PhySense: Evaluating LLMs on Foundational Physics Principles

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models (LLMs) have rapidly advanced and are increasingly capable of tackling complex scientific problems, including those in physics. Despite this progress, current LLMs often fail to emulate the concise, principle-based reasoning characteristic of human experts, instead generating lengthy and opaque solutions. This discrepancy highlights a crucial gap in their ability to apply core physical principles for efficient and interpretable problem solving. To systematically investigate this limitation, we introduce *PhySense*, a novel *principle-based physics reasoning* benchmark designed to be easily solvable by experts using guiding principles, yet deceptively difficult for LLMs without principle-first reasoning. Our evaluation across multiple state-of-the-art LLMs and prompt types reveals a consistent failure to align with expert-like reasoning paths, providing insights for developing AI systems with efficient, robust and interpretable principle-based scientific reasoning.

1 Introduction

Large language models (LLMs) have emerged as powerful tools, profoundly impacting numerous aspects of scientific discovery [1, 2, 3, 4]. Recent advancements in their reasoning capabilities have been particularly transformative, with notable applications in the domain of physics [5, 6, 7, 8]. Within physics, LLMs have demonstrated the ability to engage with problems ranging from those requiring real-world physical intuition [9] to complex theoretical challenges [10].

Despite these impressive strides, a critical challenge lies in ensuring that the reasoning processes of LLMs align with expert intuition and fundamental physical principles. Current LLMs tend to generate solutions with long-horizon reasoning pathways, which are opaque, convoluted, or divergent from the parsimonious and principle-driven thinking characteristic of human physicists. Such phenomena has also been identified as over-thinking [11]. In contrast, physicists master *principle-based reasoning* with principle-driven problem solving and principle-based verification. Principle-driven problem solving is a forward process where fundamental principles simplify the problem-solving space, directly guiding towards a solution. Principle-based verification is a routine where physics principles establish criteria that a correct solution must meet, ensuring its validity. This divergence between LLMs and human physicists raises concerns about the *efficiency, robustness and interpretability* of current LLMs for scientific reasoning, especially in a field where clarity, intuition and explainability of a solution is as crucial as the correctness of solution itself.

This work investigates LLMs' tendency to miss simple, intuitive solutions in physics problems that are apparent to human physicists. We posit that an incomplete grasp or misapplication of physical principles leads LLMs to unnecessarily complex reasoning, contrasting with human experts who leverage these fundamental ideas for elegant and efficient solutions (e.g., analyzing through symmetry instead of intricate numerical computation). This expert approach, which organizes knowledge around crystallized principles for efficient problem-solving, is well-documented in cognitive science [12, 13, 14]. Emulating this in LLMs could foster more aligned, efficient, and interpretable reasoning, guiding them towards computationally leaner and conceptually sound 'shorter paths.'

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

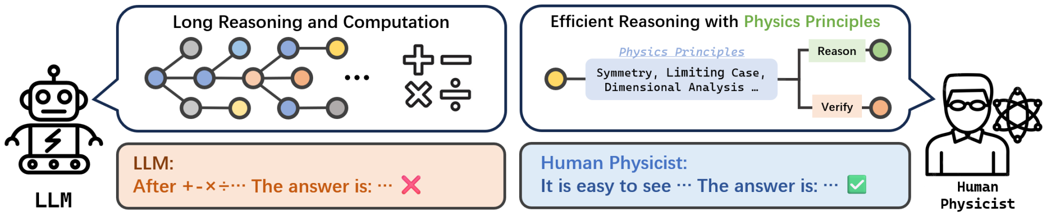


Figure 1: Illustrating how LLMs use lengthy, complex reasoning for physics problems intuitively straightforward to scientists applying core physical concepts.

