Assessment Design in the AI Era: Applying Psychometric Theory to Identify Items on Which Humans and Chatbots Diverge

Licol Zeinfeld[1,*,†], Alona Strugatski[1,†], Ziva Bar-Dov[1], Ron Blonder[1], Shelley Rap[1] and Giora Alexandron[1]

[1]*Weizmann Institute of Science, Rehovot, Israel*

Abstract

The rapid adoption of large language models (LLMs) in education raises profound challenges for assessment validity. Current evaluations often rely on descriptive statistics comparing human and chatbot performance, but little research has examined more theoretically-established, generalizable ways of measuring LLM capabilities in education. Drawing on psychometric theory, this paper introduces a statistically principled approach to identifying items where humans and LLMs diverge. We apply Differential Item Functioning (DIF) analysis—traditionally used to detect item-level differences across demographic groups—and evaluate this approach on responses from human learners and three leading chatbots (ChatGPT-4o, Gemini 1.5 Pro, Claude 3.5 Sonnet) to two instruments: a high school chemistry diagnostic test and a university entrance exam. Subject-matter experts then analyzed DIF-flagged items to characterize task dimensions associated with chatbot over- or under-performance. Results show that DIF-informed analytics provide a robust framework for understanding LLM capabilities and highlight their value for guiding the design of valid, reliable, and fair assessments in the AI era.

Keywords

Assessment, Generative AI, Item Response Theory, Differential Item Functioning

1. Introduction

The rapid adoption of generative AI (GenAI) tools in education has created both opportunities and risks. While these systems, particularly Large Language Models[1] (LLMs) such as ChatGPT, can provide personalized explanations, feedback, and support for learners, their growing use also poses a profound threat to the validity of educational assessments [1, 3]. By making it exceptionally easy for students to generate fluent and convincing answers, GenAI lowers the barriers to academic dishonesty in ways that traditional safeguards struggle to address [1, 3, 7]. This risk is amplified in unproctored or remote environments, where monitoring is limited and misuse is harder to detect [20]. Importantly, the concern is no longer theoretical: recent studies document that students are already using GenAI for assignments and exams, directly undermining the reliability of assessment results and raising urgent challenges for educational practice and policy [13, 16]. These changes point to a key observation: assessments must not only be resistant to GenAI misconduct but also capture authentic signals of student understanding. To achieve this, we need to understand the capabilities and limitations of LLMs with respect to various types of assessment tasks, and specifically, where they differ from human learners.

Current evaluations of LLM performance on assessment tasks are largely shaped by technical benchmarks [2, 9, 18]. These benchmarks provide descriptive statistics that locate LLMs on various scales relative to human learners and provide evidence that LLMs are affected differently from humans by task features such as visual elements or sequential reasoning steps [17, 19]. Yet, little research has examined more generalizable ways of measuring LLM capabilities in relation to human learners. Assessment, by definition, is a proxy that seeks to generalize from sparse samples of student performance (e.g., across

LLM Psychometrics: Applying Assessment Theories to Evaluate LLMs and LLM-Supported Learners, LAK 2026 Workshop

*Corresponding author.

†These authors contributed equally.

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]In this paper, we use the term LLMs when referring to specific types of models and architectures, GenAI when referring to the broader class of generative systems, and Chatbots when referring to agents powered by LLMs.

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-15

CEUR-WS.org/Vol-4330/paper6.pdf

time or content). For human learners, the design and interpretation of such proxies have been refined through decades of assessment research and implementation. For GenAI agents, however, it remains far less clear what inferences assessments support.

Mature frameworks from educational measurement—such as, but not limited to, Item Response Theory (IRT) [4]—may offer principled ways to assess the capabilities of GenAI. Recent studies [14, 15] have shown that psychometric modeling can be extended beyond its traditional applications to capture systematic differences between human learners and chatbots, mainly for identifying GenAI cheating. Our work moves the center of attention from examinees to items: our goal is to identify items on which chatbots and humans exhibit differential behavior. Identifying such items serves two important purposes. First, to better understand the dimensions that make tasks easy or difficult for LLMs relative to human learners. Second, to apply these insights to the design of assessments that are better adapted—in terms of validity, reliability, and fairness—to scenarios in which students may work (legitimately or not) with GenAI.

