Physics Education

PURPOSE-LED PUBLISHING™

PAPER • **OPEN ACCESS**

Can artificial intelligence solve school-level physics exam questions? A decade-based evaluation using Turkish university entrance exams

To cite this article: Güler Çörekci and Ahmet Murat Çörekci 2026 *Phys. Educ.* **61** 055011

View the article online for updates and enhancements.

You may also like

- An Automatic Knowledge Graph Construction Method from Astronomical Literatures via Large Language Models
Jia-Qi Wang, Jin-Qu Zhang and Zhi-Qiang Zou
- Lightweight language models are prone to reasoning errors for complex computational phenotyping tasks
Sarah Pungitore, Shashank Yadav, David Maughan et al.
- Unveiling the power of multimodal large language models for radio astronomical image understanding and question answering
Fuyong Zhao, Yuyang Li, Zhenyu Liu et al.

This content was downloaded from IP address 161.9.143.196 on 20/09/2026 at 22:16

OPEN ACCESS

PAPER

Phys. Educ. **61** (2026) 055011 (10pp) iopscience.org/ped

Can artificial intelligence solve school-level physics exam questions? A decade-based evaluation using Turkish university entrance exams

Güler Çörekci[1] **and Ahmet Murat Çörekci**[2,*]

[1] Haluk Ündeğer Anatolian High School, Zeytinburnu, İstanbul, Türkiye
[2] Institute of Graduate Studies in Health Sciences, İstanbul University, İstanbul, Türkiye

E-mail: ahmetmurat.corekci@ogr.iu.edu.tr

Abstract

Background. Large language models LLMs are increasingly used by students as informal learning tools; however, their reliability in solving school-level physics questions remains uncertain. *Aim.* This study evaluated the performance of three LLMs—ChatGPT-5, Gemini 2.5 Flash, and DeepSeek V3.1—in solving physics questions from Turkish university entrance examinations administered between 2016 and 2025, comprising the Yükseköğretime Geçiş Sınavı (YGS; Transition to Higher Education Examination) in 2016–2017 and the Temel Yeterlilik Testi (TYT; Basic Proficiency Test) in 2018–2025. *Methods.* A total of 84 publicly available physics questions—28 from the 2016–2017 YGS examinations and 56 from the 2018–2025 TYT examinations—were presented to each model under identical testing conditions using the same neutral student-style prompt, without revealing the examination source. Responses were scored for final-answer correctness using official answer keys. Correct-answer accuracy

* Author to whom any correspondence should be addressed.

Original content from this work may be used under the terms of the CreativeCommonsAttribution4.0licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

 © 2026 The Author(s). Published by IOP Publishing Ltd

and net scores were calculated, and model performance was compared using Cochran's Q test. *Results.* DeepSeek V3.1 achieved the highest numerical correct-answer accuracy at 78.6% (95%CI 68.7%–86.0%), followed by Gemini 2.5 Flash at 69.0% (95% CI 58.5%–77.9%) and ChatGPT-5 at 65.5% (95% CI 54.8%–74.8%). However, the difference in correct-answer performance among the three models was not statistically significant (Cochran's $Q = 4.732$, df $= 2$, $p = 0.094$). *Conclusion.* LLMs may support physics learning and exam preparation, but their responses should be used critically and under teacher guidance rather than as substitutes for structured physics instruction.

Keywords: artificial intelligence, large language models, physics education, ChatGPT, Gemini, DeepSeek, university entrance examination

Supplementary material for this article is available online

1. Introduction

Artificial intelligence (AI) has rapidly become part of everyday teaching and learning. Large language models (LLMs) such as ChatGPT, Gemini, and DeepSeek can generate explanations, answer questions, provide step-by-step problem solutions, and offer feedback in ways that resemble human tutoring [1, 2]. Early reviews of ChatGPT in education indicate that these tools have been explored for homework support, concept clarification, revision, lesson preparation, and formative feedback, while also raising concerns about accuracy, over-reliance, and academic integrity [1, 2]. Because these tools are freely accessible and conversational, they are increasingly becoming part of students' informal learning environments, often without explicit teacher guidance [2].

