt-5/ on 06 Nov 2025.

[5] Anthropic. 2025. Retrieved from https://www.anthropic.com/news/claude-4 on 09 Nov 2025.

[6] Deepseek. 2025. Retrieved from https://api-docs.deepseek.com/news/news1226 09 Nov 2025.

[7] Google Blog. 2025. Retrieved from https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/ on 06 Nov 2025.

[8] Arxiv.org. 2023 Retrieved from https://arxiv.org/abs/2308.02441 on 02 Oct 02 2023.

[9] Krumsvik RJ. 2025. GPT-4's capabilities for formative and summative assessments in Norwegian medicine exams-an intrinsic case study in the early phase of intervention. Front. Med. (Lausanne). 12:1441747. doi: 10.3389/fmed.2025.1441747

[10] Omopekunola MO, Kardanova EY. 2024. Automatic generation of physics items with large language models (LLMs). REID (Research and Evaluation in Education). 10(2):4. doi: 10.21831/reid.v10i2.76864

[11] Newton P, Xiromeriti M. 2023. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. Assess. Eval. High. Educ. doi: 10.1080/02602938.2023.2299059

[12] Gilson A, Safranek C, Huang T, Socrates V, Chi L, Taylor R, Chartash D. 2022. How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment. doi: 10.1101/2022.12.23.22283901

[13] Jin HK, Lee HE, Kim EY. 2024. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med. Educ. 24(1). doi: 10.1186/s12909-024-05944-8

[14] Kung TH, Cheatham M, Medenilla A, Silos C, De Leon L, Elepano C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J. 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digital Health. 2(2). doi: 10.1371/JOURNAL.PDIG.0000198

[15] Wang YM, Shen HW, Chen TJ. 2023. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. Journal of the Chinese Medical Association. 86(7): 653–658. doi: 10.1097/JCMA.0000000000000942

[16] Zong H, Li J, Wu E, Wu R, Lu J, Shen B. 2024. Performance of ChatGPT on Chinese national medical licensing examinations: a five-year examination evaluation study for physicians, pharmacists and nurses. BMC Med. Educ. 24(1). doi: 10.1186/s12909-024-05125-7

[17] Sumbal A, Sumbal R, Amir A. 2024. Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT's Performance in Academic Testing. J Med Educ Curric Dev. 11:23821205241238640. https://doi.org/10.1177/23821205241238641

[18] Jaleel A, Aziz U, Farid G, Zahid Bashir M, Mirza T, Khizar Abbas S, Aslam S, Sikander R. 2025. Evaluating the potential and accuracy of ChatGPT-3.5 and 4.0 in medical licensing and in-training examinations: systematic review and meta-analysis. JMIR Med. Educ. 11: e68070. doi: 10.2196/68070

[19] Akolekar H, Jhamnani P, Kumar V, Tailor V, Pote A, Meena A, Kumar K, Chaila J, Kumar D. 2025. The role of generative AI tools in shaping mechanical engineering education from an undergraduate perspective. Sci. Rep. 15(1):1–14. doi: 10.1038/S41598-025-93871-Z;SUBJMETA

[20] Al-Shakarchi NJ, Haq IU. 2023. ChatGPT performance in the UK medical licensing assessment: how to train the next generation? Digital Health. 1(3):309–310. doi: 10.1016/J.MCPDIG.2023.06.004

[21] Pursnani V, Sermet Y, Kurt M, Demir I. 2023. Performance of ChatGPT on the US fundamentals of engineering exam: comprehensive assessment of proficiency and potential implications for professional environmental engineering practice. Computers and Education: Artificial Intelligence. 5. doi: 10.1016/j.caeai.2023.100183

[22] Orgill BD, Nolin J. 2022. Learning taxonomies in medical simulation. StatPearls, Florida, USA. https://www.ncbi.nlm.nih.gov/books/NBK559109/.