Within psychometrics, there is a collection of methods known as Differential Item Functioning (DIF). DIF refers to situations in which an assessment item functions differently for subgroups of learners distinguished by a characteristic unrelated to the construct being measured (typically a demographic characteristic). In DIF terminology, the baseline group is referred to as the reference group, while the group of interest is referred to as the focal group. DIF methods go beyond simply comparing overall performance between the reference and focal groups by controlling for student ability, thereby pointing out item bias rather than general group-level performance differences.

DIF analysis is widely used in assessment design to flag items that may be biased or poorly constructed, as such items compromise the validity and fairness of the assessment [12]. For example, a mathematics item requiring advanced reading comprehension may disadvantage learners with equivalent math ability but weaker language skills (e.g., second language learners). Methods to detect DIF include both non-parametric approaches, such as the Mantel–Haenszel procedure [11], or parametric ones, such as those based on IRT or logistic regression [10].

The rationale of the current work is to examine whether DIF methods can be useful to identify items on which human learners and chatbots differ. The observation is that chatbots can be referred to as the "focal" group, while humans are the "reference" (baseline) group. This is formulated through the following research questions (RQs):

RQ1: Can DIF techniques, specifically LR-DIF, identify items that function differently for human learners and chatbots?

RQ2: What are the key characteristics that subject-matter experts identify in items on which chatbots exhibit differential performance compared to human learners?

To study these questions, we applied LR-DIF analysis using a per-item rest score as an ability-matching variable and trimming non-overlapping ability ranges between humans and chatbots (see Section 4.1), to identify items on which chatbots and human learners show differential behavior. We then iteratively refined this LR-DIF analysis pipeline and evaluated it on human learners and GenAI responses to two assessment instruments taken from two very different contexts: a high school chemistry test administered as a formative assessment, and the quantitative section of a high-stakes psychometric entrance exam for higher education. The GenAI responses were generated using three chatbots: ChatGPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet. Following that, subject-matter experts conducted a qualitative analysis of the chemistry DIF items to identify key dimensions that may lead to the differential behavior. The key contribution of this paper is demonstrating that a theory-inspired and statistically robust application of LR-DIF—using per-item rest-score matching and trimming non-overlapping ability ranges—can identify items that exhibit differential behavior for chatbots relative to students, and that the resulting DIF outputs can provide actionable item analytics to those who wish to incorporate GenAI-related considerations into assessment design.

Insert Figure 1 here: Implemented methodological framework for human–LLM DIF analysis.

Figure 1: Implemented methodological framework for human–LLM DIF analysis.

2. Differential Item Functioning (DIF)

DIF methods test whether an item functions differently across groups of respondents once respondents' abilities have been controlled for. Here, "ability" refers to a respondent's overall proficiency on the instrument and is measured by the per-item rest score, i.e., the sum of the respondent's correct responses across all items excluding the target item. In DIF terminology, the reference group is the baseline for comparison (humans, in our case), and the focal group is the group tested for differences (chatbots, in our case). For analysis, we denote each response as $X_{ij} \in \{0, 1\}$ for respondent i on item j, where 1 indicates a correct answer and 0 an incorrect one. Conceptually, an item exhibits DIF if two respondents with the same ability, but from different groups, have systematically different probabilities of answering it correctly. Throughout, we adopt a consistent direction convention: POS (positive) indicates a chatbot advantage (higher conditional probability of answering correctly than humans at the same ability level), and NEG (negative) indicates a chatbot disadvantage (lower conditional probability given ability).

2.1. Logistic Regression DIF (LR-DIF)

LR-DIF provides a flexible framework that models the probability of a correct response as a function of ability, group membership, and their interaction:

$$\text{logit}\, P(X_{ij} = 1) = \beta_0 + \beta_1 A_{i,-j} + \beta_2 G_i + \beta_3 (A_{i,-j} \times G_i), \tag{1}$$

where $G_i = 0$ for humans and $G_i = 1$ for chatbots, and $A_{i,-j}$ is the per-item rest score used as an ability proxy. Nested model comparisons allow us to test separately for uniform DIF (via the main group effect β_2) and non-uniform DIF (via the interaction term β_3). In cases where the interaction implies different directions across the ability range, we label the effect as Sign-change DIF. From the fitted GLMs we compute likelihood-ratio p-values, McFadden's ΔR^2, and at-median odds ratios, which indicate how much more or less likely chatbots are to answer correctly relative to humans at the median ability level after trimming. We adopt $\Delta R^2 \geq 0.035$ as the threshold for a meaningful effect size [6], and use these measures together to detect and characterize DIF effects.