Within science education, and particularly physics education, the implications of LLMs are especially worth examining. Physics learning requires more than factual recall: it depends on conceptual understanding, the ability to translate between verbal, mathematical, graphical, and diagrammatic representations, and multi-step quantitative reasoning [3–5]. Solving even a typical school-level physics question often demands selecting an appropriate model of the situation, identifying relevant variables, applying physical laws, and carrying out algebraic or numerical work without losing track of meaning. These cognitive demands make physics both a challenging and a highly informative domain in which to evaluate AI tools, because superficial pattern matching is unlikely to be sufficient for consistent success [6].

The use of LLMs in physics teaching presents a mix of opportunities and risks that physics teachers should consider explicitly. On the opportunity side, LLMs may support individualized explanations, immediate feedback, accessibility for learners who struggle to ask questions in class, and self-regulated study outside school hours [1, 2, 7]. They can also generate alternative worked examples or rephrase difficult concepts, which is a routine but time-consuming part of teachers' work [8]. On the risk side, LLMs are known to produce hallucinations and confidently stated but incorrect reasoning, especially when problems require precise quantitative steps or careful attention to physical assumptions [9, 10]. Recent work in physics specifically has shown that even strong models can fail on conceptually simple but contextually nuanced problems, sometimes producing solutions that are fluent yet physically wrong [11, 12]. For school students, who may not yet have the disciplinary expertise to detect such errors, this combination of fluency and unreliability can reinforce misconceptions rather than correct them.

One practical way to study LLM performance in education is through nationally administered or curriculum-based assessment questions. Such questions provide a transparent benchmark for evaluating problem-solving accuracy because they are aligned with expected learning outcomes and are often accompanied by official scoring criteria or answer keys. Studies in medicine, mathematics, introductory physics, and computing education have used national, institutional, or curriculum-based assessments to examine what LLMs can and cannot do [6, 13–15]. In physics

education, related assessment-based studies have also shown that AI-generated responses may perform well in some contexts while raising concerns about reasoning quality and assessment validity [12, 16]. More recently, Tschisgale *et al* evaluated LLM-generated solutions across six German Physics Olympiad problems, generating 20 independent solutions per problem and prompting condition, and found that both GPT-4o and the reasoning-optimized o1-preview model outperformed Olympiad participants on average, with o1-preview showing the strongest overall performance [17]. At the same time, researchers have warned that exam-based scoring captures only the correctness of final answers; it does not, on its own, reveal whether the underlying reasoning is sound or whether the explanation would be pedagogically useful for a learner [7, 12]. Both dimensions ultimately matter for physics education.

Despite a growing literature, several gaps remain. Many published evaluations focus on a single model, typically a version of ChatGPT, and on a single test, course context, or narrow assessment format [6, 16]. To our knowledge, comparative evaluations of multiple current LLMs on school-level physics using a coherent set of official, nationally administered examination questions spanning several years remain limited. There is also limited evidence about how stable LLM performance is across different question cohorts: does a model that performs well on one year's examination also perform well on others, or does accuracy fluctuate with question style, topic mix, or representational demands? For physics teachers advising students about AI use in examination preparation, this consistency question is at least as important as overall accuracy.

The present study addresses this gap by evaluating three contemporary LLMs—ChatGPT-5, Gemini 2.5 Flash, and DeepSeek V3.1—on 84 publicly available physics questions from Turkish national university entrance examinations administered between 2016 and 2025. The dataset comprised YGS questions from 2016–2017 and TYT questions from 2018–2025. The questions were obtained from official ÖSYM examination booklets, and the corresponding official answer keys were used as the reference standard [18]. All models were tested under identical, neutral student-style prompting conditions, without revealing the source of the questions, and responses were judged on final-answer correctness. The aim was to compare the accuracy and year-to-year consistency of these models on a decade-long set of curriculum-based physics items. The findings are intended to help physics teachers and education researchers form a more grounded view of where current AI tools may support exam preparation and physics learning, and where their use still requires caution and explicit teacher guidance.