[23] Gunawan I, Palupi AR. 2016. TAKSONOMI BLOOM – REVISI RANAH KOGNITIF: KERANGKA LANDASAN UNTUK PEMBELAJARAN, PENGAJARAN, DAN PENILAIAN," Premiere Educandum: Jurnal Pendidikan Dasar dan Pembelajaran. 2(2). doi: 10.25273/PE.V2I02.50

[24] Nakajima N, Fujimori T, Furuya M, Kanie Y, Imai H, Kita K, Uemura K, Okada S. 2024. A comparison between GPT-3.5, GPT-4, and GPT-4V: can the large language model (ChatGPT) pass the Japanese Board of Orthopaedic surgery examination? Cureus. 6(3). doi: 10.7759/CUREUS.56402

[25] Susnjak T, McIntosh TR. 2024. ChatGPT: The end of online exam integrity? Educ. Sci. (Basel). 14(6). doi: 10.3390/educsci14060656

[26] Yudovich MS, Makarova E, Hague CM, Raman JD. 2024. Performance of GPT-3.5 and GPT-4 on standardized urology knowledge assessment items in the United States: a descriptive study. J. Educ. Eval. Health Prof. 21. doi: 10.3352/JEEHP.2024.21.17

[27] Bahroun Z, Anane C, Ahmed V, Zacca A. 2023. Transforming education: a comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. Sustainability. 15(17): 12983. doi: 10.3390/SU151712983

[28] Szabó A, Laein GD. 2025. Comparative evaluation of large language models performance in medical education using urinary system histology assessment. Sci. Rep.15(1). doi: 10.1038/S41598-025-17571-4.

[29] Amoah N, Fianko S, Dake S, Agyemang K, Nyame I, Adjaye-Gyamfi O, Nooni I, Agbemaba E, Agropah F, Zuma F, Zaglago L, Atiase D, Amposah R, Lartey R. 2024. The impact of Ai chatbots on the landscape of professional accountancy examination: an experimental study. doi: 10.2139/SSRN.4991304

[30] Eulerich M, Sanatizadeh A, Vakilzadeh H, Wood D. 2023. Can artificial intelligence pass accounting certification exams? ChatGPT: CPA, CMA, CIA, and EA? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4452175

[31] Hartman H, Mutmansky J. 2002. Introductory mining engineering. John Wiley & Sons.

[32] Dwivedi C. 2019. A study of selected-response type assessment (MCQ) and essay type assessment methods for engineering students. Journal of Engineering Education Transformations. 32(3). doi: 10.16920/JEET/2019/V32I3/143024

[33] Bremner DJ, Le Kernec J, Fioranelli F, Dale VHM. 2018. The use of multiple-choice questions in 3rd-year electronic engineering assessment: a case study. 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering. p. 887–892. doi: 10.1109/ TALE.2018.8615153

[34] Visual Capitalist. 2026 Retrieve from https://www.visualcapitalist.com/the-10-most-used-ai-chatbots-in-2025/on 25 Jan 2026.

