Evaluating the Capabilities of Large Language Models in Solving Subject-specific Problems: A Case Study of High School Physics Mechanics

Yuqi Li

Golden Apple Jincheng No.1 Secondary School, Chengdu 610095, China
Email: liyuqi907@gmail.com

***Abstract*—As generative artificial intelligence is developing at an unprecedented pace, the evaluation of its logical reasoning abilities in vertical disciplines has become a research focus. This paper takes high school physics mechanics as a case study to systematically assess the subject-specific problem-solving abilities of Large Language Models (LLMs). To ensure an unbiased and accurate assessment of their problem-solving abilities, an evaluation set comprising 45 classical mechanics problems was constructed for this study. The evaluation set covers three difficulty levels (basic, medium, advanced) in a balanced manner and adopts a dual ten-point scoring scheme that assigns five points to the final answer and five points to the solution process. Under this scheme, empirical quantitative assessments were performed on DeepSeek, ChatGPT, and Gemini. The results show that Gemini achieved the highest overall score rate (95.78%), demonstrating exceptional stability in multimodal diagram interpretation and higher-order reasoning. ChatGPT exhibited the strongest algebraic derivation resilience in high-difficulty computational tasks, with a score rate of 93.33%. DeepSeek attained an overall score rate of 85.56%, but encountered capability bottlenecks in high-difficulty and multimodal diagram tasks due to visual perception biases and deficiencies in solution standardization.**

***Keywords*—large language model, physics education, problem solving assessment, multi-dimensional scoring, high school mechanics**

I. Introduction

As artificial intelligence develops rapidly, the application of AI to solving problems and providing detailed explanations of knowledge has become increasingly widespread in the field of intelligent education, especially in science and engineering disciplines.

Assessments of Large Language Models (LLMs) in these disciplines have gradually shifted from general common-sense questions and answers to deeper, more specialized domains. At the macro level, research has confirmed that LLMs are approaching the human levels in the ability of resolving university-level scientific questions and explaining physical puzzles [1, 2]. At the meso-level, specialized evaluations of physical perception and reasoning, such as SCIBENCH [3] and PHYBench [4], as well as benchmarks for high school physics Olympiads like High School Physics Olympiad Benchmark (HIPHO) [5], have been proposed. However, rigorous assessment of standardized tests for specific regions (such as Chinese middle school and high school physics problems) displays that LLMs are unable to authentically possess complete "Physical Science Thinking" [6]. In the education of mechanics and physical engineering, LLMs regularly have potential confusion of the interpretation of mechanics concepts [1], while their highly fluent text can easily obscure their lack of scientific accuracy [7].

Previous LLMs display apparent deficiencies in fine-grained evaluation of high school physics mechanics, a field that emphasizes multimodal image perception, the construction of complex physical models, and the multi-step rigorous integration of mathematics. Under the current circumstances this study focuses on assessing the capabilities of large language models in the specific discipline of high school physics and mechanics. By constructing a question set that can precisely test, DeepSeek, ChatGPT, and Gemini are tested and evaluated, and the scores of each large language model (including the final answer and process score) are quantified. Ultimately, this study aims to clarify the performance and logical approach of mainstream large language models in solving specific disciplinary problems.

This research will reveal the strengths and weaknesses in the vertical disciplines of mathematical logic and conceptual understanding of mainstream large language models, providing a fundamental basis for the optimization of personalized Q&A and automated correction algorithms in future "AI education".

II. Methodology

A. Evaluation Set Design and Multi-label Mapping of Question Types

To ensure that all the questions in the assessment set are

Manuscript received November 11, 2026; accepted November 25, 2026; published November 26, 2026.

representative and the question-setting logic is scientific, the questions from three places were collected: the practice questions in Khan Academy Physics 1, the past real questions of the Physics Bowl, and the practice questions of AP Physics C Mechanics in College Board. These three question sources almost cover all the knowledge points of mechanics in high school physics. Fifteen questions are collected from each source to ensure that sample size variations across sources would not skew the results.

Before constructing the test set, this study developed a classification system for both difficulty levels and question types. The difficulty is divided into three levels: Basic, Medium and Advanced. The classification criteria are detailed in Table I.

