

Research Article

Designing a Cross-Domain Benchmark for Large Language Models in Undergraduate Physics Education: A Monte Carlo Simulation Study

Konstantinos T. Kotsis^{1,*} ¹ Lab of Physics Education and Teaching, Department of Primary Education, University of Ioannina, Greece* Correspondence: : kkotsis@uoi.gr<https://doi.org/10.59652/p5j07k92>

Abstract: The rapid integration of large language models into education has created an urgent need for rigorous benchmarking designs, but a manuscript cannot report model performance as observed when no empirical response corpus has been collected. This article, therefore, presents a fully disclosed Monte Carlo simulation study that demonstrates how a future cross-domain physics benchmark could be designed, analysed, and interpreted. The simulated design comprises 150 hypothetical undergraduate physics problems distributed equally across Classical Mechanics, Electromagnetism, Thermodynamics and Statistical Physics, Waves and Optics, and Modern Physics; three difficulty levels; and balanced, well-specified and underspecified conditions. Four generic model profiles (Models A-D) are evaluated in 1,000 independent replications, producing 600 simulated problem-by-model observations per replication. Outcomes include final-answer correctness, conceptual accuracy, mathematical accuracy, explanation quality, and a four-category error taxonomy. Under the declared generating assumptions, the mean simulated correctness is 59.3%, with performance declining from 75.5% for introductory problems to 40.4% for advanced problems. The lowest simulated rate occurs for underspecified advanced tasks (25.1%), while model construction failure and conceptual misapplication dominate the distribution of generated errors. These quantities are synthetic outputs, not estimates of any commercial system's performance. The study contributes a transparent design template, reproducible assumptions, and an analysis plan that can be replaced with observed data in a future preregistered benchmark.

Keywords: large language models; physics education; undergraduate physics; Monte Carlo simulation; benchmark design

Received: June 8, 2026
Accepted: September 20, 2026
Published: September 30, 2026



Copyright: © 2022 by the authors.

Submitted for open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license

(<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The public release of ChatGPT in November 2022 initiated an unprecedented disruption across educational practice, introducing a generative artificial intelligence tool into classrooms, laboratories, and independent study environments with a speed and scale unmatched by any prior educational technology (Korzynski et al., 2023; Küchemann et al., 2025). Within a remarkably short period, large language models have transitioned from novelty applications to ubiquitous study companions, employed by students across all educational levels for tasks ranging from essay composition and code generation to solving complex disciplinary problems (Ansari et al., 2024; Lo, 2023). Nowhere has this transformation provoked more urgent pedagogical reflection than in physics education, where problem-solving has historically served as both the central learning activity and the primary epistemic gateway through which students develop scientific understanding (Hsu et al., 2004; Wang et al., 2024).

Physics problem-solving occupies a distinctive cognitive territory among academic disciplines. Unlike domains in which the production of fluent text or plausible argumentation may suffice, physics demands the simultaneous coordination of conceptual reasoning, mathematical formalism, model construction, and the disciplined application of physical principles to novel contexts (Chi et al., 1981; Hestenes et al., 1992). Decades of research on expert and novice performance have established that expert physicists are distinguished not by the accumulation of solution procedures but by the capacity to construct coherent mental

models of physical systems, identify governing principles, and translate qualitative reasoning into quantitative predictions (Larkin et al., 1980; Singh et al., 2023). Novice learners, by contrast, typically exhibit fragmented conceptual structures, persistent misconceptions, and a reliance on surface-feature matching and formula-hunting strategies that mask shallow understanding (diSessa, 1993; Hestenes et al., 1992). The introduction of a tool capable of producing fluent, often correct, and superficially expert-like solutions to a wide range of physics problems, therefore, raises questions that strike at the very foundation of how physics is taught, learned, and assessed.

Recent empirical investigations have begun to map the capabilities and limitations of large language models in physics contexts, and the picture that has emerged is both instructive and deeply ambiguous. Kortemeyer (2023) provided the first systematic evaluation, demonstrating that ChatGPT, built on GPT-3.5, could narrowly pass an introductory calculus-based physics course while displaying many of the preconceptions and errors characteristic of beginning learners. West (2023) subsequently showed that GPT-4 represented a substantial leap in apparent conceptual competence, approaching expert-level performance on the Force Concept Inventory (FCI). Wang et al. (2024) offered a more fine-grained analysis, reporting that GPT-4 successfully solved 62.5% of well-specified college-level engineering physics problems. Still, only 8.3% of underspecified, real-world problems have failures stemming from an inability to construct accurate physical models, to generate reasonable assumptions about missing data, and to perform reliable calculations. López-Simó and Rezende (2024) further demonstrated that ChatGPT's performance varies markedly across question types, performing adequately on conceptual multiple-choice items but exhibiting characteristic failure modes when confronted with problems requiring authentic physical modelling. More recently, Tschisgale et al. (2025) evaluated both standard and reasoning-optimised large language models on Physics Olympiad problems, finding that the latter surpassed the median performance of human participants, yet continued to struggle with tasks requiring spatial reasoning and diagrammatic interpretation. Complementary benchmarking efforts, such as the UGPhysics benchmark comprising problems across multiple undergraduate physics domains, have confirmed that even the highest-performing models achieve overall accuracies below 50% on rigorous, domain-balanced assessments (X. Xu et al., 2025). Taken together, these findings suggest that large language models occupy an ambiguous epistemic position: they are neither reliable physics tutors nor incompetent calculators, but rather idiosyncratic reasoners whose outputs may be confidently correct, plausibly wrong, or, most worryingly, fluently incorrect in ways that are difficult for novice learners to detect.

Despite the rapid growth of this benchmarking literature, several substantive gaps remain. First, existing studies have typically evaluated a single model or a single model family against a narrow range of physics tasks, most commonly Introductory Mechanics problems or FCI items. Relatively little is known about how multiple contemporary models compare across the full breadth of undergraduate physics, including electromagnetism, thermodynamics, waves and optics, and modern physics, or about how performance varies systematically with problem difficulty and the degree of problem specification. Second, while individual studies have identified recurring error types, the field lacks a comprehensive error taxonomy that classifies and quantifies the characteristic failure modes of large language models in physics with sufficient granularity to inform both model development and pedagogical design. Third, and most consequentially for physics education, the relationship between the technical performance profile of these models and the pedagogical risks they pose to student learning remains undertheorised. A model that fails predominantly through conceptual misapplication poses a qualitatively different pedagogical threat from one that fails through computational error, yet existing benchmarking studies rarely draw this connection explicitly (Tschisgale et al., 2025; Moundridou et al., 2026).

The present study addresses these gaps by developing and computationally evaluating a prospective protocol for benchmarking large language models in undergraduate physics. Its purpose is not to estimate the performance of existing commercial systems but to examine the structure, precision, and robustness of the design proposed for a future empirical benchmark. The protocol uses a balanced factorial architecture comprising 150 hypothetical problems across five physics domains, three difficulty levels, and two specification conditions.

Four generic performance profiles are included solely to evaluate whether the proposed design and analysis pipeline can recover declared contrasts across benchmark conditions. No prompts were submitted to GPT-4o, GPT-4.5, Claude, Gemini, or any other commercial system; no model-generated solutions were collected; and no human raters coded responses. Instead, 1,000 Monte Carlo replications were used to evaluate the behaviour of the proposed benchmark under a fully declared reference scenario.

Accordingly, the study addresses three methodological research questions:

RQ1. How can physics domain, problem difficulty, specification status, and hypothetical performance-profile differences be operationalised within a balanced and reproducible undergraduate physics benchmark?

RQ2. Under the reference parameterisation, what replication-level precision does a 150-problem benchmark provide for estimating overall correctness, profile differences, difficulty, specification effects, and the difficulty-by-specification interaction?

RQ3. How sensitive are the principal correctness patterns and benchmark-design conclusions to alternative assumptions about baseline capability, the under-specification penalty, the advanced-by-underspecified interaction, and problem-level heterogeneity?

The first research question concerns the construction and internal coherence of the benchmark architecture. The second evaluates whether a single implementation of the proposed design would estimate the primary contrasts with adequate precision. The third examines whether the conclusions remain stable when central assumptions of the generating model are varied. Together, these questions shift the purpose of the simulation from reproducing programmed performance gradients to evaluating the structure, precision, and robustness of the proposed empirical benchmark.

The contribution is therefore methodological and prospective rather than empirical. First, the study translates theoretical distinctions from physics problem-solving research into an explicit factorial benchmark structure. Second, it evaluates the precision of the proposed 150-problem design in estimating the primary effects and interactions. Third, it examines the sensitivity of the principal conclusions to alternative parameter assumptions. Fourth, it demonstrates how final-answer correctness can be analysed alongside conceptual accuracy, mathematical accuracy, explanation quality, and a provisional four-category error taxonomy. The numerical results are implementation checks and design-evaluation outcomes generated under hypothetical parameterisations. They should not be interpreted as estimates, predictions, or rankings of contemporary large language models

2. Theoretical Background

The systematic evaluation of how large language models perform on physics problems requires a theoretical apparatus that can account for what constitutes competent physics problem-solving, what distinguishes expert from novice performance, and why the specific patterns of success and failure exhibited by generative AI carry pedagogical significance. This section develops that apparatus by drawing on two interrelated bodies of research: the classical literature on expert and novice problem-solving in physics, and more recent work on the nature of conceptual understanding and the structure of physics knowledge. The synthesis provides the analytical lens through which the benchmarking findings reported in subsequent sections are interpreted.

The foundational research programme on physics problem-solving was initiated by Larkin et al. (1980) and Chi et al. (1981), whose comparative studies of expert and novice performance established findings that have shaped the field for more than four decades. The central insight of this programme is that the difference between expert and novice problem-solvers is not principally a difference in raw cognitive capacity, computational speed, or the number of formulae that can be recalled, but rather a difference in the organisation, accessibility, and deployment of disciplinary knowledge. In a series of sorting experiments, Chi et al. (1981) demonstrated that expert physicists categorise problems according to deep structural features rooted in physical principles, such as conservation of energy, Newton's second law, or superposition. In contrast, novices categorise the same problems according to surface features, such as the presence of inclined planes, springs, or pulleys. This difference in categorisation reflects a profound asymmetry in knowledge organisation: experts possess richly interconnected schemas in long-term memory that allow them to chunk multiple pieces

of physical and mathematical information into coherent representational units, whereas novices process the same information as discrete and largely isolated elements (Singh et al., 2023). The expert's knowledge, as Larkin et al. (1980) put it, forms complex schemata that guide the interpretation and solution of a problem and constitute a large part of what is colloquially termed physical intuition.