To systematically analyze this, we introduce *PhySense*, a benchmark of 380 carefully curated physics problems designed to be straightforward for human physicists with core principles but appears to be challenging for LLMs to solve efficiently. In contrast to other physics reasoning benmark which focuses on reasoning on specific domain or challenging calculations, *PhySense* focuses on short reasoning chains where single principles are crucial. Our findings aim to highlight the need for developing LLMs that are not only accurate, but also exhibit interpretable, robust reasoning aligned with fundamental physical principles. Our key contributions are summarized as follows:

- We introduce *PhySense*, *the first novel, human-curated principle-based physics reasoning benchmarking dataset* of over 380 problems that are straightforward for experts using fundamental principles but challenging to LLMs unless they adopt direct, principle-first reasoning shortcuts.
- We quantify not only whether an LLM arrives at the correct answer, but also how closely its reasoning cost matches with principle-based solutions via both accuracy and token efficiency metrics.
- We evaluate a range of state-of-the-art LLMs under zero-shot, hint, and no-computation prompts, uncovering LLMs' systematic lack of capability in applying principles and offering guidance for training LLMs toward more efficient, robust and interpretable principle-based physics reasoning.

2 Related Work

Benchmarks for General Scientific Reasoning As LLMs are increasingly considered as important tools in scientific inquiry, understanding their true capabilities and limitations in scientific reasoning becomes paramount. Early benchmarks such as AI2 ARC [15], MMLU [16], IconQA [17] and ScienceQA [18] focused on general scientific context, surface-level reasoning, and basic factual knowledge. As model capabilities have grown, newer evaluations target deeper, multi-step problem solving and domain-specific expertise—either by repurposing advanced human exams and problem set (e.g., AGIEval [19], JEEBench [20], SciBench [21]) or by probing complex reasoning dimensions (e.g., MMLU-Pro [22], SciEval [23], TheoremQA [24]), up to the extreme challenges posed by capstone-style assessments like Humanity's Last Exam [25]. Some of the general science reasoning datasets like OlympicBench [26] and OlympicAreana [27] provides advanced physics problems but with limited scope.

Benchmarks for Physics Reasoning The landscape of physics-reasoning benchmarks for LLMs has rapidly evolved from primarily general problem sets to multifaceted collections that probe deeper conceptual, procedural, and physics-specific understanding. Efforts like PhyQA [28] and UGPhysics[29] assemble thousands of structured introductory problems, while other benchmarks such as PhysBench [30] and PhysReason [31] introduce problems require longer reasoning steps. More research-oriented suites like TP-Bench [10], CURIE [32] and multi-modal benchmarks like MM-PhyQA [33] and domain specific benchmarks like FEABench [34] further pushes the understanding of LLM's physics capability with more research-oriented settings. In contrast to multi-modal approaches, our work deliberately focuses on theoretical, text-only problems where all relevant information is conveyed textually. This design choice allows for a targeted evaluation of conceptual and algebraic reasoning, isolating these core competencies from confounding factors of image or diagram understanding. We discovered that single-modality benchmark already reveals significant limitations in current LLMs. Increased attention is also being directed towards fine-grained evaluation methodologies for the precise assessment of many-step reasoning including Expression Edit Distance (EED) Score [30].

Reasoning in LLMs and "Over-Thinking" Recent advances in LLMs, sometimes characterized by "slow thinking" capabilities demonstrated since models like GPT o1 [35], have showcased stronger abilities in solving STEM problems. This improvement is often attributed to post-training techniques and reinforcement learning. Models like DeepSeek-R1[36], Gemini-2.0-Flash-Thinking[37], and versions of Claude [38] and Qwen [39] have demonstrated enhanced reasoning. However, while these models can generate longer reasoning chains (i.e., use more tokens), this does not always equate to more efficient or accurate reasoning. The phenomenon of "over-thinking" [11], where models may engage in unnecessarily complex or incorrect reasoning paths, remains a challenge.

3 Dataset Generation

"The universe is an enormous direct product of representations of symmetry groups."