3. Methodology

This section describes how LR-DIF was applied in order to identify items that function differently for humans and chatbots (RQ1), as well as the process through which DIF items were subjected to qualitative analysis by subject matter experts to characterize human–chatbot DIF behavior (RQ2).

3.1. LR-DIF Analysis Design

3.1.1. Task Definition and Modeling

Our task is to apply LR-DIF to detect and characterize multiple-choice items exhibiting systematic response differences between human learners and LLM-based chatbots. Adopting the DIF terminology, we refer to humans as the reference group and chatbots as the focal group.

3.1.2. Design Process

To develop an analysis pipeline for identifying true item-level differences between humans and chatbots, we followed the process outlined in Figure 1. Below, we briefly describe the key steps. In the Experiments and Results section, we describe how these steps were implemented, and their results.

Table 1
Fraction correct per item for students and chatbots.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19
Students	.836	.531	.603	.699	.921	.918	.968	.968	.924	.944	.740	.563	.820	.827	.783	.429	.484	.487	.374
Chatbots	.983	.683	.850	.717	.917	.267	.283	.983	1.000	.967	.517	1.000	.983	.233	.133	.967	.983	.733	.983

Insert Figure 2 here: Human–Chatbot ability distribution.

Figure 2: Human–Chatbot ability distribution.

Step 1 – DIF Analysis: Given response data from humans and chatbots, we preprocessed the data and applied LR-DIF on the preprocessed data. This application of LR-DIF identifies items for which humans and chatbots with the same ability level (i.e., respondents' overall proficiency on the instrument) differ in the probability of providing a correct response. We interpret the items flagged in this step as DIF items.

Step 2 – Qualitative Expert Analysis: Subject-matter experts analyzed DIF-flagged chemistry items to characterize task dimensions associated with chatbot over- or under-performance.

3.2. Instruments and Data

We used two assessment instruments from two very different contexts. The first was a chemistry diagnostic test administered as a formative assessment activity in preparation for the high school matriculation exam. It consisted of 22 multiple-choice items answered by 931 students. The second was the quantitative section of a psychometric entrance exam used for higher-education admissions, containing 40 items with over 4,800 respondents. Both instruments were multi-modal, with some questions including figures, images, or formulas.

To generate chatbot data, we collected responses from three LLMs: ChatGPT-4o, Gemini 1.5 Pro, and Claude 3.5 Sonnet. Each instrument was uploaded with a prompt requesting only the final answers. The models typically returned numbered answer lists, which we exported to CSV. This process was repeated 20 times per model, simulating 20 "artificial students" for each chatbot. Runs were conducted in separate sessions to prevent memory effects, and variation across responses was introduced by using a non-zero temperature setting. Since both assessments included multimodal content, it was important that all models supported visual inputs. When a model skipped an item or did not provide a valid option, the response was counted as incorrect (similar to how student responses were treated).

The ability distributions of the chatbots and human learners on both instruments are presented in Figure 2. Table 1 shows the fraction correct of each group on the chemistry items. As can be seen, although the overall success rate of the students is higher, the chatbots have higher success rates on most items. These relations can be handled by the LR-DIF, as we later demonstrate.

3.3. Qualitative Analysis by Domain Experts

The qualitative analysis of the DIF items was applied to the chemistry instrument. It was conducted by two subject-matter experts who lead a team responsible for analyzing the results of the high-stakes national matriculation exams in chemistry each year. The experts received a table listing the DIF properties of the items flagged by the LR-DIF method and the instrument. The two experts independently reviewed the DIF items, integrated their observations through discussion, and provided qualitative judgments regarding their characteristics.

4. Experiments and Results

This section applies the pipeline outlined in Section 3 and illustrated in Figure 1. The computational part was applied to the psychometric and chemistry data. The qualitative part was applied to the chemistry

Table 2
LR-DIF: Summary of DIF-flagged items. Q: uniform DIF; $\widetilde{Q}$: non-uniform DIF; items with no accent: sign-change DIF.