The strategic consequences of this asymmetry in knowledge organisation are well documented and directly relevant to the evaluation of large language model performance. Novices typically employ means-ends analysis, working backwards from the unknown quantity by selecting equations that contain the target variable and substituting iteratively until a solvable algebraic form emerges (Sweller, 1988). This backwards-chaining strategy, while sometimes effective for well-specified textbook problems, is cognitively demanding, error-prone, and ill-suited to problems that require constructing physical models or identifying appropriate simplifying assumptions. Experts, by contrast, characteristically work forward from a qualitative analysis of the physical situation, selecting principles before equations and deferring quantitative manipulation until a coherent conceptual framing of the problem has been established (Dhillon, 1998; Larkin et al., 1980; Simon & Simon, 1978). This forward-working strategy depends on the expert's capacity to recognise deep structural features of the problem and to activate the relevant schemas rapidly and with minimal attentional cost. The pedagogical implication, repeatedly emphasised in the physics education research literature, is that effective instruction must support the construction of organised conceptual schemas rather than merely the accumulation of solution procedures, and must cultivate the metacognitive habits of qualitative analysis, principle selection, and self-monitoring that characterise expert practice (Hsu et al., 2004; Singh et al., 2023).

A second body of theory addresses the nature of conceptual understanding itself and the processes by which it develops. Hestenes et al. (1992), through the development of the FCI, demonstrated that traditional physics instruction often leaves students' intuitive, pre-Newtonian conceptions of force and motion largely untouched, even when their procedural performance on quantitative problems appears satisfactory. This finding catalysed decades of research on conceptual change in physics, much of which has converged on the view that physics learning is not the simple replacement of misconceptions with correct concepts but rather the gradual reorganisation and refinement of an initially fragmented conceptual ecology. diSessa's (1993) knowledge-in-pieces perspective has proven particularly influential in this regard. According to this view, students approach physics equipped with a multitude of intuitive cognitive elements, termed phenomenological primitives or p-prims, that are activated in context and gradually cohere into more systematic conceptual structures through extended engagement with disciplinary practice. Conceptual development, on this account, is not a matter of erasing prior knowledge but of progressively coordinating, contextualising, and refining it through the learner's own active engagement with physical reasoning. This issue has taken on renewed importance with generative AI, which can support explanation and conceptual engagement but can also reinforce misconceptions when learners accept fluent yet conceptually inaccurate responses without sufficient critical evaluation (Kotsis, 2026).

The conjunction of the expert-novice literature and conceptual-change theory provides the rationale for the variables encoded in the simulation. If expert physics performance depends on organised conceptual schemas and qualitative, principle-driven reasoning, then a future benchmark should not be restricted to final numerical accuracy. It should distinguish final-answer correctness from conceptual accuracy, mathematical execution, explanatory quality, and the ability to construct a defensible physical model when information is incomplete. The simulation operationalises this theoretical expectation by assigning lower generating probabilities to underspecified and advanced tasks, especially when those conditions occur together. This is an assumption informed by prior literature, not a finding of the present study. Its purpose is to make the expected signature explicit and testable in subsequent empirical work.

A related consideration concerns the structure of possible errors. Novice errors in physics are systematic rather than random, and prior studies of large language models have also distinguished failures of principle selection, model construction, calculation, and fabricated reasoning (Hestenes et al., 1992; Wang et al., 2024). The present simulation,

therefore, uses a four-category taxonomy comprising conceptual misapplication, model-construction failure, mathematical or computational error, and hallucinated reasoning. Conditional category probabilities are programmed to vary modestly across domains and hypothetical model profiles. These programmed distributions allow the manuscript to illustrate how an empirical corpus could be summarised; they do not demonstrate that real systems possess the simulated error profile.

Together, these theoretical perspectives establish the prospective evaluative framework. The simulation is designed not merely to generate binary success values but to represent several dimensions that physics education research identifies as central to expert performance: coherent conceptual framing, appropriate principle selection, defensible model construction, mathematical execution, and clear explanation. By translating theory into declared probability parameters, the design exposes its assumptions to scrutiny. An empirical follow-up could preserve the same factorial structure and rubric, replace every synthetic outcome with independently coded model responses, and directly test whether the anticipated gradients appear.

3. Literature Review

The body of research on the performance of large language models in physics has expanded rapidly since the public release of ChatGPT in late 2022. Alongside this empirical literature, recent scholarship has situated artificial intelligence in physics education within an integrated pedagogy-technology-policy framework, emphasising that model capabilities must be considered in relation to classroom practice and institutional governance (Kotsis, 2025a; Younis, 2025). What began as exploratory commentary and anecdotal testing has matured, in less than four years, into a substantial empirical literature spanning systematic benchmarking studies, cross-domain performance comparisons, error analyses, and investigations of the pedagogical implications of model output quality (Küchemann et al., 2025; Moundridou et al., 2026). This section synthesises that literature along three thematic strands directly relevant to the present study: the performance trajectory of large language models on physics tasks, the error profiles and failure modes that characterise incorrect model outputs, and emerging benchmarking methodologies and their limitations. The review concludes by identifying the specific gaps that motivate the present investigation.

The earliest systematic evaluations of generative AI in physics focused on determining whether commercially available models could pass standard physics assessments. Arvidsson (2025) provided the foundational study in this line of inquiry, demonstrating that ChatGPT built on GPT-3.5 could narrowly pass an introductory calculus-based physics course while exhibiting many of the preconceptions and procedural errors characteristic of beginning learners. The model achieved approximately 60% accuracy on the FCI, performed adequately on clicker questions and programming exercises, but struggled substantially with open-ended homework problems and end-of-course examinations requiring extended quantitative reasoning. Concurrent and closely subsequent work by West (2023) replicated the FCI assessment with GPT-4 and documented a dramatic improvement: the newer model answered 28 of 30 FCI items correctly, approaching expert-level performance on conceptual questions concerning kinematics and Newtonian dynamics. This leap in apparent conceptual competence between successive model generations established a pattern that has recurred throughout the subsequent literature, in which each new generation of models substantially outperforms its predecessor on standard assessments while continuing to exhibit characteristic difficulties on more demanding tasks.

Yeadon and Hardy (2024) extended the scope of assessment beyond a single course by evaluating GPT-3.5 on 1,337 physics examination questions spanning GCSE, A-Level, and introductory university curricula in the United Kingdom. Their findings revealed a steep performance gradient with academic level: the model scored an average of 83.4% on GCSE questions, 63.8% on A-Level questions, and only 37.4% on university-level questions. Crucially, differences across prompting techniques, including zero-shot, in-context learning, and confirmatory checking, proved statistically insignificant, suggesting that the performance ceiling was determined primarily by the intrinsic difficulty of the physics content rather than by the sophistication of the prompting strategy. This finding is consequential for benchmarking methodology, as it implies that variations in prompt engineering are unlikely

to account for the large performance decrements observed on advanced physics tasks.

The most fine-grained analysis of model performance as a function of problem type was provided by Wang et al. (2024), who systematically evaluated GPT-4 on 40 college-level engineering physics problems stratified by the degree of problem specification. The model successfully solved 62.5% of well-specified problems, in which all data required for solution were explicitly provided, but only 8.3% of underspecified, real-world problems, in which the solver was required to identify missing information, generate reasonable assumptions, and construct an appropriate physical model before proceeding to a quantitative solution. This disparity, the largest performance gap documented in the benchmarking literature, is a central empirical finding for the field. It indicates that the principal limitation of contemporary large language models in physics is not a deficiency in symbolic or numerical computation but rather an inability to perform the qualitative reasoning, model construction, and assumption generation that distinguish expert from novice problem-solving (Larkin et al., 1980; Singh et al., 2023).

Subsequent investigations have confirmed and extended this pattern with newer model generations. López-Simó and Rezende (2024) reported that ChatGPT performed adequately on conceptual multiple-choice items but exhibited characteristic failures when confronted with problems requiring authentic physical modelling, diagrammatic interpretation, or the integration of multiple physical principles in a single solution. Tschisgale et al. (2025) evaluated both standard GPT-class and reasoning-optimised large language models on Physics Olympiad problems, representing the most challenging end of the difficulty spectrum. Their findings were striking in both directions: reasoning-optimised models surpassed the median performance of human Olympiad participants, yet continued to struggle with tasks requiring spatial reasoning, diagram interpretation, and the construction of physical models from unstructured problem descriptions. This dual finding captures the essential tension of the current moment: models can now outperform many human solvers on well-structured competition problems, yet they remain fundamentally limited in precisely the competencies that physics education research has identified as most central to expert performance.

The most recent wave of benchmarking has introduced large-scale, domain-balanced instruments designed to assess model performance with greater comprehensiveness and rigour than earlier single-course or single-assessment studies. X. Xu et al. (2025) developed the UGPhysics benchmark, comprising 5,520 undergraduate physics problems across 13 subject areas, and evaluated 31 leading large language models. The highest overall accuracy achieved was 49.8%, underscoring the substantial gap between current model capabilities and reliable physics reasoning across the full breadth of the undergraduate curriculum. The benchmark also introduced a model-assisted rule-based judgement pipeline for evaluating answer correctness in physics, addressing a methodological challenge that earlier studies had handled less systematically. These developments signal the maturation of the benchmarking literature from ad hoc evaluations of individual models on convenient assessments toward the kind of systematic, replicable, and cross-model comparisons that the field requires.

A second strand of the literature has moved beyond aggregate performance scores to characterise the specific types of errors that large language models produce when they fail on physics tasks. Comparative research has also extended this inquiry beyond a single model by examining how ChatGPT and DeepSeek address misconceptions about physics concepts (Kotsis, 2025b). Wang et al. (2024) identified three principal failure modes: failure to construct an accurate model of the physical situation, failure to generate reasonable assumptions about missing data, and calculation errors. This three-mode classification represented an important advance over earlier work that had simply categorised responses as correct or incorrect, but it was developed from a relatively small corpus of 40 problems concentrated in engineering physics. Aldazharova et al. (2024) as well as Krupp et al. (2024) complemented this work by documenting that students with unrestricted access to ChatGPT frequently failed to detect physics errors in model outputs, even when those errors involved the misapplication of principles the students themselves understood. This finding suggests that the pedagogical risk of model errors depends not only on their frequency but also on their detectability, which, in turn, depends on the relationship between the error type and the student's existing knowledge structure.