— Steven Weinberg, Nobel laureate in physics

Principle-based Reasoning Physics principles such as symmetries, conservation laws, and dimensional analysis remain cornerstones of modern physics research and problem solving. They not only simplify complex systems and reduce computational costs, but also illuminate the nature of various phenomena and provide a unified understanding across diverse contexts. Therefore, an LLM's proficiency in applying these principles serves as a reliable gauge of its understanding of physics. Principle-based physics reasoning can (1) **efficiently yield the correct answer** (2) **robustly validate potential solutions** (3) **provide clear interpretability beyond calculation**. We demonstrate this with the following example.

Example 1

A 5x5 square grid of nodes: $x \in \{0, 1, 2, 3, 4\}$, $y \in \{0, 1, 2, 3, 4\}$ connected by resistors r between nearest neighbors. Connect node $V_{(0,0)} = 0$, node $V_{(4,4)} = V$, node $V_{(0,4)} = V/2$. Which of the following is true?
(a) $V_{(1,3)} = V/2$ (b) $V_{(2,2)} = V/2$ (c) $V_{(1,1)} = V/4$ (d) $V_{(3,3)} = 3V/4$ (e) $V_{(4,0)} = V/2$

Answer 1

Answer by symmetry principle:
A trained physicist would notice the circuit together with added voltages has a reflection symmetry along the diagonal $x + y = 4$. One can then deduce directly that (a,b,e) is correct.

Answer by explicit calculation:
Without using symmetries, one has to solve Kirchhoff equations for the whole system (22 unknown voltages),

$$3V_{0,1} - V_{1,1} - V_{0,2} = 0,\ 3V_{0,2} - V_{1,2} - V_{0,1} - V_{0,3} = 0,\ 3V_{0,3} - V_{1,3} - V_{0,2} = V/2$$
$$3V_{1,0} - V_{2,0} - V_{1,1} = 0,\ 3V_{2,0} - V_{1,0} - V_{3,0} - V_{2,1} = 0,\ 3V_{3,0} - V_{2,0} - V_{4,0} - V_{3,1} = 0$$
$$2V_{4,0} - V_{3,0} - V_{4,1} = 0,\ 3V_{4,1} - V_{4,0} - V_{4,2} - V_{3,1} = 0,\ 3V_{4,2} - V_{4,1} - V_{4,3} - V_{3,2} = 0$$
$$3V_{4,3} - V_{4,2} - V_{3,3} = V,\ 3V_{1,4} - V_{2,4} - V_{1,3} = V/2,\ 3V_{2,4} - V_{1,4} - V_{3,4} - V_{2,3} = 0$$
$$3V_{3,4} - V_{2,4} - V_{3,3} = V,\ 4V_{i,j} - V_{i-1,j} - V_{i+1,j} - V_{i,j-1} - V_{i,j+1} = 0 \text{ for } 1 \leq i, j \leq 3.$$

Solving all the equations above numerically, one gets $V_{1,3} = V_{2,2} = V_{4,0} = V/2$, $V_{3,3} \approx 0.6702V$, $V_{1,1} \approx 0.3298V$. Thus the answer is (a,b,e). Clearly, this "standard" approach is much more complicated than using the symmetry principle.

Despite the power of physical principles, existing benchmarks (see e.g. Sec. 2), while challenging, do not evaluate whether LLMs truly apply these principles. Do LLMs genuinely understand physics, or are they merely leveraging greater computational power than humans? To address this gap, we have developed a new problem set of 380 physics questions spanning electricity and magnetism, electric circuits, quantum spin/fermion chains, quantum dynamics, topological insulators, the renormalization group, and conformal field theory. These problems are crafted according to the following criteria.

Principle-based physics reasoning A key feature of *PhySense* is its design to test LLMs' understanding on fundamental principles and capability on principle-based reasoning. Our dataset is different than previous physics reasoning dataset, since we do not aim to test LLMs' knowledge in a specific domain or cabability of reasoning with long calculation. While our problems may be challenging or could be solved with lengthy calculation, we design the problems to be solved easily using physics principle reasoning.