TABLE I. CRITERIA FOR QUESTION DIFFICULTY CLASSIFICATION

Difficulty Level	Criteria
Basic	• Direct application of formulas (e.g., F = ma, W = Fd) • Number of steps ≤ 2 • Simple, idealized scenarios (no friction, flat surfaces) • No critical conditions or assumptions involved
Medium	• Requires concatenation of two or more core formulas (e.g., first use Newton's second law to find a, then use kinematic equations to find *v*) • Involves basic force analysis or energy process analysis • Contains a common "turning point" (e.g., the threshold where an object just starts to slide) • Requires understanding and converting graphical information (e.g., calculating impulse from an *F–t* graph)
Advanced	• Involves two or more physical processes (e.g., collision followed by circular motion) • Requires setting up custom equations or establishing geometric relationships • Involves precise differentiation between easily confused concepts (e.g., instantaneous velocity vs. average velocity, tension vs. normal force) • Contains hidden conditions or traps that require deep comprehension of the problem statement to uncover • Involves moderately complex mathematical manipulation (e.g., solving quadratic equations, trigonometric transformations)

Question types are classified into four categories based on the framework of core physics literacy: calculation, graph recognition, and concept and reasoning.

B. Dual-linkage Scoring Mechanism

To evaluate the performance of large language models in the problem-solving process more comprehensively and accurately, this research adopts "5 + 5" two-dimension scoring mechanism, with a full score of 10 points. Among them, 5 points are allocated to the accuracy of the final answer, and 5 points are allocated to the completeness and logic of the solution process as shown in Tables II and III.

This scoring design serves two primary purposes. First, decoupling procedural evaluation from final answers mitigates false positives resulting from flawed logic, thereby enhancing assessment reliability. Second, independent process scoring provides a more precise evaluation of model reasoning in terms of formula selection, step organization, and reasoning chain construction.

TABLE II. DIMENSION A: CONCEPTUAL UNDERSTANDING AND PROCEDURAL REASONING (5 POINTS)

Score	Criteria
5 (Excellent)	All relevant physical concepts are correctly identified; formulas are applied without error; the reasoning chain is complete and logically coherent.
3–4 (Good)	Core concepts are accurate; main steps are present, though minor omissions or redundancies may occur.
1–2 (Poor)	Key concepts are misapplied or unclear; essential steps are missing or flawed.
0	The response is irrelevant or reflects a fundamental misunderstanding of the underlying physics.

TABLE III. DIMENSION B: FINAL ANSWER ACCURACY (5 POINTS)

Score	Criteria
5 (Perfect)	The final answer is numerically correct, includes appropriate units, and fully addresses all parts of the question.
3 (Partially Correct)	The core result is essentially correct but contains one or more minor defects, such as: giving only a ratio or expression without computing the numerical value; correct numerical value but missing or incorrect units; answering the primary question but neglecting a secondary sub-question.
1 (Related but Incorrect)	An attempt was made to produce an answer, but the computation contains a substantive error.
0	No answer is provided, or the answer is completely incorrect.

Total Score = Dimension A Score + Dimension B Score (maximum 10 points).

The total score will be calculated after assessing the performance of LLMs.

C. Prompt Design and Interaction Strategy

To ensure the objectivity and repeatability of the entire process, this study controlled several variables: question format and prompts, as well as input method for problems.

First, to eliminate potential biases in results arising from differences in phrasing, a fully standardized prompt template was adopted, as follows:

"You are a professional physics tutor who answers physics questions in a rigorous, accurate, and step-by-step manner. Please first understand the problem, then derive the solution step by step, and finally provide a clear final answer. Respond according to the following format:

(1) Analysis: Briefly describe the core physical concepts involved and the given conditions.
(2) Steps: Clearly present the derivation and calculation process, with each step supported by sound physical reasoning.
(3) Answer: In the final line, begin with 'Final Answer:' followed by the numerical value (with units) or expression. Ensure all calculations are accurate, physical concepts are correctly applied, and the response is complete—do not truncate it midway."

Second, regarding the input format of the questions, since each question varies in whether it requires image recognition, this study directly provides the questions from

the evaluation set to LLMs as images. These images include the question stem, the problem statement, and, for certain image-based questions, the images that need to be recognized.