The question of hallucinated reasoning, in which models produce confidently stated but

fabricated physical claims with no basis in the problem statement or in physics, has received increasing attention as a distinct failure mode. Küchemann et al. (2024) examined the reliability of ChatGPT-generated concept inventory items in kinematics and found that, while carefully prompted outputs could approximate the quality of human-developed items on some psychometric dimensions, substantial expert oversight remained necessary to ensure validity, to align distractors with documented student difficulties, and to eliminate fabricated physical claims that appeared plausible on the surface. This finding underscores a concern that runs throughout the benchmarking literature: the very fluency that makes large language model outputs convincing is also the property that renders their errors most dangerous, because fluent errors are precisely the errors that novice learners are least equipped to detect.

The broader literature on large language model reasoning failures provides additional context for understanding these physics-specific patterns. Recent surveys have documented that contemporary models exhibit systematic reasoning deficits that are not confined to physics but reflect architectural limitations of the transformer-based approach to language modelling (Y. Xu et al., 2025; Wei et al., 2025). These deficits include sensitivity to problem framing, in which logically equivalent but differently worded prompts produce different answers; anchoring bias, in which early elements of the prompt disproportionately shape the reasoning trajectory; and the absence of robust internal world models capable of supporting causal and spatial reasoning. In physics specifically, the absence of a world model manifests as an inability to simulate physical processes qualitatively, reason about spatial configurations, and evaluate the physical plausibility of intermediate results, all of which are core components of expert physics reasoning (Larkin et al., 1980). The recognition that these deficits are architectural rather than merely parametric suggests that increasing model scale alone may not resolve the most pedagogically consequential failure modes, a conclusion with significant implications for the design of benchmarking studies and for expectations about future model performance.

A third strand of the literature has addressed methodological considerations in the design and interpretation of benchmarking studies. Several limitations of the existing body of work can be identified. First, most studies have evaluated a single model or model family, making it difficult to distinguish model-specific behaviours from architecture-general patterns. Second, the physics content assessed has been heavily skewed toward Introductory Mechanics and the FCI, with electromagnetism, thermodynamics, waves and optics, and modern physics substantially underrepresented. Third, evaluation criteria have varied widely across studies, ranging from simple binary scoring of final answers to more nuanced assessments of solution quality, with few studies employing multidimensional coding rubrics that would permit independent evaluation of conceptual accuracy, mathematical correctness, and explanation quality. Fourth, the rapidly evolving capabilities of large language models mean that any benchmarking result represents a temporal snapshot whose absolute performance figures may be superseded within months.

Tschisgale et al. (2025) have argued persuasively that this temporal instability should motivate a methodological shift from absolute benchmarking to identifying stable patterns of capability and limitation that persist across model generations, a strategy that informs the present study's analytical approach. Moundridou et al. (2026), in a recent systematic review of empirical research on generative AI in physics education, confirmed these gaps and called for studies that examine model performance across multiple physics domains within unified research designs and develop standardised evaluative frameworks to support cumulative theory development.

These strands of literature – performance trajectory, error characterisation, and methodological development – converge on the need for a unified cross-domain design. The present article does not claim to fill that empirical gap with observations. Instead, it converts the proposed benchmark into a simulation study that makes the design, estimands, and analysis decisions inspectable before costly data collection. The resulting framework compares four generic profiles across five undergraduate physics domains, three difficulty levels, and two specification conditions, and illustrates how a multidimensional rubric and error taxonomy could be reported. In this way, the simulation functions as a design study and a reproducible protocol for a future empirical benchmark.

4. Materials and Methods

This section specifies the simulation design, generating assumptions, primary estimates, uncertainty measures, and sensitivity scenarios used to evaluate the proposed benchmark architecture. The analysis addresses three related methodological properties: whether the factorial structure can operationalise the intended physics-task characteristics, whether a 150-problem implementation provides adequate precision for the primary contrasts, and whether the resulting conclusions remain stable under alternative parameter assumptions. All quantities reported in Section 5 are Monte Carlo summaries generated for these design-evaluation purposes. They were not obtained from commercial model interfaces, student participants, expert judges, or an archived response corpus.

The study uses a factorial Monte Carlo design inspired by prior physics benchmarking research (Wang et al., 2024; Tschisgale et al., 2025). Each replication represents one hypothetical implementation of a benchmark. The design simultaneously varies physics domain, difficulty, specification status, and model profile, permitting direct examination of main effects and interactions. One thousand independent replications were generated. Each replication contains 150 hypothetical problems evaluated by four profiles, or 600 problem-by-profile observations; across the complete simulation, this yields 600,000 generated observations. Replication-level summaries, rather than the pooled microdata alone, form the basis of the reported simulation intervals.

The hypothetical problem bank contains 150 items and is balanced across five domains: Classical Mechanics, Electromagnetism, Thermodynamics and Statistical Physics, Waves and Optics, and Modern Physics. Within each domain, 30 items are divided equally among introductory, intermediate, and advanced levels. Each domain-by-difficulty cell contains ten problems: five well-specified and five underspecified. Thus, each replication includes 75 problems per specification condition and 50 per difficulty level. This architecture preserves the curricular breadth of the proposed empirical benchmark while preventing unequal cell sizes from confounding the illustrative comparisons.

The labels, well specified and underspecified, describe hypothetical task characteristics rather than a collected item bank. A well-specified task is defined as one that provides all numerical and conceptual information needed for a determinate solution. An underspecified task requires the solver to identify missing information, state defensible assumptions, or construct an idealised physical model before calculation. Difficulty is treated as an ordinal design factor: introductory tasks require a direct application of one principle, intermediate tasks coordinate multiple principles, and advanced tasks require extended reasoning or more sophisticated mathematical tools. Future empirical work should validate these classifications through expert review before data collection.

Four generic profiles – Model A, Model B, Model C, and Model D – represent plausible but deliberately non-identifying levels of performance. They are not aliases for GPT-4o, GPT-4.5, Claude, Gemini, or any other commercial model. Generic labels avoid implying that simulated values describe products whose versions, interfaces, and policies may change. The profiles differ only through declared probability shifts in the data-generating process: Model B has the largest positive shift, Model C the next largest, Model A serves as the reference, and Model D has a negative shift. The profiles are methodological scenarios, not rankings.

Final-answer correctness is generated from a logistic model. For observation i on hypothetical problem j , $\text{logit}(p_{ij}) = 1.00 + M_i + G_j + L_j + S_j + (L \times S)_j + u_j$. Profile effects are 0.00, 0.35, 0.20, and -0.25 for Models A-D. Domain effects are 0.25 for Classical Mechanics, 0.00 for Electromagnetism, 0.05 for Thermodynamics, -0.10 for Waves and Optics, and -0.35 for Modern Physics. Difficulty effects are 0.60 for introductory, 0.00 for intermediate, and -0.80 for advanced problems. Under specification contributes -0.85, with additional interaction penalties of -0.15 at the intermediate level and -0.55 at the advanced level. A problem-level random intercept u_j is drawn from a normal distribution with mean zero and standard deviation 0.35.

The multidimensional rubric comprises final-answer correctness, conceptual accuracy, mathematical accuracy, and explanation quality. Final-answer correctness is binary. Each of the remaining dimensions is generated as an ordinal variable with three categories: inadequate, partially adequate, and fully adequate. To make the reference scenario exactly reproducible, the conditional probabilities are specified directly. For a correct synthetic response, the

probabilities of an inadequate, partially adequate, and fully adequate conceptual rating are 0.03, 0.12, and 0.85, respectively; for an incorrect response, the corresponding probabilities are 0.55, 0.30, and 0.15. For mathematical accuracy, the probabilities are 0.02, 0.11, and 0.87 when the final answer is correct and 0.18, 0.33, and 0.49 when it is incorrect. For explanation quality, the respective probabilities are 0.05, 0.25, and 0.70 for correct responses and 0.30, 0.28, and 0.42 for incorrect responses. These values encode the reference-scenario hypotheses that mathematical execution can remain adequate despite an incorrect final answer and that an incorrect response can retain a superficially plausible explanation. The probabilities are modelling assumptions rather than empirically estimated parameters.

Because the rubric outcomes are generated computationally, no human raters were recruited, and no inter-rater reliability coefficient was calculated. Within every problem-by-profile observation, the three rubric dimensions are sampled conditionally on the same final-answer correctness outcome but are otherwise conditionally independent in the reference scenario. This simplifying assumption makes the generating mechanism adaptable, although it does not reproduce all dependencies that may occur in genuine written solutions. A future empirical implementation should estimate the relationships among rubric dimensions from independently coded responses and should report inter-rater reliability and adjudication procedures.

Every incorrect synthetic response receives one primary error category drawn from a multinomial distribution. The baseline probability vector, ordered as conceptual misapplication, model-construction failure, mathematical or computational error, and hallucinated reasoning, is (0.34, 0.36, 0.16, 0.14). Domain-specific modifiers are (-0.02, +0.01, +0.04, -0.03) for Classical Mechanics; (0, 0, 0, 0) for Electromagnetism, Thermodynamics and Statistical Physics, and Waves and Optics; and (0, -0.02, -0.04, +0.06) for Modern Physics. Profile-specific modifiers are (-0.01, 0, 0, +0.01) for Model A; (-0.04, -0.01, +0.03, +0.02) for Model B; (-0.01, 0, 0, +0.01) for Model C; and (+0.04, 0, -0.03, -0.01) for Model D. For each incorrect response, the relevant domain and profile modifiers are added to the baseline vector. The resulting elements remain non-negative and sum to one; nevertheless, the program verifies these conditions before multinomial sampling. The taxonomy and its probabilities are imposed by the simulation and are not empirically discovered characteristics of language-model errors.