Novel problems from human experts Although the underlying concepts in our problem set are widely available online, we have crafted entirely new questions with physicists from top universities that cannot be found elsewhere, ensuring that LLMs have not been exposed to similar problems. This novelty is essential for testing an LLM's ability to generalize the application of physics principles.

A wide range of difficulties The problems span difficulty levels from undergraduate through graduate and research-level, yet none requires advanced mathematical techniques, complicated integrals, or large-scale numerical computations. This ensures we evaluate how well LLMs can think like physicists — using fundamental physical principles to understand problems — rather than merely assessing raw computational capability. We also annotate each problem with a difficulty rating (as judged by humans) for subsequent analysis.

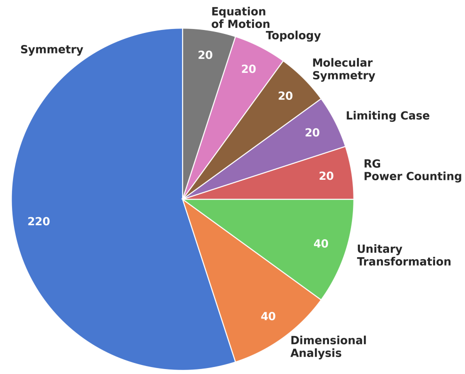


Figure 2: Distribution of physics principles in the dataset.

Conciseness for evaluation Every problem is stated and solved entirely through textual description and derivation. The physical setups are simple to describe, minimizing the risk of misinterpretation by LLMs. To eliminate ambiguity in the outputs, each question offers either multiple-choice options or expects a concise numerical answer.

3.1 Physical principles and models

Following the criteria above, we evaluate the LLM's understanding and correct application of several fundamental yet powerful principles in both classical and quantum physics. To do this, we design 19 distinct problem models

Symmetry Spatial symmetries can be leveraged to identify points where complicated integrals vanish. To evaluate this, we construct problem sets involving two-dimensional and three-dimensional electric (or magnetic) fields generated by symmetric charge (or current) distributions. These problems are categorized into the following models, each with an abbreviation:

- *2D electric field (2DEF), 2D electric field on a lattice (2DEFL), 3D electric field (3DEF), 2D magnetic field (2DBF), 3D magnetic field (3DBF)*

We also devise problems that leverage symmetries to determine voltages of certain nodes in finite and infinite circuits:

- *Infinite resistive lattices (InfRes), Circuits on a square lattice (SqGrd), Circuits on other lattices (OthGrd)*

The symmetry of molecules can determine the solubility in solvents, which leads to another model of problems:

- *Solubility comparison (Solub)*

Moreover, symmetries impose constraints on correlation functions in quantum many-body physics and statistical mechanics. We have also developed problems involving quantum spin and fermion chains and their dynamical variants, to test $\mathbb{Z}_2$, $U(1)$, and time-reversal symmetries alongside spatial symmetries such as translation and reflection:

- *Quantum spin chains (Qspin), Fermionic chains (Ferm), Quantum dynamics with symmetry and conservation laws (DynCon)*

Dimensional analysis Dimensional analysis is a powerful tool in uncovering possible relations between different physical quantities. Not only is it widely used in the context of thermodynamics, fluid mechanics, etc., its applications also extend to quantum mechanics as well. We design problems in two areas: (a) applying the Π theorem in fluid and quantum mechanics, and (b) using power counting to determine relevance in the renormalization group. This yields the following problem models:

• *Dimensional analysis using* Π *theorem, where we focus on testing LLM's ability to compute dimensions in arguments of functions such as* $\sin, \log$*, etc. (DimLS), Dimensional analysis with artificial irrelevant perturbations (WrdH), Power-counting in renormalization group analysis (RGPow)*

Limiting case Irrelevant perturbations in physical problems can be omitted to simplify the physical model. To test the LLM's ability to do so, we introduce perturbations into Model (WrdH) above and evaluate whether it correctly ignores the higher-order terms.