[35] One Little Web. 2026. Retrieved from https://onelittleweb.com/data-studies/best-ai-chatbots/ on 25 Jan 2026.
[36] Meo SA, Al-Masri AA, Alotaibi M, Meo MZ S, Meo MOS. ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare. 11(14). doi: 10.3390/HEALTHCARE11142046
[37] Agarwal M, Sharma P, Wani P. 2025. Evaluating the accuracy and reliability of large language models (ChatGPT, Claude, DeepSeek, Gemini, Grok, and Le Chat) in answering item-analyzed multiple-choice questions on blood physiology. Cureus. doi: 10.7759/cureus.81871
[38] Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. 2025. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. BMC Oral Health. 25(1). doi: 10.1186/s12903-025-05926-2
[39] Law AK K, So J, Lui C, Cheung K, Kei-ching Hung K, Graham C. AI versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. BMC Med. Educ. 25(1). doi:10.1186/s12909-025-06796-6
[40] Nikolic S, Daniel D, Haque R, Belkina M, Hassan G, Grundy S, Lyden S, Neal P, Sandison C. 2023. ChatGPT versus engineering education assessment: a multidisciplinary and multi-institutional benchmarking and analysis of this generative artificial intelligence tool to investigate assessment integrity. European Journal of Engineering Education. 48(4): 559–614. doi: 10.1080/03043797.2023.2213169
[41] Córdova-Esparza DM. 2025. AI-powered educational agents: opportunities, innovations, and ethical challenges. Information. 16(6). doi: 10.3390/INFO16060469
[42] Güvel MC, Kıyak YS, Varan HD, Sezenöz B, Coşkun Ö, Uluoğlu C. 2025. Generative AI vs. human expertise: a comparative analysis of case-based rational pharmacotherapy question generation. Eur. J. Clin. Pharmacol. 81(6): 875–883. doi: 10.1007/S00228-025-03838-2/METRICS
[43] Kıyak YS, Emekli E. 2024. ChatGPT prompts for generating multiple-choice questions in medical education and evidence on their validity: a literature review. Postgrad. Med. J. 2024:1–8. doi: 10.1093/POSTMJ/QGAE065
[44] Bolgova O, Ganguly P, Mavrych V. 2025. Comparative analysis of LLMs performance in medical embryology: a cross-platform study of ChatGPT, Claude, Gemini, and Copilot. Anat. Sci. Educ. 18(7):718–726. doi: 10.1002/ASE.70044
[45] Funk PF, Hoch C, Knoedler S, Knoedler L, Cotofana S, Sofo G, Bashiri Dezfouli A, Wollenberg B, Guntinas-Lichius O, Alfertshofer M. 2024. ChatGPT's response consistency: a study on repeated queries of medical examination questions. Eur. J. Investig. Health Psychol. Educ. 14(3): 657–668. doi: 10.3390/ejihpe14030043
[46] Rodriguez-Echeverria R, Gutierrez JD, Conejero JM, Prieto AE. 2024. Analysis of ChatGPT performance in computer engineering exams. Revista Iberoamericana de Tecnologias del Aprendizaje. 19:71–80. doi: 10.1109/RITA.2024.3381842
[47] Xu P, Wu Y, Jin K, Chen X, He M, Shi D. 2025. DeepSeek-R1 outperforms Gemini 2.0 Pro, OpenAI o1, and o3-mini in bilingual complex ophthalmology reasoning. Advances in Ophthalmology Practice and Research. 5(3):189–195. doi: 10.1016/J.AOPR.2025.05.001
[48] Gotta J, Le Hong Q, Koch V, Gruenewald L, Geyer T, MartinS, Scholtz J, Booz C, Santos D, Mahmoudi S, Eichler K, Gruber-Rouh T, Hammerstingl R, Biciusca T, Juegrens L, Hohne R, Mader C, Vogl T, Reschke P. 2024. Large language models (LLMs) in radiology exams for medical students: Performance and consequences. RoFo Fortschritte auf dem Gebiet der Rontgenstrahlen und der Bildgebenden Verfahren. 197(9): 1057–1067. doi: 10.1055/a-2437-2067
[49] Takagi S, Watari T, Erabi A, Sakaguchi K. 2023. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. JMIR Med Educ. 9(1):e48002. doi: 10.2196/48002
[50] Levin G, Horesh N, Brezinov Y, Meyer R. 2024. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. BJOG. 131(3): 378–380, doi: 10.1111/1471-0528.17641
[51] Meo SA, Abukhalaf FA, Eltoukhy RA, Sattar K. 2025. Exploring the role of DeepSeek-R1, ChatGPT-4, and Google Gemini in medical education: how valid and reliable are they? Pak. J. Med. Sci. vol. 41(7):1887–1892. doi: 10.12669/pjms.41.7.12183
[52] Li W, Li L, Xiang T, Liu X, Deng W, Garcia N. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024. pp. 2819–2834.https://arxiv.org/pdf/2403.17752

[53] Epstein RH, Dexter F. 2023. Variability in large language models' responses to medical licensing and certification examinations. Comment on 'How Does ChatGPT perform on the United States Medical Licensing Examination. The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 9(1):e48305. doi: 10.2196/48305

[54] Su M-C, Lin L-E, Lin L-H, Chen Y-C. 2024. Assessing question characteristic influences on ChatGPT's performance and response-explanation consistency: insights from Taiwan's nursing licensing exam. Int. J. Nurs. Stud. p. 104717. doi: 10.1016/J.IJNURSTU.2024.104717