Instrument	POS	NEG
Psychometric	Q11, $\widetilde{Q}$13, Q34, $\widetilde{Q}$35, Q40	$\widetilde{Q}$4, $\widetilde{Q}$5, $\widetilde{Q}$6, $\widetilde{Q}$7, $\widetilde{Q}$10, $\widetilde{Q}$16, $\widetilde{Q}$28, $\widetilde{Q}$31, $\widetilde{Q}$32, $\widetilde{Q}$33
Chemistry	$\widetilde{Q}$3, $\widetilde{Q}$16, Q17, $\widetilde{Q}$18, Q19	Q5, $\widetilde{Q}$6, $\widetilde{Q}$7, $\widetilde{Q}$9, $\widetilde{Q}$11, Q14, Q15, $\widetilde{Q}$21, Q22

dataset, and is described in Subsection 4.3. We then conclude with the resulting pipeline for detecting human:GenAI DIF items.

4.1. Preprocessing

The following properties were computed per item j. Let $X_{ij} \in \{0, 1\}$ denote correctness for respondent i on item j.

(i) Ability proxy. We defined an ability proxy for respondent i as the rest score $R_{i,-j}$, the respondent's sum of correct answers across all items except item j. Using rest score for respondent i, as opposed to using the global score of respondent i, avoids contamination from the target item and provides a consistent matching variable.

(ii) Trimming non-overlapping ability ranges. Following DIF analysis guidelines to ensure that both groups were compared only within the overlapping strata of the ability proxy, we removed the human responses in the ability ranges not populated by the chatbots.

4.2. LR-DIF Analysis and Results

We applied logistic regression DIF (LR-DIF) on the preprocessed data, with humans as the reference group and chatbots as the focal group. For each item, we fit three nested logistic regression models, computed likelihood-ratio test p-values for uniform and non-uniform DIF, calculated McFadden's ΔR^2, flagged items as DIF if ($p_{uniform} < .05$ or $p_{nonuniform} < .05$) and $\Delta R^2 > 0.035$, summarized flagged items by log-odds contrast, odds ratio, and probability gap between humans and chatbots at median ability, and classified direction as POS, NEG, or sign-change if the group effect changed sign at different ability levels.

The items flagged as DIF by LR-DIF are presented in Table 2. The experimental results support a theory-informed application of LR-DIF in which examinees are matched using a per-item rest score and comparisons are restricted to overlapping ability ranges between humans and chatbots. The resulting DIF items provide a robust foundation for qualitative expert analysis of task dimensions associated with chatbot over- and under-performance.

4.3. RQ2: Domain Expert Analysis and Results

As described in Subsection 3.3, the qualitative analysis of the DIF items was conducted by two subject-matter experts who received the instrument and a table that listed the properties of the LR-DIF items, taken from Table 2.

As shown in Table 2, Items 3, 16, 17, and 19 had POS DIF, meaning that chatbots had higher likelihood of answering them correctly than humans at the same ability level. On item 3, the chatbots successfully circumvented a common alternative conception according to which heating causes molecules to decompose. Instead, it distinguished accurately between a physical phase transition, entailing the disruption of intermolecular forces, and a chemical transformation, which requires the cleavage of covalent bonds within molecules. On item 16, which involves a sequence of solution manipulations, the chatbots effectively maintained consistency by monitoring the relationship between solute moles,

solution volume, and concentration across several experimental stages. On item 17 the chatbots demonstrated a scientifically appropriate interpretation of the law of conservation of matter by situating the assessment of mass change at the level of the entire experimental system, rather than restricting the analysis just to the original metal. Item 19 further highlighted the chatbots' capacity to apply algorithmic reasoning: the multiple-choice structure facilitated accurate assignment of oxidation states, identification of electron transfer, and correct designation of oxidizing and reducing agents. Collectively, these examples underscore the chatbots' relative strength in tasks that prioritize rule-governed reasoning and conceptual clarity, where adherence to canonical chemical principles prevents susceptibility to intuitive but scientifically inaccurate interpretations.