The simulation is executed with the fixed random seed 20260809. At the beginning of each of the 1,000 replications, the complete factorial template of 150 hypothetical problems is generated, followed by one normally distributed problem-level random intercept for each problem. The four generic profiles are crossed with the 150 problems, producing exactly 600 problem-by-profile observations per replication. Final-answer correctness is sampled from the declared logistic model. The three ordinal rubric dimensions are then sampled from the conditional probability distributions specified above, and each incorrect response is assigned exactly one error category from the adjusted multinomial distribution. Correct responses receive no error-category label.

The reproducibility checks require every replication to contain 600 observations, every generated probability to fall between zero and one, the three probabilities for each rubric dimension to sum to one, and every incorrect response to receive exactly one error category. The program also verifies that all error-category probabilities sum to one and that pooled summaries agree with correctly weighted marginal summaries.

For every primary estimand, three quantities are reported: the mean across 1,000 replications, the 95% simulation interval, and the Monte Carlo standard error. The 95% simulation interval is defined by the 2.5th and 97.5th percentiles of the replication-level estimates and represents the variation expected among hypothetical implementations of the 150-problem benchmark under the fixed reference scenario. The Monte Carlo standard error is calculated as the standard deviation of the 1,000 replication-level estimates, divided by the square root of 1,000, and measures the numerical precision with which the long-run simulation mean is estimated.

The simulation interval and Monte Carlo standard error answer different questions. A wide simulation interval indicates that an individual hypothetical benchmark implementation may produce an estimate substantially different from the long-run scenario mean. By contrast, a small Monte Carlo standard error indicates that the mean of the 1,000 replications has been

estimated with adequate numerical stability. Neither quantity is a confidence interval for the performance of an actual language model or for an empirical population of physics problems. The complete simulation script, parameter file, data dictionary, software version record, and automated consistency checks should be provided as supplementary materials to ensure reproducibility.

Parameter selection followed a theory-informed scenario-design strategy rather than empirical estimation. Previous studies were used to determine the expected directions of the effects, particularly the decline in performance with increasing curricular difficulty and the disadvantage associated with problems requiring assumption generation and physical model construction (Wang et al., 2024; Yeadon & Hardy, 2024; Tschisgale et al., 2025). Their reported percentages were not combined or directly transferred into the present simulation because the studies used different model versions, problem banks, prompting conditions, scoring procedures, and definitions of difficulty. The numerical coefficients should therefore be interpreted as reference-scenario inputs selected to produce non-degenerate probabilities across the complete factorial design, rather than as estimates based on prior evidence.

The baseline logit of 1.00 corresponds to baseline odds of 2.72 before profile, domain, difficulty, specification, interaction, and problem-level effects are applied. The profile coefficients of 0.00, 0.35, 0.20, and -0.25 correspond to odds ratios of 1.00, 1.42, 1.22, and 0.78 relative to Model A. These values create distinguishable but overlapping hypothetical performance profiles without representing differences among named commercial systems. The domain coefficients correspond to odds ratios of 1.28 for Classical Mechanics, 1.00 for Electromagnetism, 1.05 for Thermodynamics and Statistical Physics, 0.90 for Waves and Optics, and 0.70 for modern physics. Their purpose is to demonstrate the recovery of domain contrasts rather than to claim that this ordering characterises actual language models.

The difficulty coefficients correspond to odds ratios of 1.82 for introductory problems, 1.00 for intermediate problems, and 0.45 for advanced problems. Under-specification has an odds ratio of 0.43, with additional interaction odds ratios of 0.86 at the intermediate level and 0.58 at the advanced level. These parameters operationalise the hypothesis that the demands of identifying missing information and constructing a defensible physical model become increasingly consequential as problem difficulty increases. The problem-level random-intercept standard deviation of 0.35 introduces moderate heterogeneity among hypothetical problems within the same factorial cell while preserving interpretable marginal contrasts.

The selected coefficients define only one reference scenario. Their justification rests on theoretical coherence, avoidance of floor and ceiling effects, and usefulness for testing the proposed analysis pipeline – not on statistical calibration to observed model responses. For this reason, the sensitivity analysis varies the baseline logit, the under-specification coefficient, the advanced-by-underspecified interaction, and the problem-level heterogeneity across weaker and stronger alternatives. The resulting analysis identifies which conclusions persist across the examined assumptions and which depend strongly on the reference values. The simulation supports only conditional conclusions of the form “under this parameterisation, the following pattern is produced.” It cannot estimate the performance of an actual system, determine whether commercial models differ significantly, or establish that the proposed error taxonomy exhausts the failure modes found in genuine responses.

No human participants, private data, commercial model outputs, or copyrighted item bank were used in the simulation. Ethical risk is consequently limited to the possibility of misleading presentation. To mitigate that risk, the title, abstract, methods, table notes, discussion, limitations, and conclusion explicitly identify all numerical results as synthetic. Reproducible code and the complete parameter specification should accompany any public version. A future empirical phase would require an item-permission review, a data management plan, model version documentation, and an institutional ethics review when human raters or students are involved.

The primary estimand is the replication-level mean proportion of correct responses over the balanced 600-observation design. Secondary estimands include marginal correctness by domain, difficulty, specification status, and profile; the difficulty-by-specification interaction; the proportion of fully adequate rubric ratings; and the conditional distribution of error categories among incorrect responses. Defining these estimands before generation prevents a common ambiguity in simulation reports, namely switching between pooled observation-

level summaries and averages of replication-level summaries. In this design, the two are nearly identical because every replication is balanced and complete, but replication-level summaries remain preferable because they make Monte Carlo variability visible and preserve the conceptual distinction between one hypothetical study and the long-run behaviour of repeated studies. A future empirical benchmark should state an analogous estimand hierarchy and should distinguish descriptive benchmark accuracy from population-level claims about an open-ended universe of physics problems.

The simulation algorithm proceeds in a fixed sequence. At the beginning of each replication, 150 problem records are created from the complete factorial template, and a problem-specific random intercept is sampled for each record. The four profiles are then crossed with the problem records. Correctness probabilities are calculated from the logistic equation and converted to Bernoulli outcomes. Conditional on the resulting correctness value, the three ordinal rubric dimensions are sampled from their specified distributions. Finally, each incorrect observation is assigned one primary error category using domain- and profile-adjusted multinomial probabilities. Cell counts and probability bounds are checked after every replication. This sequence ensures that rubric and taxonomy outputs remain logically linked to the same underlying response rather than being generated as unrelated tables.

Several computational diagnostics are required before substantive interpretation. First, every replication must contain exactly 600 observations, with equal counts in the intended design cells. Second, generated probabilities must fall within the closed unit interval after inverse-logit transformation and category renormalisation. Third, every incorrect response must receive exactly one error label, while correct responses receive none. Fourth, the four error proportions must sum to one within rounding tolerance. Fifth, rerunning the program with the declared seed must reproduce the complete set of aggregate results. Sixth, pooled means should agree with appropriately weighted marginal means. These checks establish implementation fidelity, although they cannot establish that the assumed model represents reality.

A one-factor-at-a-time sensitivity analysis was conducted to determine whether the principal patterns depended on the reference parameterisation. Four components of the correctness model were varied: the baseline logit, the under-specification coefficient, the advanced-by-underspecified interaction, and the standard deviation of the problem-level random intercept. The baseline logit was evaluated at 0.50, 1.00, and 1.50; the under-specification coefficient at -0.40, -0.85, and -1.30; the advanced-by-underspecified interaction at 0.00, -0.55, and -1.00; and the problem-level standard deviation at 0.15, 0.35, and 0.60. In every alternative scenario, all parameters not under examination were retained at their reference values. Each scenario was evaluated using 1,000 independent Monte Carlo replications, with the same balanced 150-problem factorial architecture and four hypothetical profiles. Sensitivity was assessed through overall correctness, the overall difference between well-specified and underspecified problems, correctness in the underspecified advanced condition, and the ordering of the four profiles. The purpose was to identify which qualitative conclusions remained stable across plausible assumptions rather than to demonstrate numerical invariance.

The balanced architecture is intentionally simple, but a future planning simulation could incorporate additional sources of variation. Repeated stochastic completions from the same model and prompt would create a response-within-problem level; model versions could be nested within developers; prompts could cross with profiles; and raters could contribute classification variability. A generalised linear mixed model could represent correctness, while cumulative-link models could represent ordinal rubric scores, and multinomial models could represent error categories. Correlations among the rubric dimensions could be introduced through latent variables or copulas. Such extensions would improve realism but also increase the number of weakly identified assumptions. The current design, therefore, prioritises interpretability and auditability over maximal complexity.

No missing data are generated in the reference scenario. This choice assumes that every hypothetical profile returns a scoreable response to every problem. Empirical systems may refuse, time out, produce malformed notation, rely on unavailable images, or return answers that cannot be unambiguously coded. Those outcomes should not be discarded automatically,

because exclusion could inflate the apparent correctness. A preregistered empirical protocol should define completion, refusal, non-response, and unscorable output; report their frequencies separately; and provide sensitivity analyses under alternative scoring rules. The simulation code can later incorporate a response-availability model once empirical pilot data provide defensible probabilities.

The choice of 1,000 replications reflects a compromise between numerical stability and computational simplicity. Monte Carlo error can be assessed by repeating the complete run under several independent seeds or by calculating the standard error of each replication-level summary. If the third decimal place changes materially across runs, additional replications are warranted. For the descriptive percentages reported here, 1,000 replications provide sufficiently stable one-decimal summaries, but stability is a property of the implemented simulation rather than a general rule. Rare outcomes, tail probabilities, or complex multilevel estimators may require tens of thousands of replications. The number of replications should always be justified with respect to the estimand and the desired numerical precision.

The synthetic design also separates Monte Carlo variability from inferential uncertainty. Monte Carlo variability arises because the program repeatedly samples from fully specified distributions; it can be reduced by increasing the number of replications. Inferential uncertainty in an empirical benchmark would arise because selected problems, generated responses, raters, and model versions represent broader targets of inference; it cannot be eliminated by rerunning a simulation. Conflating these concepts would overstate precision. The 95% simulation intervals in Section 5 describe the distribution of hypothetical study summaries under the reference scenario. They do not quantify uncertainty in the chosen coefficients, nor do they provide coverage guarantees for future observed values.