2. Methodology

This study evaluated 84 physics questions from Turkish national university entrance examinations administered between 2016 and 2025. The 2016 and 2017 questions were obtained from the *Yükseköğretime Geçiş Sınavı* (YGS; Transition to Higher Education Examination), whereas the 2018–2025 questions were obtained from the *Temel Yeterlilik Testi* (TYT; Basic Proficiency Test), the first session of the *Yükseköğretim Kurumları Sınavı* (YKS; Higher Education Institutions Examination). The dataset included 28 YGS questions and 56 TYT questions. The official examination booklets and corresponding answer keys were obtained from the ÖSYM archives [18, 19].

Both YGS and TYT were centrally prepared and nationally administered by the Turkish Measurement, Selection and Placement Center (ÖSYM) as part of Türkiye's national higher education admission system [18, 19]. The questions included in this study were obtained from official examination booklets, and the corresponding official answer keys were used as the reference standard. In the present study, the term 'nationally standardized examination questions' refers to their central preparation, uniform national administration, and official scoring procedures. It does not imply that question difficulty or psychometric properties were equated across examination years.

Three LLMs were assessed: ChatGPT-5, Gemini 2.5 Flash, and DeepSeek V3.1. All three models were evaluated on 24 August 2025. The models were accessed through their publicly available consumer web interfaces, and no API endpoints were used. Model names and versions

were recorded as displayed by the respective platforms at the time of testing. Backend build identifiers were not exposed through these consumer interfaces and therefore could not be recorded. Each model was asked to solve the same set of 84 physics questions under standardized conditions. To reduce the potential influence of question order, the questions were presented in randomized order rather than chronologically. Each session was started in a new private/incognito browser window or a reset conversation environment to minimize carry-over effects from previous interactions.

To reduce contextual bias, the models were not informed that the questions were taken from a university entrance examination. Each question was presented using the same neutral student-style prompt: 'I could not solve this question, can you solve it for me?' This approach was used to simulate a typical student–AI interaction while avoiding additional cues about the source, level, or examination context of the questions.

Model responses were evaluated solely on the basis of answer correctness. Correct answers were scored as 1 point, and incorrect answers were scored as 0 points. No partial credit was awarded. The official answer keys were used as the reference standard for determining whether each response was correct. Explanation quality, linguistic clarity, reasoning depth, and pedagogical usefulness were not assessed in the present study.

For each model, the total number of correct and incorrect answers was recorded. Net scores were calculated using the correction formula commonly applied in Turkish multiple-choice examinations, in which four incorrect answers reduce the score by one correct answer:

Net score = Number of correct answers − (Number of incorrect answers/4)

Correct-answer accuracy was calculated as the proportion of correct answers among all evaluated questions. Correct-answer proportions were reported with 95% confidence intervals calculated using the Wilson score method. In addition, net-score-based success rates were calculated by dividing the net score by the total number of questions. Yearly correct-answer accuracy was also calculated to examine performance variation across the 2016–2025 examination period.

Each question was additionally classified according to its primary physics topic based on the dominant physics concept required for its solution. The predefined categories were mechanics, electricity and magnetism, optics, heat and thermodynamics, waves, modern physics/nuclear physics, and other. The other category included questions primarily addressing the nature of science, measurement and SI units, and properties of matter such as density, buoyancy, and surface tension. Topic-specific correct-answer frequencies and percentages were calculated separately for each model.

Statistical analyses were performed to compare the models' correct-answer performance. Since the same 84 questions were presented to all three models and each response was coded as correct or incorrect, Cochran's Q test was used to compare the paired binary outcomes across the three related groups [20]. If a statistically significant overall difference had been detected, pairwise McNemar tests with Bonferroni correction were planned for post-hoc comparisons. Associations between primary physics topic and answer correctness were evaluated separately for each model. Because several expected cell counts were below 5, the Fisher–Freeman–Halton exact test with Monte Carlo estimation based on 100 000 sampled tables was used for these topic-based analyses. A p value below 0.05 was considered statistically significant.

3. Results

A total of 84 physics questions from Turkish university entrance examinations were evaluated using three LLMs: ChatGPT-5, Gemini 2.5 Flash, and DeepSeek V3.1. The dataset comprised 28 YGS questions from 2016–2017 and 56 TYT questions from 2018–2025. The models were compared according to the number of correct answers, incorrect answers, correct-answer accuracy, net scores, and net-score-based success rates. The overall performance results are presented in table 1.