Conservation law Conservation law plays a crucial role in quantum field theory. Especially in free fermion conformal field theories, equation of motion, together with the fermionic statistics, provides a powerful tool to determine whether an operator is primary, descendant, or merely vanishing.

• *Operator properties in conformal field theories (CFTOp)*

Topology Topological phases of matter is a central topic in modern condensed matter physics. It typically exhibits gapless edge spectrum, and sensitive to the boundary condition of the system. We design problems to evaluate if LLMs can understand the stability of symmetry-protected topological phases from the edge spectrum perspective:

• *Edge spectrum in topological insulators (GpEdg)*

We also compose problems in counting the ground state degeneracy of (generalized) spin chain with antiperiodic or periodic boundary condition. In particular, in these problems, applying finite-depth local unitary circuits, which does not alter the topological property including the ground state degeneracy, greatly simplifies the calculation.

• *Ground state degeneracy of spin chains (GSDeq), Ground state degeneracy of generalized spin chains (GSDGen)*

4 Experiments

This section details the experiments conducted to evaluate the scientific problem-solving capabilities of LLMs. We begin by outlining the experimental setup, including the models tested and the prompting strategies employed to simulate scientific reasoning scenarios.

4.1 Experiment Setup

We evaluated seven unimodal LLMs on our benchmark. These included four **reasoning models** optimized for reasoning: GPT o4-mini-high [40], Claude Sonnet 3.7 Thinking [38], Gemini 2.5 Pro [41], and DeepSeek R1 [36]. Additionally, we tested three regular **non-reasoning models**: GPT 4.1 [42], Claude 3.7 Sonnet [43], and DeepSeek V3 [44]. For all the models, we use the API-based services with default hyperparameter setting. We utilized three common prompting strategies in scientific applications to test LLMs:

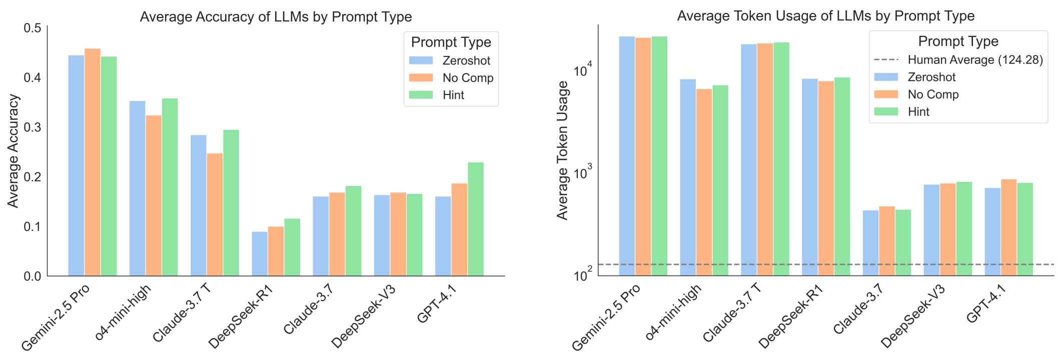
Zeroshot Prompting Zero-shot prompting tests a model's intrinsic reasoning by providing only the problem statement, without examples or hints. This method gauges the model's ability to apply existing knowledge, making it a good test for scientific discovery.

Hint Prompting Hint prompting provides models with guidance on relevant physical principles, helping to see if they can use explicit direction to solve problems. This is useful when models fail to apply the correct principles on their own.

No Computation Prompting In this approach, models are instructed to avoid complex calculations and instead focus on principle-based reasoning. This assesses their ability to prioritize simpler, conceptual solutions over complicated computational methods.

4.2 Metrics

We employ two primary metrics for evaluation: **accuracy** and **token usage**. For accuracy, LLMs were instructed to provide their final answer within a boxed environment for automated extraction and comparison against ground truth solutions. The problems fall into two categories with different evaluation implementation: (1) numerical: Answers are compared to the ground truth allowing for a 5% tolerance. (2) multiple choice: The selected option must exactly match the correct choice. For



(a) Average accuracy across models.