In contrast, items 6, 7, 14, and 15 had NEG DIF, meaning that chatbots had lower likelihood of answering them correctly than humans at the same ability level. These items place greater demands on visual interpretation, understanding linguistic nuances, and the execution of complex, multi-step procedures. Items 6 and 7 require learners to interpret partial microscopic models, such as diagrams omitting solvent molecules, or to map symbolic labels (A–C) onto microscopic and macroscopic representations of aqueous solutions. These forms of representation, which depend heavily on implicit visual conventions, posed substantial challenges for the chatbots, whereas students were able to draw upon prior representational knowledge. Item 14 illustrates a further point of divergence: the solution involves a multi-stage stoichiometric calculation, proceeding from gas volume to moles, through mole ratios, and finally to mass determination, with each step contingent on the correctness of the one before it. Such extended procedures are particularly vulnerable to intermediate errors in algorithmic reasoning. Item 15 compounds this difficulty, as errors originating in the previous calculation can propagate forward, whereas students often display the metacognitive awareness required to detect and correct inconsistencies in their own work.

Taken together, these findings suggest that the chatbots excelled in tasks governed by formalized conceptual rules. By contrast, students retain a relative advantage in tasks that demand flexible interpretation of visual representations, sensitivity to linguistic subtleties, and careful monitoring of extended quantitative reasoning processes.

5. Discussion

The LR-based DIF approach developed and piloted in this research provides a method and conceptual framework for identifying assessment items that show differential behavior between humans and chatbots. In two assessment contexts and on key leading chatbots, we demonstrated that the method yields meaningful item-level signals. A key observation that DIF, and this research, makes, is that differential behavior is a more nuanced property than merely looking at overall performance gaps between groups, since the DIF analysis controls for ability level, and can identify items as POS DIF (chatbots over-perform learners of the same skill) although the difference between the mean performance of both groups is in the opposite direction. The LR-DIF that our method uses offers a modeling that also enables detection of non-uniform DIF patterns in which either the magnitude or direction of group differences vary across ability levels, including sign-change patterns.

Using the DIF-method to deliver content analytics to the subject-matter experts (RQ2) drove an analysis that yielded interesting insights. As reported in Subsection 4.3, analyzing the POS DIF items, we found that the chatbots managed to circumvent distractors aimed at surfacing alternative conceptions (sometimes referred to as "misconceptions") commonly held by students. Given that LLMs can be cognitively biased due to their training data [5], which likely included student data also representing wrong or incomplete knowledge [20], the fact that the chatbots dodged these designated "traps" is interesting, and is in disagreement with previous work that found moderate alignment between LLMs and student misconceptions [8]. On the contrary, experts' analysis of the NEG DIF items revealed that the chatbots underperformed on items requiring visual interpretation and connecting visual and textual information, items that their wording required understanding linguistic nuances, and those requiring performing multi-step problem solving procedures. These findings reinforce previous work

about GenAI problem solving in STEM domains [17, 19].

More generally, this research demonstrates a fruitful application of assessment and measurement theory to learning analytics, with the purpose of establishing robust approaches for understanding GenAI capabilities, and for developing assessment in the GenAI era.

Limitations. In terms of internal validity, because the chatbot responses did not span the full range of ability levels, we trimmed non-overlapping ability ranges (Section 4.1) and restricted the analysis to the overlapping rest-score range between humans and chatbots. This reduced the effective sample size and may have affected the robustness of the LR-DIF results. The main limitation to external validity is that the results are based on a small number of instruments and three specific GenAI tools. Overall, more research is required to increase the credibility and generalizability of this approach. Additionally, our DIF findings are based on a single detection approach (LR-DIF). While LR-DIF is widely used, other DIF frameworks (e.g., Mantel–Haenszel or IRT-based DIF) may flag partially different sets of items or characterize effects differently, so the scope of conclusions is necessarily limited to the LR-DIF lens.

Contribution and future work. The main contribution of this work is the introduction of a theory-driven and statistically sound method for detecting items that exhibit differential behavior between chatbots and human learners, and the presentation of evidence of its ability to yield meaningful analytics for subject-matter experts who seek to integrate GenAI considerations into assessment design. In future work, we will (1) study the task dimensions of Human–GenAI DIF items to better characterize GenAI capabilities, (2) incorporate Human–GenAI DIF analysis into assessment design, and (3) extend our analyses beyond LR-DIF by applying additional DIF detection approaches (e.g., Mantel–Haenszel) to assess the robustness and consistency of the identified DIF items.