As shown in table 1, DeepSeek V3.1 demonstrated the highest numerical performance,

Table 1. Overall performance of the three large language models on Turkish university entrance physics questions administered between 2016 and 2025.

Model	Correct answers, *n*	Incorrect answers, *n*	Correct-answer accuracy, % (95% CI)	Net score	Net-score-based success rate, %
ChatGPT-5	55	29	65.5 (54.8–74.8)	47.75	56.8
Gemini 2.5 Flash	58	26	69.0 (58.5–77.9)	51.50	61.3
DeepSeek V3.1	66	18	78.6 (68.7–86.0)	61.50	73.2

Note: A total of 84 questions were evaluated, comprising 28 YGS physics questions from 2016–2017 and 56 TYT physics questions from 2018–2025. Correct-answer accuracy was calculated as the number of correct answers divided by the total number of questions. The 95% confidence intervals for correct-answer accuracy were calculated using the Wilson score method. Net score was calculated as correct answers − (incorrect answers/4). Net-score-based success rate was calculated by dividing the net score by the total number of questions. CI, confidence interval; YGS, *Yükseköğretime Geçiş Sınavı*; TYT, *Temel Yeterlilik Testi*.

Table 2. Yearly correct-answer accuracy of the three large language models on YGS and TYT physics questions administered between 2016 and 2025.

Year	Questions, *n*	ChatGPT-5, correct *n* (%)	Gemini 2.5 Flash, correct *n* (%)	DeepSeek V3.1, correct *n* (%)
2016 (YGS)	14	11 (78.6)	10 (71.4)	9 (64.3)
2017 (YGS)	14	9 (64.3)	12 (85.7)	13 (92.9)
2018 (TYT)	7	4 (57.1)	4 (57.1)	6 (85.7)
2019 (TYT)	7	7 (100.0)	5 (71.4)	6 (85.7)
2020 (TYT)	7	5 (71.4)	5 (71.4)	4 (57.1)
2021 (TYT)	7	2 (28.6)	5 (71.4)	5 (71.4)
2022 (TYT)	7	4 (57.1)	6 (85.7)	5 (71.4)
2023 (TYT)	7	3 (42.9)	3 (42.9)	5 (71.4)
2024 (TYT)	7	3 (42.9)	2 (28.6)	7 (100.0)
2025 (TYT)	7	7 (100.0)	6 (85.7)	6 (85.7)

Note: Questions from 2016 and 2017 were drawn from the YGS, whereas questions from 2018 to 2025 were drawn from the TYT, the first session of the YKS. Yearly percentages represent the proportion of correct answers within each year's question set. The 2016 and 2017 examinations included 14 physics questions each, whereas each examination from 2018 to 2025 included 7 physics questions. YGS, Yükseköğretime Geçiş Sınavı; TYT, Temel Yeterlilik Testi; YKS, Yükseköğretim Kurumları Sınavı.

answering 66 questions correctly and 18 incorrectly. This corresponded to a correct-answer accuracy of 78.6% (95% CI 68.7%–86.0%), a net score of 61.5, and a net-score-based success rate of 73.2%. Gemini 2.5 Flash ranked second, with 58 correct and 26 incorrect answers, corresponding to a correct-answer accuracy of 69.0% (95% CI 58.5%–77.9%), a net score of 51.5, and a net-score-based success rate of 61.3%. ChatGPT-5 answered 55 questions correctly and 29 incorrectly, corresponding to a correct-answer accuracy of 65.5% (95% CI 54.8%–74.8%), a net score of 47.75, and a net-score-based success rate of 56.8%.

To compare the correct-answer performance of the three models on the same set of 84 questions, Cochran's Q test was performed. Although DeepSeek V3.1 had the highest correct-answer rate, the difference among the three models did not reach statistical significance (Cochran's $Q = 4.732$, df $= 2$, $p = 0.094$). Therefore, no post-hoc pairwise McNemar comparisons were performed.

The yearly distribution of correct-answer accuracy is summarized in table 2. Across the ten-year period, all three models showed noticeable year-to-year variability.

Yearly performance varied substantially across the examination period (table 2 and figure 1). ChatGPT-5 ranged from 28.6% to 100.0%, Gemini 2.5 Flash from 28.6% to 85.7%, and DeepSeek V3.1 from 57.1%

to 100.0%, with no model consistently achieving the highest accuracy across all years.