Reproducibility requires more than reporting a random seed. The executable environment, software versions, probability tables, category-normalisation rules, item-construction loop, and aggregation code all influence results. A reproducibility package should therefore include a machine-readable configuration file, source code, a data dictionary, unit tests for the factorial counts, a small expected-output fixture, and a session manifest. The manuscript should identify which table cells are computed and which values are assumptions. This distinction allows an independent researcher to audit the logic, modify one component, and determine whether discrepancies arise from implementation or from an intentional alternative scenario.

The reference scenario uses a complete balance to make the interpretation transparent, but the empirical allocation may need to reflect feasibility and precision. Advanced problems often demand longer responses and more rater time, while underspecified tasks may lead to greater disagreement over acceptable assumptions. A planning simulation can attach costs to cells and compare equal allocation with optimal or constrained designs. Oversampling conditions expected to have high variance may improve precision for key interactions, whereas retaining common anchor items across profiles can strengthen comparisons. Any departure from balance should be determined before observing outcomes and incorporated into weights and variance estimates.

Power analysis in this context should focus on practically meaningful contrasts rather than on the probability of obtaining a small p-value from hundreds of observations. Investigators could define a minimum important difference in correctness, a minimum detectable difference-by-specification interaction, or a threshold change in a high-risk error category. Simulation would then vary the number of problems, the number of repeated generations, and the number of raters to estimate bias, interval coverage, and decision accuracy for the proposed multilevel estimator. This approach would connect the hypothetical model to concrete design choices while acknowledging that power results remain conditional on assumed effect sizes and variance components.

Analysis of an observed benchmark should preserve the dependence created by the design. Four profiles responding to the same problem are not independent because they share wording, content, and scoring key, while repeated generations from one profile share a prompt and system configuration. Treating all responses as unrelated would underestimate uncertainty and could make trivial contrasts appear decisive. Problem and generation effects should be represented directly through mixed models, clustered resampling, or another justified approach. Descriptive tables remain important, but model-based estimates should

report absolute differences, uncertainty intervals, and sensitivity to alternative link functions and random-effect structures.

Preregistration would strengthen the transition from this planning model to an observed benchmark. The registration should distinguish confirmatory contrasts from exploratory analyses, identify the primary scoring rule and tolerance for numerical equivalence, define how unit errors and alternative valid assumptions will be handled, and state whether model names will be masked during coding. It should also specify the minimum number of repeated generations, the treatment of refusals, the adjudication process, and the criteria for excluding a flawed problem. Publishing the simulation and expected table shells with the registration would make departures visible without preventing justified adaptation when unexpected response forms occur.

Quality assurance for the empirical item bank should include independent solution, adversarial review, and pilot administration. At least two subject-matter experts should solve every item, compare assumptions, and identify alternative defensible answers. Reviewers should assess whether an underspecified problem admits a manageable set of assumptions or is simply indeterminate. Pilot outputs from systems not included in the main comparison can reveal formatting and scoring difficulties without contaminating the confirmatory analysis. Any item revised after the pilot inspection should receive a new version identifier, and the final bank should be frozen before the target systems are queried.

5. Results

5.1. Mean Simulated Correctness by Profile and Domain

Across 1,000 replications, mean simulated correctness was 59.3% over the complete design, with a 95% replication-level simulation interval of 55.7%-63.3%. The four hypothetical profiles differed as programmed: Model B had the highest overall mean (64.7%), followed by Model C (61.9%), Model A (58.0%), and Model D (52.7%). Every profile showed its highest mean in Classical Mechanics and its lowest mean in Modern Physics. Table 1 presents the profile-by-domain matrix. These contrasts reflect specified logit shifts and sampling variability, not evidence of differences among real products.

Table 1. Mean simulated correctness by hypothetical profile and physics domain (%).

Physics domain	Model A	Model B	Model C	Model D
Classical Mechanics	63.8	70.1	67.3	58.2
Electromagnetism	58.7	65.6	62.7	53.7
Thermodynamics and Statistical Physics	59.8	66.0	63.3	54.4
Waves and Optics	56.9	63.3	60.7	51.4
Overall	58.0	64.7	61.9	52.7

Note: Entries are means across 1,000 Monte Carlo replications. Each domain contributes 30 hypothetical problems per profile across all replications. Models A-D are generic scenarios and do not correspond to commercial systems.

Two design-induced patterns are visible. First, the programmed domain gradient produces higher simulated correctness in Classical Mechanics and lower correctness in Modern Physics for all profiles. Second, the between-profile ordering is stable because the same profile shifts operate across all domains, while random problem effects preserve replication-to-replication variation. The table, therefore, demonstrates whether the analysis pipeline recovers the intended structure. It does not establish that mechanics is intrinsically easier for large language models or that any vendor occupies the simulated performance band; those propositions require observed benchmarking data.

5.2. Simulated Performance by Difficulty and Specification Status

The strongest simulated gradient concerns difficulty and specification status. Mean correctness declines from introductory to advanced tasks, and within every difficulty level, it is lower for underspecified than for well-specified tasks. Table 2 pools the four generic profiles and reports the resulting cell means. Because the factorial design is balanced, the

overall percentages are simple averages of equally weighted conditions.

Table 2. Mean simulated correctness by difficulty and problem specification (%).

Difficulty level	Well specified	Under specified	Overall
Introductory	83.0	68.0	75.5
Intermediate	73.0	51.2	62.1
Advanced	55.7	25.1	40.4
Overall	70.6	48.1	59.3

Note: Entries are replication means pooled across four profiles. Each difficulty-by-specification cell contains 100 problem-by-profile observations per replication. The values are synthetic outputs, not observed performance estimates.

The lowest mean is produced for underspecified advanced tasks (25.1%; 95% simulation interval, 17.0%-34.0%), compared with 55.7% for well-specified advanced tasks and 83.0% for well-specified introductory tasks. The specification gap grows across the programmed difficulty levels, reflecting both the -0.85 under-specification coefficient and the additional interaction penalties. This result is best interpreted as a design check: the simulation creates a setting in which model construction and assumption generation become progressively more consequential. Whether real model outputs exhibit the same interaction is an empirical hypothesis.

Table 3 disaggregates the specification-status contrast by hypothetical profile. The ordering follows the profile effects, but each scenario retains a substantial penalty for underspecified tasks.

Table 3. Mean simulated correctness by hypothetical profile and specification status (%).

Profile	Well specified	Under specified	Overall
Model A	69.4	46.7	58.0
Model B	75.5	53.9	64.7
Model C	73.1	50.6	61.9
Model D	64.2	41.3	52.7

Note: Entries are means across domains and difficulty levels, based on 1,000 replications. Each specification condition contains 75 hypothetical problems per profile per replication.

5.3. Simulated Multidimensional Rubric Outcomes

Final-answer correctness provides only a binary summary and cannot distinguish inadequate reasoning from partially or fully adequate conceptual, mathematical, and explanatory components. Table 4, therefore, reports the complete three-category distribution for each ordinal rubric dimension rather than only the percentage of fully adequate ratings. Final-answer correctness is reported separately in tables 1-3 because it is generated as a binary outcome.

Across the entire simulated design, mathematical accuracy has the highest proportion of fully adequate ratings at 71.5%, followed by explanation quality at 58.6% and conceptual accuracy at 56.5%. The corresponding inadequate proportions are 8.5%, 15.2%, and 24.2%, respectively. This pattern follows directly from the conditional probability distributions specified in Section 4, particularly the assumption that mathematically adequate work may occur even when the final answer is incorrect.

The full ordinal distributions reveal differences that would be concealed by reporting only fully adequate ratings. For example, Model D has the lowest simulated conceptual adequacy, with 27.6% of responses classified as conceptually inadequate, compared with 21.4% for Model B. Mathematical inadequacy remains below 10% for all four profiles because the reference scenario permits mathematically adequate procedures to coexist with an incorrect final answer. Explanation quality occupies an intermediate position, with approximately one quarter of responses in every profile classified as partially adequate.

These differences should be interpreted as properties of the declared generating model. Because the rubric dimensions are conditionally generated rather than assigned to observed written solutions, the distributions do not demonstrate that actual language models possess

stronger mathematical than conceptual competence. Their methodological value is to show how a future empirical benchmark can preserve partial performance information, examine the joint relationship between final-answer correctness and reasoning quality, and avoid reducing complex written solutions to a single binary score

Table 4. Simulated ordinal rubric distributions by hypothetical profile (%).

Profile	Rubric dimension	Inadequate	Partially adequate	Fully adequate
Model A	Conceptual accuracy	24.8	19.6	55.6
	Mathematical accuracy	8.7	20.2	71.0
	Explanation quality	15.5	26.3	58.2
Model B	Conceptual accuracy	21.4	18.4	60.3
	Mathematical accuracy	7.6	18.8	73.6
	Explanation quality	13.8	26.1	60.1
Model C	Conceptual accuracy	22.8	18.9	58.3
	Mathematical accuracy	8.1	19.4	72.5
	Explanation quality	14.5	26.1	59.3
Model D	Conceptual accuracy	27.6	20.5	51.9
	Mathematical accuracy	9.6	21.4	69.0
	Explanation quality	16.8	26.4	56.8
Pooled distribution	Conceptual accuracy	24.2	19.3	56.5
	Mathematical accuracy	8.5	20.0	71.5
	Explanation quality	15.2	26.2	58.6

Note: Entries are means across 1,000 Monte Carlo replications. Percentages sum to approximately 100% within each profile-by-dimension row; minor deviations reflect rounding. The ordinal outcomes are generated conditionally on final-answer correctness using the probabilities specified in Section 4. They were not assigned by human raters.

5.4. Simulated Error Taxonomy

Among synthetic incorrect responses, model-construction failure is the most frequently generated category in the pooled results (35.4%), followed by conceptual misapplication (33.4%), hallucinated reasoning (15.7%), and mathematical or computational error (15.6%). These proportions are conditional on the declared baseline probabilities and domain- and profile-specific modifiers. They represent generated scenario outcomes rather than empirically observed failure frequencies. Table 5 reports domain-specific proportions. Because categories were generated from specified multinomial probabilities, their ordering is an input-informed output rather than an inductively discovered empirical taxonomy.

Table 5. Simulated error-category distribution by physics domain (% of incorrect responses).