Acknowledgments

This work was supported by the Knell Family Institute for Artificial Intelligence, Israel. The authors thank the National Institute for Testing and Evaluation for providing access to psychometric exam data.

Ethics statement

The research was approved by the Institutional Review Board.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to perform grammar, spelling checks and citation management. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Noor Akbari. 2024. The AI cheating crisis: Education needs its anti-doping movement. Retrieved from https://www.edweek.org/technology/opinion-the-ai-cheating-crisis-education-needs-its-anti-doping-movement/2024/02.

[2] Beatriz Borges et al. 2024. Could ChatGPT get an engineering degree? Evaluating higher education vulnerability to AI assistants. *Proceedings of the National Academy of Sciences* 121, 49 (2024).

[3] Debby R. E. Cotton, Peter A. Cotton, and J. Reuben Shipway. 2024. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International* 61, 2 (2024), 228–239. doi:10.1080/14703297.2023.2190148

[4] Rafael Jaime De Ayala. 2013. *The theory and practice of item response theory.* Guilford Publications.

[5] Jessica Maria Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive Bias in Decision-Making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

[6] Michael G. Jodoin and Mark J. Gierl. 2001. Evaluating Type I Error and Power Rates Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection. *Applied Measurement in Education* 14, 4 (2001), 329–349. doi:10.1207/S15324818AME1404_2

[7] Mohammad Khalil and Erkan Er. 2023. Will ChatGPT Get You Caught? Rethinking of Plagiarism Detection. In *International Conference on Human-Computer Interaction*. Springer, 475–487.

[8] Naiming Liu, Shashank Sonkar, and Richard Baraniuk. 2025. Do LLMs Make Mistakes Like Students? Exploring Natural Alignments Between Language Models and Human Error Patterns. In *International Conference on Artificial Intelligence in Education*. Springer, 364–377.

[9] Pan Lu et al. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *Advances in Neural Information Processing Systems* 35 (2022), 2507–2521.

[10] David Magis, Sebastien Beland, Francis Tuerlinckx, and Paul De Boeck. 2010. A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods* 42, 3 (2010), 847–862.

[11] Nathan Mantel and William Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22, 4 (1959), 719–748.

[12] Patrícia Martinková, Adéla Drabinová, Yuan-Ling Liaw, Elizabeth A. Sanders, Jenny L. McFarland, and Rebecca M. Price. 2017. Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments. *CBE Life Sciences Education* 16, 2 (2017).

[13] Arianna Prothero. 2024. New Data Reveal How Many Students Are Using AI to Cheat. https://www.edweek.org/technology/new-data-reveal-how-many-students-are-using-ai-to-cheat/2024/04

[14] Benjamin Sorenson and Kenneth Hanson. 2024. Identifying Generative Artificial Intelligence Chatbot Use on Multiple-Choice, General Chemistry Exams Using Rasch Analysis. *Journal of Chemical Education* 101, 8 (2024), 3216–3223.

[15] Alona Strugatski and Giora Alexandron. 2025. Applying IRT to Distinguish Between Human and Generative AI Responses to Multiple-Choice Assessments. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference, LAK 2025*. ACM, 817–823.

[16] Teo Susnjak and Timothy R. McIntosh. 2024. ChatGPT: The End of Online Exam Integrity? *Education Sciences* 14, 6 (2024).

[17] Karen D. Wang, Eric Burkholder, Carl Wieman, Shima Salehi, and Nick Haber. 2024. Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. In *Frontiers in Education*, Vol. 8. Frontiers Media SA, 1330486.

[18] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. arXiv preprint arXiv:2307.10635 (2023).

[19] Elad Yacobson, Yael Schleifer, Ziva Bar-Dov, Shelley Rap, Ron Blonder, and Giora Alexandron. 2026. Benchmarking AI on Standard Chemistry Exams: LLMs Still Underperform Compared to High School Students. *Journal of Science Education and Technology* (2026), 1–18.

[20] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* 55, 1 (2024), 90–112.