Model performance according to primary physics topic is presented in table 3. Fisher–Freeman–Halton exact tests with Monte Carlo estimation showed no statistically significant association between physics topic and answer correctness for ChatGPT-5 ($p = 0.717$), Gemini 2.5 Flash ($p = 0.187$), or DeepSeek V3.1 ($p = 0.906$). The modern physics/nuclear physics category contained only one question and was therefore interpreted descriptively.

Overall, DeepSeek V3.1 ranked first in total numerical performance, followed by Gemini 2.5 Flash and ChatGPT-5. However, the overall differences among models were not statistically significant. Although descriptive differences were observed across physics topics and examination years, physics topic was not significantly associated with answer correctness for any of the three models.

4. Discussion

This study compared three LLMs on Turkish university entrance physics questions administered between 2016 and 2025. DeepSeek V3.1 achieved the highest numerical performance, followed by Gemini 2.5 Flash and ChatGPT-5; however, the overall difference was not statistically significant. Thus, DeepSeek V3.1's higher numerical accuracy should not be interpreted as definitive evidence of superiority, particularly because LLM performance may vary with task characteristics and reasoning demands [11, 12].

All three models also showed substantial year-to-year variability (table 2 and figure 1). These fluctuations may reflect differences in conceptual difficulty, mathematical reasoning demands, and representation requirements, although the present study was not designed to isolate these factors. Similar task-dependent variability has been reported in previous physics education studies [21–23].

The topic-based analysis provided an additional perspective on this variability. Descriptive differences were observed across physics domains: for example, Gemini 2.5 Flash performed particularly well in Heat and Thermodynamics, whereas DeepSeek V3.1 showed its highest accuracy in Optics. Nevertheless, the Fisher–Freeman–Halton exact analyses showed no statistically significant association between primary physics topic and answer correctness for any of the three models. Therefore, the observed topic-specific differences should be interpreted descriptively rather than as evidence that any model is systematically stronger or weaker in a particular area of physics. The small number of questions in some categories, particularly waves and modern physics/nuclear physics, also limits the precision of topic-level comparisons.

From a physics education perspective, this variability is important because problem solving requires conceptual understanding, appropriate representation, selection of relevant principles, mathematical reasoning, and evaluation of the physical meaning of results [5, 24]. The observed fluctuations suggest that high overall accuracy does not necessarily indicate consistent reliability across all areas or formats of school-level physics [21, 24].

The findings also highlight the potential role of AI tools as supplementary resources in physics learning. Students may benefit from using LLMs to check answers, receive alternative explanations, or practise problem-solving outside the classroom, as previous studies have suggested that LLMs may support individualized learning and provide accessible educational assistance [2, 25]. For teachers, these tools may provide examples, support discussion, or help identify how students might interact with AI-generated solutions, particularly when used to foster critical reflection rather than passive answer acceptance [26, 27]. However, the present results also show that AI-generated answers should not be accepted without verification. Teachers should encourage students to question the reasoning behind AI responses, compare them with classroom instruction, and focus on understanding the underlying physics rather than simply obtaining the final answer.

Another important point is that this study evaluated only final-answer correctness. In educational practice, correctness alone is not sufficient to determine the usefulness of an AI-generated solution. Previous physics education