Moreover, to prevent LLMs from retaining previous context when answering questions, this study adopted a single-turn dialogue mode.

D. Selection of LLMs

Three representative LLMs—ChatGPT, Gemini, and DeepSeek—were selected based on their technical architectures and benchmark standing:

- ChatGPT (represented by the GPT-4 series): serves as an industry benchmark for closed-source commercial models. Technical reports confirm its expert-level mathematical logic and Chain-of-Thought (CoT) reasoning in standardized examinations [8, 9].
- Gemini (represented by the Gemini 1.5 series): incorporates a native multimodal Mixture-of-Experts (MoE) architecture, demonstrating strong context-aware capabilities in Science Reasoning Tasks involving visual diagrams [10].
- DeepSeek (represented by the DeepSeek-R1 series): Represents cutting-edge open-source Reinforcement Learning (RL) reasoning models. Driven by pure RL, it exhibits strong self-reflection and deep reasoning capabilities in verifiable STEM tasks [11].

III. Quantifying LLM Performance in Mechanics Problem Solving

After several days of testing and analysis, this study obtained the performance of each major language model in terms of problem-solving logic, computational power, and graph recognition ability from the responses of the three major language models in the evaluation set.

A. Overall Problem-solving Performance and Distribution of Score Characteristics

As shown in Table IV and Fig. 1, the results indicate that, in terms of total scores and overall accuracy, all three models demonstrated extremely high consistency (the total score of process points is all above 200 points), and the total score has even exceeded the 85% threshold of the subject ability, which proves that large language models have already demonstrated high feasibility in handling high school physics mechanics subjects that are highly logical and require strong formula derivation. Statistically, the median score for each model was a perfect score (10.00 points), indicating a significant negatively skewed distribution. This means that the model does not "deduct a little point for each question", but "most questions can get full marks, and only when encountering specific visual or logical blind spots will there be catastrophic deduction of points". Gemini has the smallest standard deviation (1.71), and its pass rate (93.33%) and full score rate (93.33%) are highly aligned, indicating that its performance is the most stable. In contrast, DeepSeek has a standard deviation as high as 2.82, with a minimum score of 1 point, suggesting frequent catastrophic errors.

TABLE IV. Overall Performance Metrics and Score Distribution of Tested LLMs

Evaluation Criteria	DeepSeek	ChatGPT	Gemini
Total Questions	45	45	45
Total Score (Final Answer)	183	207	213
Total Score (Process)	202	214	218
Total Score	385	421	431
Mean	8.56	9.36	9.58
Median	10	10	10
Standard Deviation	2.82	1.96	1.71
Maximum Score	10	10	10
Minimum Score	1	2	2
Total Score Rate	85.56%	93.56%	95.78%
Pass Rate (≥8 points)	82.22%	88.89%	93.33%
Perfect Score Rate (10 points)	73.33%	88.89%	93.33%

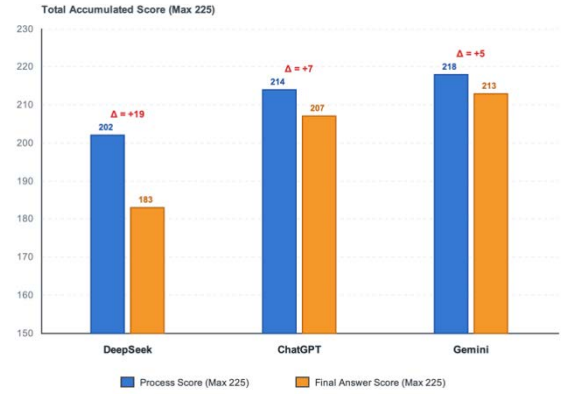

Fig. 1. Decoupled comparison of process scores and final answer scores.

B. Model Performance Differentiation Under the Progressive Difficulty Ladder

By comparing performance across the three difficulty levels—basic (B), medium (M), and advanced (A)—this study explores the performance ceilings and resilience of different models as the physical context grew more complex.

According to the Table V and Fig. 2, the performance of the three large language models in the three difficulty levels of questions can be analyzed.