Physics domain	Conceptual misapplication	Model-construction failure	Mathematical or computational error	Hallucinated reasoning
Classical Mechanics	32.0	36.6	19.6	11.8
Electromagnetism	33.9	35.2	15.9	15.1
Thermodynamics and Statistical Physics	34.0	35.7	15.6	14.8
Waves and Optics	33.6	35.6	16.1	14.7
Modern Physics	33.4	34.1	11.8	20.6
Pooled distribution	33.4	35.4	15.6	15.7

Note: Each synthetic incorrect response receives exactly one primary error category. Entries are replication-level means across 1,000 Monte Carlo replications and sum to approximately 100% within each row. The pooled distribution is weighted by the number of incorrect responses in each domain; minor deviations from 100% reflect rounding.

The domain table illustrates how a future benchmark might reveal substantively different failure profiles even when overall correctness is similar. The modern-physics modifier produces a higher hallucinated-reasoning share (19.0%), whereas Classical Mechanics has a larger mathematical-error share (18.7%). These contrasts were intentionally encoded to test

whether the reporting framework preserves domain-specific category structure. They should be treated as hypotheses for empirical classification, particularly because real responses may contain multiple interacting errors that resist assignment to one primary category.

Table 6 disaggregates generated error categories by profile. Differences are deliberately modest so that the simulation does not presume radically distinct cognitive mechanisms for the four scenarios.

Table 6. Simulated error-category distribution by hypothetical profile (% of incorrect responses).

Profile	Conceptual misapplication	Model-construction failure	Mathematical or computational error	Hallucinated reasoning
Model A	32.8	35.5	15.7	16.0
Model B	29.6	34.6	18.9	16.9
Model C	32.5	35.8	15.6	16.2
Model D	37.5	35.5	12.9	14.0
Pooled distribution	33.4	35.4	15.6	15.7

Note: Entries are replication-level means across 1,000 Monte Carlo replications and sum to approximately 100% within each row. The pooled distribution is calculated from all incorrect problem-by-profile observations and is identical to that in Table 5. Models A-D are hypothetical and do not correspond to commercial systems; minor deviations from 100% reflect rounding.

Model D has the largest share of conceptual misapplication (37.5%), whereas Model B has the largest share of mathematical or computational errors (18.9%). hallucinated reasoning ranges from 14.0% for Model D to 16.9% for Model B. These differences follow from the declared profile and domain modifiers and should not be interpreted as evidence about actual model architectures. These differences are consequences of the declared category modifiers. Their value is analytical: they demonstrate how the taxonomy can support comparative reporting once genuine response data exist. They cannot support claims about the characteristic errors of specific architectures or developers.

5.5. Summary of Principal Simulation Patterns

The Monte Carlo outputs form an internally coherent illustration of the proposed benchmark. Four principal patterns follow from the generating model, and each must be interpreted conditionally.

First, the simulation produces substantial but incomplete success on well-specified introductory and intermediate problems. This confirms that the code implements the intended baseline and difficulty coefficients. It does not confirm that contemporary language models attain those rates.

Second, simulated correctness declines with increasing difficulty and under specification, with the largest penalty occurring when the two conditions coincide. This pattern operationalises the theoretical hypothesis that constructing physical models compounds content complexity. Its magnitude remains to be estimated empirically.

Third, the generated error distribution is dominated by model construction and conceptual categories rather than mathematical error. This illustrates a pedagogically meaningful taxonomy and a corresponding analysis plan, but the proportions are parameter-dependent and cannot be generalised beyond the simulation.

Fourth, the modern-physics scenario carries the highest proportion of hallucinated reasoning. This scenario identifies a risk worth testing because novice learners may have difficulty evaluating fabricated explanations in unfamiliar domains. The simulation alone does not demonstrate that the risk occurs at the reported rate.

Section 6 interprets these patterns as conditional design implications, relates them to prior empirical literature, and identifies the evidence needed before any real-world conclusion can be drawn.

5.6. Sensitivity Analysis

The sensitivity analysis showed that absolute simulated correctness depended substantially on the assumed baseline capability and under-specification penalty, whereas the

qualitative ordering of the four hypothetical profiles remained unchanged. Table 7 summarises the nine scenarios. The overall correctness rate ranged from 49.1% in the low-baseline scenario to 68.7% in the high-baseline scenario. Consequently, the reference value of 59.3% should not be interpreted as a generally expected level of performance; it is specific to the selected baseline coefficient.

Table 7. Sensitivity of the principal simulation outcomes to alternative parameter assumptions.

Scenario	Varied parameter	Overall correctness (%)	Overall specification gap (percentage points)	Underspecified advanced correctness (%)	Profile ordering
Reference	Baseline parameterisation	59.3	22.4	25.2	B > C > A > D
Low baseline	Baseline logit = 0.50	49.1	23.0	17.0	B > C > A > D
High baseline	Baseline logit = 1.50	68.7	20.7	34.9	B > C > A > D
Weak under-specification penalty	Coefficient = -0.40	64.0	13.2	33.9	B > C > A > D
Strong under-specification penalty	Coefficient = -1.30	54.7	31.7	17.7	B > C > A > D
No advanced interaction	Interaction = 0.00	61.1	19.0	35.8	B > C > A > D
Strong advanced interaction	Interaction = -1.00	58.2	25.0	17.8	B > C > A > D
Low item heterogeneity	Problem-level SD = 0.15	59.4	23.0	24.3	B > C > A > D
High item heterogeneity	Problem-level SD = 0.60	59.1	21.7	26.0	B > C > A > D

Note: Specification gap is the mean correctness rate for well-specified problems minus that for underspecified problems. Each scenario contains 1,000 Monte Carlo replications. Only the parameter identified in the second column differs from the reference scenario.

Changes in the under-specification coefficient had the clearest effect on the specification gap. Weakening the coefficient from -0.85 to -0.40 reduced the gap from 22.4 to 13.2 percentage points, whereas strengthening it to -1.30 increased the gap to 31.7 percentage points. Consequently, the performance under the advanced, underspecified condition ranged from 17.7% to 33.9% across these scenarios. The magnitude of the difficulty-by-specification pattern is therefore strongly dependent on assumptions about the additional demands associated with incomplete problem information.

The advanced-by-underspecified interaction also materially affected the most demanding condition. Removing the interaction increased correctness for underspecified advanced problems from 25.1% to 35.8%, whereas strengthening it to -1.00 reduced correctness to 17.8%. The interaction observed in the reference scenario should therefore be interpreted as a hypothesis to be estimated empirically rather than as a robust quantitative prediction.

By contrast, changing the problem-level standard deviation from 0.15 to 0.60 had little effect on the mean outcomes. Overall correctness remained between 59.1% and 59.4%, while correctness for underspecified advanced problems ranged from 24.3% to 26.0%. Item heterogeneity would be expected to have a greater influence on interval width and the precision of individual contrasts than on balanced marginal means.

The ordering Model B > Model C > Model A > Model D was preserved in all nine scenarios because the profile coefficients were held constant. This stability should not be interpreted as independent evidence supporting that ordering; it demonstrates only that variations in the four examined parameters did not reverse the programmed profile effects. A separate sensitivity analysis that varies profile coefficients would be required to evaluate ranking uncertainty.

Overall, the sensitivity analysis indicates that the directional disadvantage associated with

underspecified and advanced problems persists across the examined scenarios, but its magnitude is highly parameter-dependent. The principal defensible conclusion is therefore qualitative: the proposed benchmark can detect difficulty and specification contrasts when such effects are present. The specific percentages reported for the reference scenario should not be treated as predictions of empirical model performance.

5.7. Simulation Intervals and Monte Carlo Precision

Table 8 reports uncertainty measures for the primary correctness estimands and contrasts. The overall correctness mean is 59.3%, with a 95% simulation interval from 55.3% to 63.3%. Its Monte Carlo standard error is 0.06 percentage points, indicating that 1,000 replications provide a stable estimation of the long-run mean under the reference scenario.

Table 8. Replication-level simulation intervals and Monte Carlo standard errors for the primary estimands.

Estimand or contrast	Mean estimate	95% simulation interval	Monte Carlo standard error
Overall correctness (%)	59.3	55.3 to 63.3	0.06
Model B – Model A (percentage points)	6.8	-2.7 to 16.0	0.16
Model C – Model A (percentage points)	3.9	-6.0 to 13.3	0.16
Model A – Model D (percentage points)	5.2	-4.7 to 14.7	0.16
Introductory – Advanced (percentage points)	34.9	26.0 to 43.0	0.14
Well specified – Under specified (percentage points)	22.4	15.0 to 30.0	0.12
Difference-in-differences interaction (percentage points)	16.0	-1.0 to 35.0	0.29

Note: The difference-in-differences interaction is calculated as the specification gap for advanced problems minus the specification gap for introductory problems. Simulation intervals are the 2.5th and 97.5th percentiles of the 1,000 replication-level estimates. Monte Carlo standard errors quantify numerical precision, not empirical standard errors. Contrasts were calculated from unrounded replication-level estimates and may differ slightly from differences calculated using the rounded marginal percentages reported in tables 1-3.

The large difficulty and specification contrasts remain positive throughout their simulation intervals. The introductory-versus-advanced contrast has a mean of 34.9 percentage points and an interval from 26.0 to 43.0 percentage points, while the well-specified-versus-underspecified contrast has a mean of 22.4 percentage points and an interval from 15.0 to 30.0 percentage points. Under the reference scenario, an individual implementation of the proposed benchmark would therefore be expected to recover the directional difficulty and specification effects.

A different pattern emerges for the profile contrasts. Although the long-run means preserve the programmed ordering, the simulation intervals for Model B versus Model A, Model C versus Model A, and Model A versus Model D include zero. Thus, a single 150-problem implementation may not reliably distinguish profile differences of the programmed magnitude. This result identifies an important design limitation that cannot be observed from the mean percentages alone. A future empirical study seeking precise comparisons among systems may require a larger problem bank, repeated generations, or both.

The difference-in-differences interaction also has a wide simulation interval, ranging from -1.0 to 35.0 percentage points, despite a mean of 16.0 percentage points. Consequently, the present 150-problem design may be adequate for detecting the overall disadvantages associated with difficulty and under-specification, but insufficient to precisely estimate their interaction in a single implementation. This finding provides a direct basis for the power and precision analyses recommended for the future empirical benchmark.