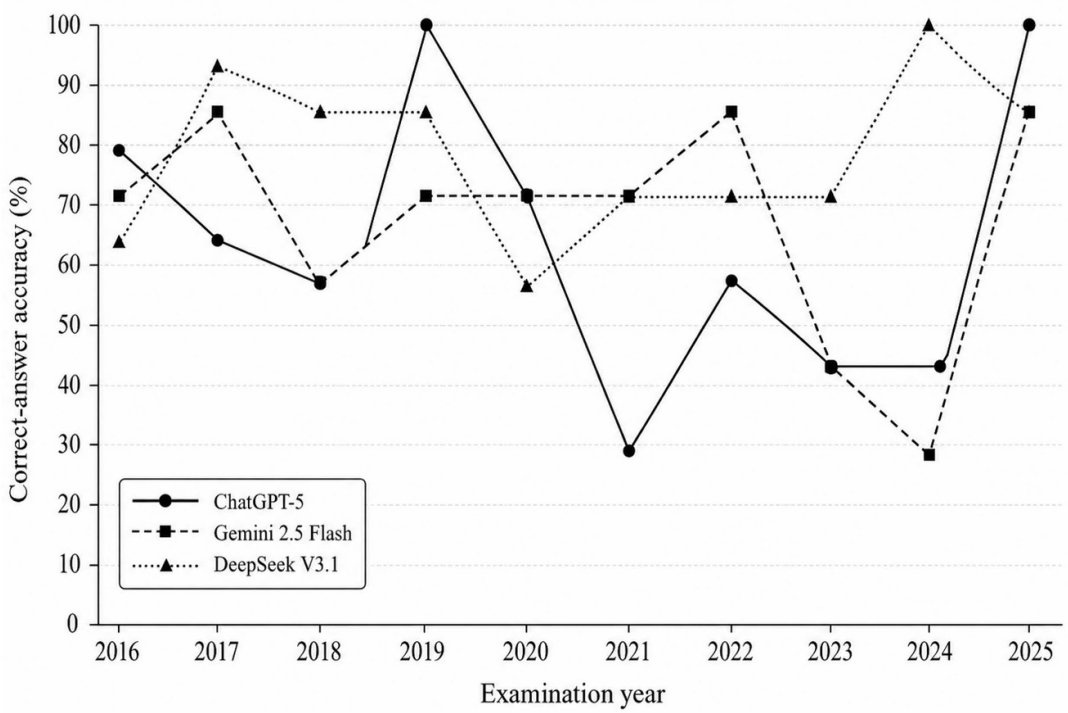

Figure 1. Annual correct-answer accuracy of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek V3.1 on Turkish university entrance physics questions administered between 2016 and 2025. Questions from 2016–2017 were drawn from the YGS, whereas questions from 2018–2025 were drawn from the TYT. Each point represents the percentage of questions answered correctly by the corresponding model in that examination year. Abbreviations: AI: artificial intelligence; YGS: *Yükseköğretime Geçiş Sınavı*; TYT: *Temel Yeterlilik Testi*; YKS: *Yükseköğretim Kurumları Sınavı*.

Table 3. Correct-answer performance of the three large language models according to primary physics topic.

Physics topic	Questions, *n*	ChatGPT-5, correct *n* (%)	Gemini 2.5 Flash, correct *n* (%)	DeepSeek V3.1, correct *n* (%)
Mechanics	24	16 (66.7)	18 (75.0)	19 (79.2)
Electricity and magnetism	15	8 (53.3)	11 (73.3)	11 (73.3)
Optics	15	9 (60.0)	7 (46.7)	13 (86.7)
Heat and thermodynamics	12	7 (58.3)	11 (91.7)	10 (83.3)
Waves	6	5 (83.3)	4 (66.7)	4 (66.7)
Modern physics/nuclear Physics	1	1 (100.0)	1 (100.0)	1 (100.0)
Other	11	9 (81.8)	6 (54.5)	8 (72.7)
Overall	**84**	**55 (65.5)**	**58 (69.0)**	**66 (78.6)**

Note: Percentages represent the proportion of questions answered correctly within each physics-topic category. Because several expected cell counts were below 5, associations between physics topic and answer correctness were evaluated separately for each model using the Fisher–Freeman–Halton exact test with Monte Carlo estimation based on 100 000 sampled tables. No statistically significant association was observed for ChatGPT-5 ($p = 0.717$), Gemini 2.5 Flash ($p = 0.187$), or DeepSeek V3.1 ($p = 0.906$). The modern physics/nuclear physics category contained only one question and should therefore be interpreted descriptively.