Across basic difficulty problems, ChatGPT and Gemini achieved flawless execution, demonstrating complete mastery over foundational, single-formula mechanics tasks, whereas DeepSeek exhibited minor performance gaps. At the medium level, a performance inversion was observed, with DeepSeek outperforming ChatGPT. This disparity is attributable to the "verbosity bias" and "length bias" inherent in strongly RLHF-aligned models [12, 13], where ChatGPT's excessively lengthy responses caused ambiguities in system boundary definitions and led to omitted terms in multi-state energy conservation. As defined by recent research on "verbosity compensation" [14], models frequently generate excessive tokens to mask underlying mathematical uncertainties, which paradoxically degrades scientific accuracy. In contrast, DeepSeek's reinforcement-learning-driven reasoning flow maintained higher formula derivation precision. For advanced problems requiring long reasoning chains, calculus, and high-order algebra, ChatGPT and Gemini demonstrated strong mathematical

resilience, successfully navigating technical ceilings. Conversely, DeepSeek's performance declined markedly, highlighting a clear performance bottleneck when handling high-difficulty comprehensive mechanics tasks.

TABLE V. MODEL PERFORMANCE DIFFERENTIATION ACROSS PROGRESSIVE DIFFICULTY LEVELS

Difficulty level	Evaluation Criteria	DeepSeek	ChatGPT	Gemini
Basic Difficulty (B)	Total Questions	15	15	15
	Total Score (Final Answer)	66	75	75
	Total Score (Process)	71	75	75
	Total Score	137	150	150
	Mean	9.13	10	10
	Total Score Rate	91.33%	100.00%	100.00%
	Pass Rate (≥8 points)	86.67%	100.00%	100.00%
	Perfect Score Rate (10 points)	80.00%	100.00%	100.00%
Medium Difficulty (M)	Total Questions	15	15	15
	Total Score (Final Answer)	66	64	70
	Total Score (Process)	68	67	72
	Total Score	134	131	142
	Mean	8.93	8.73	9.47
	Total Score Rate	89.33%	87.33%	94.67%
	Pass Rate (≥8 points)	86.67%	80.00%	93.33%
	Perfect Score Rate (10 points)	86.67%	80.00%	93.33%
Advanced Difficulty (A)	Total Questions	15	15	15
	Total Score (Final Answer)	51	68	68
	Total Score (Process)	63	72	71
	Total Score	114	140	139
	Mean	7.6	9.33	9.27
	Total Score Rate	76.00%	93.33%	92.67%
	Pass Rate (≥8 points)	73.33%	86.67%	86.67%
	Perfect Score Rate (10 points)	53.33%	86.67%	86.67%

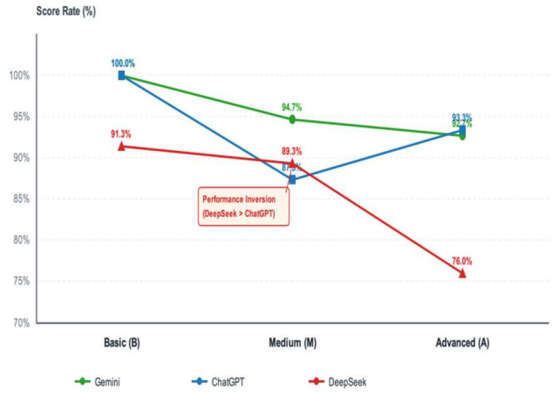

Fig. 2. Score rate trajectory across progressive difficulty levels.

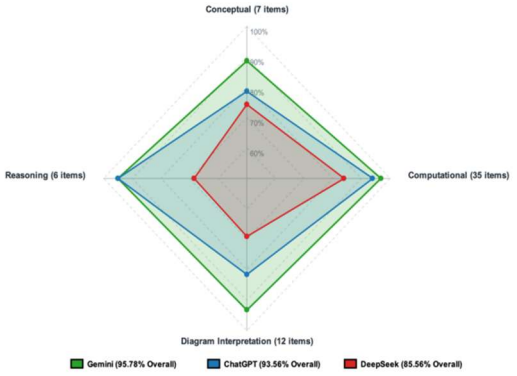

Fig. 3. Performance across core physics literacy question types.