[55] Ogunleye B, Zakariyyah KI, Ajao O, Olayinka O, Sharma H. 2024. Higher education assessment practice in the era of generative AI tools. Journal of Applied Learning and Teaching. vol. 7(1): 46–56. doi: 10.37074/jalt.2024.7.1.28

[56] Shuraiqi SA, Abdulsalam AA, Masters K, Zidoum H, AlZaabi A. 2024. Automatic generation of medical case-based multiple-choice questions (MCQs): a review of methodologies, applications, evaluation, and future directions. Big Data and Cognitive Computing. 8(10):139. doi: 10.3390/BDCC8100139

[57] Grévisse C, Pavlou MAS, Schneider JG. 2024.Docimological quality analysis of LLM-generated multiple choice questions in computer science and medicine. SN Comput. Sci. 5(5):1–14. doi: 10.1007/S42979-024-02963-6/FIGURES/12

[58] Sabaner MC, Hashas ASK, Mutibayraktaroglu KM, Yozgat Z, Klefter ON, Subhi Y. 2024. The performance of artificial intelligence-based large language models on ophthalmology-related questions in Swedish proficiency test for medicine: ChatGPT-4 omni vs Gemini 1.5 Pro. AJO International. 1(4). doi: 10.1016/j.ajoint.2024.100070

[59] Cheung BHH, Lau G, Wong G, Lee E, Kulkarni D, Seow C, Wong R, Co M. 2023. ChatGPT versus human in generating medical graduate exam multiple choice questions—a multinational prospective study (Hong Kong S. A.R., Singapore, Ireland, and the United Kingdom). PLoS One, 18(8). doi: 10.1371/JOURNAL.PONE.0290691

[60] Flores-Cohaila JA, Garcia-Vicente A, Vizcarra-Jimenez S, De la Cruz-Galan J, Gutierrez-Arratia J, Quiroga Torres B, Taype-Rondan A. 2023. Performance of ChatGPT on the Peruvian national licensing medical examination: cross-sectional study. JMIR Med. Educ. 9:e48039. doi: 10.2196/48039

[61] Prazeres F. 2025. Can ChatGPT help general practitioners become acquainted with conversations about dying? A simulated single-case study. Healthcare (Basel). 13(7). doi: 10.3390/healthcare13070835

[62] Özer NE, Balcı Y, Bölükbaşı G, İlhan B, Güneri P. 2025. Examining the role of artificial intelligence in assessment: a comparative study of ChatGPT and educator-generated multiple-choice questions in a dental exam. Eur. J. Dent. Educ. doi: 10.1111/eje.70034

[63] Moore S. Schmucker R. Mitchell T. Stamper J. 2024. Automated generation and tagging of knowledge components from multiple-choice questions. 11th ACM Conference on Learning @ Scale. p. 122–133 doi: 10.1145/3657604.3662030

[64] Moore S, Nguyen HA, Chen T, Stamper J. 2023. Assessing the Quality of multiple-choice questions using GPT-4 and rule-based methods. Lecture Notes in Computer Science. 14200: 229–245. doi: 10.1007/978-3-031-42682-7_16

[65] Gill GS, Tsai J, Moxam J, Sanghvi HA, Gupta S. 2024. Comparison of Gemini Advanced and ChatGPT 4.0's performances on the ophthalmology resident ophthalmic knowledge assessment program (OKAP) examination review question banks. Cureus. doi: 10.7759/cureus.69612

[66] Mistry NP, Saeed H, Rafique S, Le T, Obaid H, Adams SJ. 2024. Large language models as tools to generate radiology board-style multiple-choice questions. Acad. Radiol. 31(9):3872–3878. doi: 10.1016/J.ACRA.2024.06.046

[67] Sridharan K, Sequeira RP. 2024. Artificial intelligence and medical education: application in classroom instruction and student assessment using a pharmacology & therapeutics case study. BMC Med. Educ. 24(1). doi: 10.1186/S12909-024-05365-7 5-7