(b) Average token usage across models.

Figure 3: Average accuracy and token usage for different models.

token usage, we record the total number of tokens produced during the generation of the solution for each problem and model. This is a crucial metric that provides insight into the computational cost associated with each model's problem-solving process, and reflects how much principle-based reasoning each LLM acquires.

5 Results

In this section, we report the benchmarking results and present our primary observations regarding the performance of various LLMs on PhySense.

5.1 Reasoning Accuracy

We report the model performance in terms of accuracy score for each section and an average accuracy over all problems. We quantify model performance using accuracy percentage, calculated for each distinct problem category within our benchmark, alongside an overall average accuracy across all problems. This accuracy reflects the proportion of problems correctly solved by each model according to our evaluation protocol. The accuracy results are compiled in Table 1. To provide a clearer visual summary of the overall performance trends, we present histograms illustrating the distribution of average accuracy scores across the cohort of tested models in Figure 3a. To further assess LLM alignment with human physicist problem-solving, problems were categorized by human-judged difficulty (easy, medium, difficult). Figure 4 shows each model's average zero-shot accuracy across these levels. While reasoning models achieve better performance than non-reasoning models on average, all LLMs' performances are not satisfactory, reflecting their incapability of mastering principle-based reasoning.

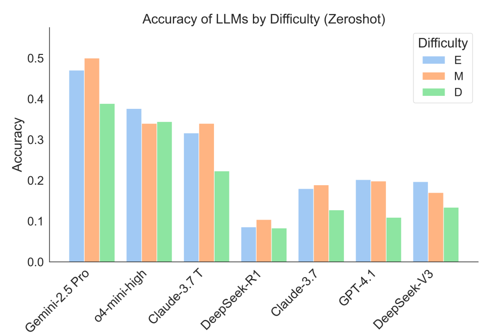


Figure 4: Zeroshot accuracy of LLMs under three difficulties.

5.2 Reasoning Token Efficiency

In parallel to accuracy, we report the average number of completion tokens produced by the models for generating solutions, both for individual sections and on average. This token usage metric provides an indication of the computational resources and reasoning complexity associated with each model's problem-solving attempts. The token usages are compiled in Table 2. Figure 3b offers a comparative overview of the token utilization patterns. Reasoning models, due to reaoning mechanisms, consume about ten times more tokens ($\sim 10^4$) than non-reasoning models ($\sim 10^3$). In stark contrast, human physicists demonstrate far greater efficiency, often solving the same problems using about a hundered times fewer tokens ($\sim 10^2$) than reasoning models. It indicates a huge gap between LLMs and human experts on efficient principle-based reasoning.

Table 1: LLM accuracy scores (as percentages) for reasoning models. The first subtable shows accuracy for the first 10 problem sets, and the second subtable shows accuracy for the remaining 9 problem sets and the overall average (AVG). The best accuracy of each section is marked in bold font. The full result, including non-reasoning models, is in the Appendix.