research has shown that students' perceptions of AI-generated responses may depend not only on scientific accuracy but also on linguistic quality, and that incomplete or misleading responses can still appear plausible to learners [28]. A model may provide the correct answer with incomplete or unclear reasoning, while another model may give an incorrect answer despite offering a partially useful explanation. This distinction is supported by Dunlap *et al*, who evaluated 165 GPT-4 responses to an introductory inclined-plane problem: although 89.7% of responses contained no mathematical error, only 42.4% had both the physical setup and final numerical answer in agreement with expert physicists [29]. Thus, a correct option may occasionally be reached through flawed, incomplete, or physically inconsistent reasoning, resulting in a 'hallucinated correct answer' that would still be counted as correct under the scoring approach used in the present study. For secondary-school students who may not yet be able to identify subtle reasoning errors, this can substantially reduce the educational reliability of an otherwise convincing AI-generated response. For physics teachers, the quality of reasoning, conceptual clarity, and pedagogical value of an explanation may be as important as the final answer itself, especially when AI tools are used to support teaching or task development [26]. Therefore, future studies should evaluate not only whether AI models can select the correct answer, but also whether they can explain the reasoning process in a way that supports meaningful learning and provides useful feedback [30].

This study has several limitations. First, the study included only physics questions from two successive Turkish university entrance examination formats—YGS and TYT—so the findings may not be directly generalizable to other national examinations, curricula, or educational systems. Second, the analysis was limited to a single corpus of official national examination questions and a single prompt format. Previous studies have shown that LLM outputs in physics and multiple-choice contexts may be sensitive to prompt design, question format, and even answer-option ordering, suggesting that different prompts, follow-up questions, or tutoring-style interactions may produce different results [11, 31]. Third, only specific versions of ChatGPT-5, Gemini 2.5 Flash, and DeepSeek V3.1 were evaluated during a defined time period. Since AI models are frequently updated, their performance may change over time. In addition, some physics-topic categories contained relatively few questions, including only one modern physics/nuclear physics item, limiting the statistical power and precision of topic-specific comparisons. Finally, explanation quality, conceptual accuracy, and reasoning depth were not assessed in this study.

5. Conclusion

LLMs were able to solve a substantial proportion of Turkish university entrance physics questions, with DeepSeek V3.1 showing the highest numerical performance. However, the overall difference among models was not statistically significant, and performance varied across examination years. Although descriptive differences were observed across physics topics, topic was not significantly associated with answer correctness for any model. These findings suggest that AI tools may support physics learning and exam preparation, but their responses should be used critically and under teacher guidance. Future studies should assess not only answer accuracy but also reasoning quality and pedagogical value.

Data availability statement

The official YGS examination booklets and answer keys from 2016–2017 and the official TYT examination booklets and answer keys from 2018–2025 are publicly accessible through the ÖSYM archives. The coded response dataset generated and analysed during this study is available from the corresponding author upon reasonable request.

All data that support the findings of this study are included within the article (and any supplementary files).

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics statement

This study used publicly available examination questions and official answer keys. It did not involve human participants, personal data, or identifiable information. Therefore, ethical approval and informed consent were not required.

Funding statement

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

Author contributions

Güler Çörekci 0009-0005-3890-312X
Conceptualization (lead), Data curation (lead), Supervision (lead)

Ahmet Murat Çörekci 0009-0008-7505-8449
Formal analysis (lead), Methodology (equal), Resources (equal), Writing – original draft (equal), Writing – review & editing (equal)

Received 10 May 2026, in final form 4 August 2026
Accepted for publication 19 August 2026
https://doi.org/10.1088/1361-6552/ae9bf4