C. The Strengths and Weaknesses of Core Subject Literacy Question Types

This study also aims to analyze whether large language models have excellent problem-solving capabilities in the core literacy question types of disciplines. Accordingly, the performance of each model was examined across the four question-type dimensions.

As shown in Table VI and Fig. 3, this study finds that large language models perform differently when handling problems in various aspects, and there are also differences between different models.

Performance across core literacy question types reveals distinct capabilities and architectural boundaries among LLMs. While all models demonstrated proficiency in computational problems through symbolic operations and algebraic solving, conceptual understanding remains an area needing further improvement.

Diagram interpretation serves as a primary performance divide, highlighting the architectural gap between end-to-end native multimodal alignment and cross-platform cascaded multimodal interaction [15]. As PHYBench and HIPHO benchmarks emphasize, physical problem solving relies on real-time vision-logic collaboration, where the primary bottleneck often lies in "physical perception" rather than symbolic deduction [4, 5]. In a cascaded pipeline, front-end visual parsing errors create an "information bottleneck" that isolated back-end inference flows cannot dynamically correct, resulting in a noticeable discrepancy between procedural reasoning and final answer accuracy.

In reasoning tasks, performance disparities stemmed from two primary mechanisms underlying models' tendency to omit explanatory steps or provide unsound rationales. First, pure symbolic contexts without numerical anchors reduce the activation rate of advanced physical

theorems [3]. Second, reinforcement learning models tend to develop "reasoning shortcuts" by prioritizing terminal rewards, causing them to converge on correct answers while skipping explicit causal justifications in their generated outputs [5].

TABLE VI. MODEL PERFORMANCE ACROSS CORE SUBJECT LITERACY QUESTION TYPES

Question Types	Evaluation Criteria	DeepSeek	ChatGPT	Gemini
Conceptual	Total Questions	7	7	7
	Total Score (Final Answer)	25	25	30
	Total Score (Process)	27	30	32
	Total Score	52	55	62
	Mean	7.43	7.86	8.86
	Total Score Rate	74.29%	78.57%	88.57%
	Pass Rate (≥8 points)	71.43%	71.43%	85.71%
	Perfect Score Rate (10 points)	71.43%	71.43%	85.71%
Computational	Total Questions	35	35	35
	Total Score (Final Answer)	138	162	168
	Total Score (Process)	156	167	171
	Total Score	294	329	339
	Mean	8.4	9.4	9.69
	Total Score Rate	84.00%	94.00%	96.86%
	Pass Rate (≥8 points)	80.00%	88.57%	94.29%
	Perfect Score Rate (10 points)	68.57%	88.57%	94.29%
Diagram Interpretation	Total Questions	12	12	12
	Total Score (Final Answer)	38	46	55
	Total Score (Process)	45	52	57
	Total Score	83	98	112
	Mean	6.92	8.17	9.33
	Total Score Rate	69.17%	81.67%	93.33%
	Pass Rate (≥8 points)	58.33%	75.00%	91.67%
	Perfect Score Rate (10 points)	58.33%	75.00%	91.67%
Reasoning	Total Questions	6	6	6
	Total Score (Final Answer)	16	28	28
	Total Score (Process)	25	29	29
	Total Score	41	57	57
	Mean	6.83	9.5	9.5
	Total Score Rate	68.33%	95.00%	95.00%
	Pass Rate (≥8 points)	66.67%	83.33%	83.33%
	Perfect Score Rate (10 points)	16.67%	83.33%	83.33%

IV. CONCLUSION

Evaluating DeepSeek, ChatGPT, and Gemini on high school physics mechanics problems through a dual procedural-answer scoring scheme yields key conclusions that systematically address the core research questions.

Regarding overall subject feasibility, all three LLMs demonstrated strong problem-solving capabilities, maintaining pass rates above 85.56%. However, process scores systematically surpassed final answer scores, exposing a cascading error propagation effect in autoregressive derivations and highlighting a structural mismatch between strong symbolic modeling and downstream arithmetic accuracy. Across varying difficulty levels and task types, model performance exhibited uneven degradation as problem complexity increased: ChatGPT and Gemini achieved full mastery on basic tasks, whereas medium-level problems exhibited a performance inversion in which DeepSeek's concise argumentation outperformed ChatGPT, which suffered from verbosity bias. On advanced problems, DeepSeek hit a pronounced performance ceiling, with its perfect score rate dropping to 53.33%. Furthermore, examining the cognitive root causes of errors indicates that diagram interpretation and reasoning tasks marked the primary performance divide among the models, where DeepSeek's lower scores in these categories resulted from visual-textual alignment loss within a third-party cascaded pipeline (Tencent Yuanbao), alongside reinforcement-learning-driven reasoning shortcuts that omitted explicit causal justifications during final text generation.