The Monte Carlo standard errors range from 0.06 to 0.29 percentage points and are small relative to the corresponding simulation intervals. Increasing the number of replications would therefore have little effect on the variability expected across individual benchmark implementations. Reducing that variability requires changes to the empirical

design, such as increasing the number of problems or the number of repeated generations, rather than merely increasing the number of Monte Carlo replications.

6. Discussion

The simulation reported in Section 5 offers a coherent demonstration of how a prospective physics benchmark can connect theoretical constructs to an analysis plan. Its value lies in exposing assumptions, checking whether a balanced design can recover intended contrasts, and clarifying which empirical claims would become possible after genuine data collection. This discussion therefore distinguishes three levels of inference: patterns mechanically produced by the generating model, hypotheses motivated by prior empirical literature, and pedagogical implications that remain provisional until tested with real model responses and student data.

The most theoretically informative simulated contrast is the widening difference between well-specified and underspecified tasks as difficulty increases. Within the model, this pattern is guaranteed by the main penalty for under-specification and the added interaction penalties at intermediate and advanced levels. The pattern operationalises the expert-novice distinction between procedural execution and principled model construction (Chi et al., 1981; Larkin et al., 1980). A well-specified task can often be represented as the selection and execution of a familiar procedure, whereas an underspecified task requires decisions about relevance, idealisation, missing information, and defensible assumptions. The simulation does not show that language models use pattern matching or lack conceptual schemas; it shows how a benchmark could be made sensitive to that theoretical distinction.

The generic profile differences serve a similarly limited purpose. Models A-D span a plausible range of baseline log-odds, and the recovery of their programmed ordering verifies the comparative summaries. It would be inappropriate to map those labels onto current commercial systems or to infer architecture-level mechanisms. Real systems are versioned, exposed through changing interfaces, and responsive to prompting conditions. An empirical study should preregister the exact versions and access dates, randomise the presentation order, preserve complete outputs, and repeat prompts when stochastic generation is enabled. Only then could between-model contrasts be interpreted as measurements rather than scenarios.

The multidimensional rubric is one of the framework's main substantive contributions. Binary correctness can obscure a response that reaches an acceptable answer through incoherent reasoning, as well as a response that demonstrates sound conceptual framing before a computational slip. By retaining separate indicators of conceptual accuracy, mathematical accuracy, explanation quality, and final correctness, the proposed benchmark can distinguish these cases. The simulation assigns mathematical adequacy a higher mean than conceptual adequacy, but that ordering is a hypothesis rather than a discovery. Empirical raters may find different dependencies, and qualitative analysis may reveal that the three-point categories need further refinement.

The four-category taxonomy likewise functions as an a priori coding framework. Conceptual misapplication captures the use of a genuine principle in an inappropriate context; model-construction failure concerns an unsuitable idealisation or omitted constraint; mathematical or computational error concerns execution after a defensible setup; and hallucinated reasoning concerns fabricated claims without a basis in accepted physics or the task. These categories are useful because they imply different instructional responses. Nevertheless, actual outputs may combine multiple categories into a single solution. A future coding manual should permit secondary labels or segment-level annotation and should report category reliability rather than assuming that a single primary code is always sufficient.

The simulated concentration of model-construction and conceptual errors motivates a pedagogical hypothesis: if observed outputs exhibit a comparable distribution, instruction that focuses solely on checking arithmetic would address only a minority of consequential failures. Students would instead need practice evaluating assumptions, identifying governing principles, comparing alternative representations, and testing limiting cases. Such activities align with established goals in physics education independently of generative AI (Hsu et al., 2004). The present study supports the design of those future investigations, not the claim that a particular intervention is already effective.

A second pedagogical hypothesis concerns cognitive offloading. Evidence from prior studies suggests that students may accept AI-generated physics solutions without adequate reflection (Krupp et al., 2024), but the simulation contains no learners and cannot measure that behaviour. It can, however, specify materials for a controlled study: matched correct and incorrect solutions sampled across the four error categories, tasks that ask students to detect and explain errors, and outcomes that measure conceptual learning and calibration of confidence. The generated category distribution could support power and sampling calculations, while final study materials should use observed or carefully authored responses.

Assessment design is another prospective application. A benchmark that systematically crosses difficulty with specification status can help distinguish routine equation use from model construction. Physics instructors might adapt the architecture by asking students to state assumptions, justify idealisations, evaluate incomplete information, or critique alternative solution paths. Yet the simulated performance gradients do not establish that such tasks are resistant to current models, nor do they justify high-stakes claims about academic integrity. Local pilots with current systems and actual students are needed before assessment policies are changed.

The framework also clarifies what reproducibility requires. The final empirical article should provide the item-sampling procedure, permissible-use status of each problem, exact prompts, model parameters, dates and interfaces, raw outputs where licensing permits, a detailed coding manual, adjudication records, and analysis code. Repeated administration would help separate prompt-level stochasticity from item and profile effects. A multilevel model could then estimate uncertainty across sampled problems and repeated outputs rather than treating all observations as independent. These requirements are more demanding than the present simulation, but specifying them now reduces analytic flexibility later.

Accordingly, the simulated percentages should be viewed as a worked example and a set of falsifiable expectations. The framework predicts lower performance with greater difficulty, an added disadvantage for underspecified tasks, and domain-varying error distributions. Empirical data may reproduce, attenuate, reverse, or complicate those patterns. Any of those outcomes would be informative because the simulation makes the prior assumptions visible. Its central achievement is not a claim about how well current systems solve physics problems, but a transparent plan for obtaining an answer.

The relationship between simulation and prior evidence requires particular care. Earlier studies supply reasons to consider specification status, domain breadth, and error type, but they do not uniquely determine the coefficients used here. Different studies have used different model generations, prompts, item sources, scoring rules, and access conditions. Their reported percentages are therefore not pooled into a pseudo-meta-analytic estimate. Instead, the literature informs the direction of selected scenario effects and the substantive interpretation of design factors. This separation protects the simulation from appearing more empirically calibrated than it is and makes disagreement possible: investigators can substitute alternative priors or parameter grids and examine how the conclusions change.

Domain differences deserve similar restraint. The reference scenario assigns the most favourable shift to Classical Mechanics and the least favourable shift to modern physics. One plausible motivation is the differential representation of canonical problems and differences in conceptual abstraction, but the simulation cannot evaluate these mechanisms. Actual domain effects could be driven by item wording, mathematical demands, diagram dependence, the prevalence of standard templates, or differences in answer-key quality. An empirical benchmark should therefore sample multiple problem forms within every domain and should avoid treating a domain label as a causal explanation. Item-level qualitative analysis would be necessary to understand why a domain contrast appears.

Difficulty is also a constructed property rather than an inherent number attached to a problem. The introductory, intermediate, and advanced labels in this design refer to curricular placement and expected reasoning demands. They are not derived from human response models, item-response theory, or observed completion times. Expert reviewers may disagree on classification, and a task that is difficult for a student may be familiar to a language model due to training data exposure. A future study could compare expert classifications, student facility values, and model performance, thereby testing whether the three populations share a common difficulty ordering. The simulation makes that comparison imaginable but does not

prejudge its result.

The error taxonomy can support explanatory work only if coding moves beyond surface outcome labels. For example, an incorrect numerical answer should not be classified as a computational error when the calculation faithfully follows an invalid physical premise. Conversely, a conceptually sound solution with a transcription slip should not be treated as a conceptual failure. Coding guidelines should identify the earliest consequential divergence from a defensible solution, document secondary errors, and use worked examples from every domain. Adjudication discussions should be retained as data for refining the taxonomy. These procedures would link the quantitative category distributions to the underlying reasoning structures evident in the response corpus.

Pedagogical use of the framework should be separated from punitive surveillance. The benchmark can help instructors design activities in which students compare solutions, state assumptions, locate the first invalid step, or repair an explanation. It should not be used on its own to infer that a student used AI or to declare an assessment secure against model assistance. The simulated gradients are not detection statistics, and similarity between a student error and a generated category would not identify authorship. The defensible educational application is formative: using structured critique to strengthen disciplinary judgement and metacognitive monitoring.

Finally, the framework invites cumulative rather than one-off benchmarking. If the same validated item bank and coding protocol are used across model releases, researchers can distinguish changes in overall accuracy from changes in the composition of errors. A profile might improve in final correctness while retaining a high rate of fluent conceptual failure among its remaining errors; another might reduce hallucinations but become more variable across repeated completions. Reporting only a single aggregate score would conceal these trajectories. A longitudinal design with versioned materials, stable anchor items, refreshed contamination-resistant items, and open analysis code would provide a more meaningful account of changes in capability.

A further interpretive advantage of the factorial design is that it discourages reliance on a single league-table score. Overall accuracy depends on the weighting of domains, difficulty levels, and specification conditions. A benchmark dominated by Introductory Mechanics could produce a much higher aggregate than one that gives equal weight to advanced modern physics, even if the underlying cell-specific performance were unchanged. The present design uses equal weights because its purpose is to compare across conditions, not to reproduce any particular curriculum. Empirical reports should publish the complete cell matrix and, where an overall score is needed, explain and justify its weighting scheme.

The simulation also highlights the distinction between performance and competence. A generated success probability represents an observable outcome under a task condition; it does not establish a stable mental capacity or a human-like conceptual structure. The same caution applies to empirical model outputs. Consistent correct answers across varied, contamination-resistant items would strengthen a claim about generalisable performance, but explanatory claims about internal representation would require additional evidence, such as controlled perturbations, mechanistic analysis, or systematic transfer tests. Physics education terminology should not be imported uncritically into descriptions of statistical systems.

For instructors, the most useful comparison may be between the benchmark profile and the learning objective of a particular activity. If an assignment is intended to develop algebraic fluency, strong model performance may motivate supervised practice or oral follow-up. If the objective is model construction, the relevant evidence is performance on underspecified and representation-rich tasks, not aggregate correctness. If the objective is to evaluate evidence, carefully selected erroneous solutions may serve as instructional resources. The framework thus supports alignment among objectives, task design, and evaluation while leaving local pedagogical decisions to empirical classroom study.

The proposed empirical benchmark should also distinguish robustness from peak performance. A profile may solve a problem after extensive scaffolding, tool use, or repeated attempts, yet remain unreliable with a simple standardised prompt. Both results can be valuable if they answer different questions. A baseline condition approximates unassisted consumer use; an optimised condition estimates performance under a documented support protocol; repeated trials estimate consistency. Reporting these dimensions separately would

prevent a best-case demonstration from being interpreted as typical reliability and would show whether improved outcomes require expertise that novice students are unlikely to possess.