References

[1] Lo C K 2023 What is the impact of ChatGPT on education? A rapid review of the literature *Educ. Sci.* **13** 410
[2] Kasneci E *et al* 2023 ChatGPT for good? On opportunities and challenges of large language models for education *Learn. Individ. Differ.* **103** 102274
[3] Hestenes D, Wells M and Swackhamer G 1992 Force concept inventory *Phys. Teach.* **30** 141–58
[4] McDermott L C and Redish E F 1999 Resource letter: PER-1: physics education research *Am. J. Phys.* **67** 755–67
[5] Chi M T H, Feltovich P J and Glaser R 1981 Categorization and representation of physics problems by experts and novices *Cogn. Sci.* **5** 121–52
[6] Kortemeyer G 2023 Could an artificial-intelligence agent pass an introductory physics course? *Phys. Rev. Phys. Educ. Res.* **19** 010132
[7] Cooper G 2023 Examining science education in ChatGPT: an exploratory study of generative artificial intelligence *J. Sci. Educ. Technol.* **32** 444–52
[8] Mollick E and Mollick L 2023 Assigning AI: seven approaches for students, with prompts *arXiv (Cornell University)* 10.48550/arxiv.2306.10052
[9] Ji Z *et al* 2022 Survey of hallucination in natural language generation *ACM Comput. Surv.* **55** 1–38
[10] Bender E M *et al* 2021 On the dangers of stochastic parrots 610–23
[11] Polverini G and Gregorcic B 2023 How understanding large language models can inform the use of ChatGPT in physics education *Eur. J. Phys.* **45** 25701
[12] Gregorcic B and Pendrill A-M 2023 ChatGPT and the frustrated socrates *Phys. Educ.* **58** 35021
[13] Kiesler N and Schiffner D 2023 Large language models in introductory programming education: ChatGPT's performance and implications for assessments *arXiv (Cornell University)* 10.48550/arxiv.2308.08572
[14] Kung T H *et al* 2023 Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models *PLOS Digit. Health* February 9 **2** e0000198
[15] Frieder S *et al* 2023 Mathematical capabilities of ChatGPT *arXiv (Cornell University)* 31 January 10.48550/arxiv.2301.13867
[16] Yeadon W, Inyang O, Mizouri A, Peach A and Testrow C P 2023 The death of the short-form physics essay in the coming AI revolution *Phys. Educ.* **58** 35027
[17] Tschisgale P, Maus H, Kieser F, Kroehs B, Petersen S and Wulff P 2025 Evaluating GPT- and reasoning-based large language models on physics olympiad problems: surpassing human performance and implications for educational assessment *Phys. Rev. Phys. Educ. Res.* **21** 020115
[18] Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı 2025 2025-YKS (TYT, AYT ve YDT) Temel Soru Kitapçıkları ve Cevap Anahtarları (available at: www.osym.gov.tr/TR,33280/2025-yks-tyt-ayt-ve-ydt-temel-soru-kitapciklari-ve-cevap-anahtarlari.html) (Accessed 10 May 2026)
[19] Ölçme, Seçme ve Yerleştirme Merkezi Başkanlığı 2025 Yükseköğretim Kurumları Sınavı (available at: www.osym.gov.tr/SinavGrubu/Menu/323) (Accessed 17 August)
[20] Cochran W G 1950 The comparison of percentages in matched samples *Biometrika* **37** 256–66
[21] Polverini G and Gregorcic B 2024 Performance of ChatGPT on the test of understanding graphs in kinematics *Phys. Rev. Phys. Educ. Res.* **20** 010109

[22] Liang Y, Zou D, Xie H and Wang F L 2023 Exploring the potential of using ChatGPT in physics education *Smart Learn. Environ.* **10** 52

[23] Simó V L and Rezende M F 2024 Challenging ChatGPT with different types of physics education questions *Phys. Teach.* **62** 290–4

[24] Hegde B and Meera B 2012 How do they solve it? An insight into the learner's approach to the mechanism of physics problem solving *Phys. Rev. Spec. Top. Phys. Educ. Res.* **8** 010109

[25] Tlili A, Shehata B, Adarkwah M A, Bozkurt A, Hickey D T, Huang R and Agyemang B 2023 What if the devil is my guardian angel: chatGPT as a case study of using chatbots in education *Smart Learn. Environ.* **10** 15

[26] Küchemann S, Steinert S, Revenga N, Schweinberger M, Dinc Y, Avila K E and Kuhn J 2023 Can ChatGPT support prospective teachers in physics task development? *Phys. Rev. Phys. Educ. Res.* **19** 020128

[27] Bitzenbauer P 2023 ChatGPT in physics education: a pilot study on easy-to-implement activities *Contemp. Educ. Technol.* **15** ep430

[28] Dahlkemper M N, Lahme S Z and Klein P 2023 How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT *Phys. Rev. Phys. Educ. Res.* **19** 010142

[29] Dunlap J C, Sissons R and Widenhorn R 2025 Descending an inclined plane with a large language model *Phys. Rev. Phys. Educ. Res.* **21** 010153

[30] Chen Z and Wan T 2025 Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy *Phys. Rev. Phys. Educ. Res.* **21** 010126

[31] Pezeshkpour P and Hruschka E 2024 Large language models sensitivity to the order of options in multiple-choice questions 2006–17