Several limitations of this study should be acknowledged. First, the dataset (n = 45) is restricted solely to mechanics, omitting core domains such as electromagnetism, thermodynamics, and optics, while question-type distributions across literacy dimensions require finer balancing for stricter variable control. Second, channel heterogeneity existed because DeepSeek was evaluated via a third-party cascaded channel (Tencent Yuanbao) rather than through native end-to-end interfaces, as was the case for ChatGPT and Gemini. Finally, the methodology was limited to zero-shot Chain-of-Thought (Zero-shot-CoT) prompting, without exploring the potential effects of few-shot prompting or multi-agent collaborative strategies.

To address the aforementioned shortcomings, future research will proceed along the following two directions:

- Expansion of knowledge graphs and sample sets: In the future, a large-scale, multi-modal evaluation set covering the complete physical science knowledge system for the K-12 stage will be constructed.

- Trace verification of the native architecture: Once the official native multi-modal visual interface of DeepSeek is fully open, supplementary closed-loop experiments will be conducted to thoroughly investigate the mechanism by which end-to-end alignment improves physical perception accuracy.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] X. Wang, Z. Hu, P. Lu *et al.*, "Scibench: Evaluating college-level scientific problem-solving abilities of large language models," arXiv Preprint, arXiv:2307.10635, 2023.

[2] J. Ding, Y. Cen, and X. Wei, "Using large language model to solve and explain physics word problems approaching human level," arXiv Preprint, arXiv:2309.08182, 2023.

[3] S. Qiu, S. Guo, Z.-Y. Song *et al*., "PHYBench: Holistic evaluation of physical perception and reasoning in large language models," arXiv Preprint, arXiv:2504.16074, 2025.

[4] F. Yu, H. Wan, Q. Cheng *et al*., "HIPHO: How far are (M)LLMs from humans in the latest high school physics olympiad benchmark?" arXiv Preprint, arXiv:2509.07894, 2025.

[5] X. Zhang, "Does ChatGPT exhibit physics scientific thinking? A rigorous assessment using chinese-style middle and high school exam questions," *The Physics Teacher*, vol. 63, no. 9, pp. 693–697, Nov. 2025.

[6] J. Tian, J. Hou, Z. Wu *et al*., "Assessing large language models in mechanical engineering education: A study on mechanics-focused conceptual understanding," arXiv Preprint, arXiv:2401.12983, 2024.

[7] M. N. Dahlkemper, S. Z. Lahme, and P. Klein, "How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT," *Physical Review Physics Education Research*, vol. 19, no. 1, 010142, 2023.

[8] J. Achiam, "GPT-4 technical report," arXiv Preprint, arXiv:2303.08774, 2023.

[9] A. Hurst *et al.*, "GPT-4o system card," arXiv Preprint, arXiv:2410.21276, 2024.

[10] G. Team *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," arXiv Preprint, arXiv:2403.05530, 2024.

[11] D. Guo *et al.*, "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," arXiv Preprint, arXiv:2501.12948, 2025.

[12] A. Singhal, A. Köpf, S. Roux *et al*., "A Long way to go: investigating length correlations in RLHF," arXiv Preprint, arXiv:2403.05530, 2024.

[13] Y. Dubois, T. Galambosi, A. Compagnol *et al*., "AlpacaEval: A lightweight and high-quality evaluation benchmark with length-controlled win rates," arXiv Preprint, arXiv:2404.04123, 2024.

[14] Y. Zhang *et al*., "Demystify verbosity compensation behavior of large language models," in *Proc. the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, 2025.

[15] H. Liu, C. Li, Q. Wu *et al*., "Visual instruction tuning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 34879–34892, 2023.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).