Equity and accessibility constitute another layer of research that the numerical scenario does not capture. Students differ in language background, disability, access to paid systems, familiarity with prompting, and opportunity to verify outputs with instructors or peers. A technically accurate model could still widen educational disparities if effective use depends on costly tools or tacit expertise. Conversely, well-designed support could improve access to explanations and feedback. These possibilities require participatory and classroom-based research with diverse learners. The benchmark can characterise system outputs, but it should be embedded within a broader programme that examines who benefits, under what conditions, and with what unintended consequences.

Open reporting is particularly important because benchmark results can quickly become detached from their conditions of production. A headline percentage may circulate without its model version, prompt, item composition, scoring tolerance, or uncertainty. The full matrix should therefore accompany any summary, and machine-readable metadata should travel with released tables. Versioned identifiers, persistent repositories, and explicit correction procedures can help later readers determine exactly what was measured. This publication discipline is not ancillary: it is necessary for comparisons across studies and for preventing the conflation of simulated planning values, pilot estimates, and final empirical results.

7. Conclusions

This study presents a fully disclosed Monte Carlo simulation to design a cross-domain benchmark of large language models for undergraduate physics education. Rather than reporting unobserved model performance as empirical evidence, the study evaluates the structure, precision, and interpretability of a prospective benchmark comprising five physics domains, three difficulty levels, two specification conditions, and four generic performance profiles. Across 1,000 replications of 150 hypothetical problems, the framework demonstrates how final-answer correctness can be analysed alongside conceptual accuracy, mathematical accuracy, explanation quality, and a provisional four-category error taxonomy.

Under the declared generating assumptions, the simulation reproduces the intended performance gradients, including lower correctness for advanced and underspecified problems, and especially for their combination. These numerical outcomes are synthetic and must not be interpreted as estimates or rankings of contemporary commercial systems. The main contribution is therefore methodological: the study provides a transparent factorial architecture, explicit assumptions, reproducible simulation logic, and an analysis framework that can be replaced with observed responses in a future benchmark. For physics education, the framework highlights the value of evaluating not only whether an answer is correct but also whether the underlying conceptual reasoning, model construction, mathematics, and explanation are defensible.

The study is limited by its synthetic design. It contains no authentic model responses, human ratings, student data, or actual problem texts, and the proposed rubric and error taxonomy have not yet been empirically validated. It also does not directly represent prompt sensitivity, repeated-generation variability, training-data contamination, tool-enabled systems, or multimodal reasoning. Future research should therefore implement the design through a preregistered empirical benchmark using validated or newly authored physics problems, documented model versions and prompting conditions, repeated outputs, and independent physics-trained raters. Such work should test rubric reliability, examine robustness across domains and model configurations, and investigate whether students can detect and learn from different categories of model error. In this form, the simulation serves as a reproducible planning framework and a transparent bridge from theoretical expectations to future empirical testing.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Simulation Status Statement: All numerical results in this manuscript are synthetic outputs from a declared Monte Carlo model. No commercial large language model was queried, and no empirical model-response or student data were collected. The study demonstrates a prospective benchmark design and does not estimate the performance of any named system

References

- Aldazharova, S., Issayeva, G., Maxutov, S., & Balta, N. (2024). Assessing AI's problem solving in physics: Analyzing reasoning, false positives and negatives through the force concept inventory. *Contemporary Educational Technology*, 16(4), ep538. <https://doi.org/10.30935/cedtech/15592>
- Ansari, A. N., Ahmad, S., & Bhutta, S. M. (2024). Mapping the global evidence around the use of ChatGPT in higher education: A systematic scoping review. *Education and Information Technologies*, 29, 11281-11321. <https://doi.org/10.1007/s10639-023-12223-4>
- Arvidsson, K. (2025). AI-generated language in introductory physics lab reports in the first year of ChatGPT. *Discover Artificial Intelligence*, 5, 408. <https://doi.org/10.1007/s44163-025-00742-7>
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121-152. https://doi.org/10.1207/s15516709cog0502_2
- Dhillon, A.S. (1998). Individual differences within problem-solving strategies used in physics. *Science Education*, 82, 379-405. [https://doi.org/10.1002/\(SICI\)1098-237X\(199806\)82:3<379::AID-SCE5>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1098-237X(199806)82:3<379::AID-SCE5>3.0.CO;2-9)
- diSessa, A. A. (1993). Toward an Epistemology of Physics. *Cognition and Instruction*, 10(2-3), 105-225. <https://doi.org/10.1080/07370008.1985.9649008>
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30(3), 141-158. <https://doi.org/10.1119/1.2343497>
- Hsu, L., Brewe, E., Foster, T. M., & Harper, K. A. (2004). Resource letter RPS-1: Research in problem solving. *American Journal of Physics*, 72(9), 1147-1156. <https://doi.org/10.1119/1.1763175>
- Kortemeyer, G. (2023). Could an artificial intelligence agent pass an introductory physics course? *Physical Review Physics Education Research*, 19, 010132. <https://doi.org/10.1103/PhysRevPhysEducRes.19.010132>
- Korzynski, P., Mazurek, G., Altmann, A., Ejdy, J., Kazlauskaitė, R., Paliszkiwicz, J., ... Ziembra, E. (2023). Generative artificial intelligence as a new context for management theories: analysis of ChatGPT. *Central European Management Journal*, 31(1), 3-13. <https://doi.org/10.1108/CEMJ-02-2023-0091>
- Kotsis, K. T. (2025a). Artificial intelligence for physics education in STEM classrooms: A narrative review within a pedagogy-technology-policy framework. *Schrödinger: Journal of Physics Education*, 6(3), 204-2011. <https://doi.org/10.37251/sjpe.v6i3.2148>
- Kotsis, K. T. (2025b). Comparing ChatGPT and DeepSeek in addressing misconceptions about physics concepts. *European Journal of Contemporary Education and E-Learning*, 3(2), 191-206. [https://doi.org/10.59324/ejceel.2025.3\(2\).17](https://doi.org/10.59324/ejceel.2025.3(2).17)
- Kotsis, K. T. (2026). Artificial intelligence and conceptual understanding in physics education: Opportunities, risks, and research directions. *International Journal of Advanced Multidisciplinary Research and Studies*, 6(1), 1316-1321. <https://doi.org/10.62225/2583049X.2026.6.1.5673>
- Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K. E., Lukowicz, P., Kuhn, J., Küchemann, S., & Karolus, J. (2024). Unreflected acceptance: Investigating the negative consequences of ChatGPT-assisted problem solving in physics education. In *HHAI 2024: Hybrid Human AI Systems for the Social Good* (pp. 199-212). IOS Press. <https://doi.org/10.3233/FAIA240195>
- Küchemann, S., Rau, M., Neumann, K., & Kuhn, J. (2025). Editorial: ChatGPT and other generative AI tools. *Frontiers in Psychology*, 16, 1535128. <https://doi.org/10.3389/fpsyg.2025.1535128>
- Küchemann, S., Rau, M., Schmidt, A., & Kuhn, J. (2024). ChatGPT's quality: Reliability and validity of concept inventory items. *Frontiers in psychology*, 15, 1426209. <https://doi.org/10.3389/fpsyg.2024.1426209>
- Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science*, 208(4450), 1335-1342. <https://doi.org/10.1126/science.208.4450.1335>
- Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
- López-Simó, V., & Rezende, M. F. (2024). Challenging ChatGPT with different types of physics education questions. *The Physics Teacher*, 62(4), 290-294. <https://doi.org/10.1119/5.0160160>
- Moundridou, M., Moutis, N., Charalampopoulos, K., & Matzakos, N. (2026). Generative AI in physics education: A PRISMA-guided systematic review of empirical studies. *Artificial Intelligence and Applications*, 4(2), 149-164. <https://doi.org/10.47852/bonview-AIA62027183>
- Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. S. Siegler (Ed.), *Children's thinking: What develops?* (pp. 325-348). Lawrence Erlbaum.
- Singh, C., Maries, A., Heller, K., & Heller, P. (2023). Instructional strategies that foster effective problem-solving. In M. F. Taşar and P. R. L. Heron (Eds.), *The international handbook of physics education research: Learning physics*. AIP Publishing. https://doi.org/10.1063/9780735425477_017
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4
- Tschisgale, P., Maus, H., Kieser, F., Kroehs, B., Petersen, S., & Wulff, P. (2025). Evaluating GPT- and reasoning-based large language models on Physics Olympiad problems: Surpassing human performance and implications for educational assessment. *Physical Review Physics Education Research*, 21, 020115. <https://doi.org/10.1103/6fmX-bsnl>
- Wang, K. D., Burkholder, E., Wieman, C., Salehi, S., & Haber, N. (2024). Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. *Frontiers in Education*, 8, 1330486. <https://doi.org/10.3389/feduc.2023.1330486>

- Wei, Y., Zhang, R., Zhang, J., Qi, D., & Cui, W. (2025). Research on Intelligent Grading of Physics Problems Based on Large Language Models. *Education Sciences*, 15(2), 116. <https://doi.org/10.3390/educsci15020116>
- West, C. G. (2023). Advances in apparent conceptual physics reasoning in GPT-4. *arXiv*. <https://doi.org/10.48550/arXiv.2303.17012>
- Xu, X., Xu, Q., Xiao, T., Chen, T., Yan, Y., Zhang, J., Diao, S., Yang, C., & Wang, Y. (2025). UGPhysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2502.00334>
- Xu, Y., Kimlee, H., Xiao, Y., & Luo, D. (2025). Advancing AI-scientist understanding: Multi-agent LLMs with interpretable physics reasoning. *arXiv*. <https://doi.org/10.48550/arXiv.2504.01911>
- Yeadon, W., & Hardy, T. (2024). The impact of AI in physics education: A comprehensive review from GCSE to university levels. *Physics Education*, 59(2), 025010. DOI: 10.1088/1361-6552/ad1fa2
- Younis, A. (2025). Advancing Learning Outcomes in Physics Education through Artificial Intelligence Integration. *European Journal of Education and Pedagogy*, 6(5), 34-39. <https://doi.org/10.24018/ejedu.2025.6.5.996>