Model	**Prompt**	RGPow	SqGrd	QSpin	CFTOp	3DBF	GSDGen	WrdH	Ferm	DynCon	3DEF
DeepSeek R1	Hint	**25.0**	5.0	0.0	0.0	5.0	0.0	55.0	0.0	0.0	5.0
	No Comp	15.0	5.0	5.0	0.0	5.0	0.0	40.0	5.0	0.0	0.0
	Zeroshot	10.0	0.0	10.0	5.0	10.0	0.0	30.0	5.0	0.0	5.0
Claude 3.7 Sonnet (Thinking)	Hint	5.0	30.0	35.0	30.0	40.0	0.0	65.0	20.0	5.0	50.0
	No Comp	10.0	25.0	30.0	20.0	45.0	0.0	25.0	10.0	0.0	30.0
	Zeroshot	10.0	45.0	35.0	**35.0**	35.0	0.0	30.0	25.0	5.0	30.0
O4-Mini-High	Hint	5.0	20.0	45.0	25.0	35.0	15.0	70.0	**50.0**	0.0	40.0
	No Comp	**25.0**	10.0	45.0	20.0	45.0	20.0	50.0	35.0	5.0	50.0
	Zeroshot	15.0	15.0	35.0	15.0	45.0	20.0	80.0	15.0	10.0	65.0
Gemini 2.5 Pro (Preview)	Hint	10.0	**50.0**	65.0	25.0	**50.0**	5.0	**100.0**	30.0	25.0	65.0
	No Comp	20.0	40.0	65.0	25.0	**50.0**	**25.0**	**100.0**	25.0	20.0	**70.0**
	Zeroshot	10.0	35.0	**70.0**	25.0	40.0	15.0	95.0	25.0	**30.0**	50.0

Model	**Prompt**	DimLS	GpEdg	GSDeg	Solub	2DEF	2DEFL	OthGrd	2DBF	InfRes	AVG
DeepSeek R1	Hint	15.0	5.0	30.0	0.0	10.0	5.0	10.0	10.0	40.0	11.6
	No Comp	0.0	0.0	20.0	0.0	5.0	0.0	15.0	5.0	70.0	10.0
	Zeroshot	5.0	10.0	30.0	0.0	5.0	0.0	0.0	0.0	45.0	8.9
Claude 3.7 Sonnet (Thinking)	Hint	45.0	5.0	15.0	60.0	40.0	15.0	35.0	15.0	50.0	29.5
	No Comp	45.0	5.0	20.0	65.0	45.0	15.0	40.0	5.0	35.0	24.7
	Zeroshot	50.0	5.0	30.0	65.0	35.0	15.0	40.0	15.0	35.0	28.4
O4-Mini-High	Hint	**55.0**	10.0	**45.0**	45.0	45.0	45.0	65.0	30.0	35.0	35.8
	No Comp	40.0	5.0	25.0	40.0	35.0	40.0	**75.0**	35.0	15.0	32.4
	Zeroshot	50.0	10.0	15.0	40.0	45.0	40.0	60.0	**65.0**	30.0	35.3
Gemini 2.5 Pro (Preview)	Hint	40.0	10.0	25.0	75.0	**65.0**	**55.0**	30.0	30.0	**85.0**	44.2
	No Comp	**55.0**	**15.0**	30.0	**80.0**	60.0	40.0	35.0	30.0	**85.0**	**45.8**
	Zeroshot	**55.0**	**15.0**	40.0	70.0	**65.0**	**55.0**	30.0	35.0	**85.0**	44.5

Table 2: LLM token usage for reasoning models. The first subtable shows the token usage for the first 10 problem sets, and the second subtable shows the token usage for the remaining 9 problem sets and the overall average (AVG). The full result, including non-reasoning models, is in the Appendix.

Model	**Prompt**	RGPow	SqGrd	QSpin	CFTOp	3DBF	GSDGen	WrdH	Ferm	DynCon	3DEF
DeepSeek R1	Hint	7748.4	6085.5	7632.0	6711.2	11 371.5	12 606.7	10 490.3	9499.3	8516.8	9784.7
	No Comp	6183.9	3830.3	8215.8	7438.9	10 062.0	10 215.3	10 626.1	9688.8	9623.3	9504.5
	Zeroshot	7317.8	3968.2	6483.7	7647.8	11 879.9	12 193.4	10 163.2	9927.0	8581.3	9954.6
Claude 3.7 Sonnet (Thinking)	Hint	17 735.1	15 298.6	18 616.3	18 987.8	19 144.3	20 767.3	20 272.2	21 516.2	13 818.2	19 036.2
	No Comp	16 001.6	15 081.4	17 857.8	17 108.1	20 267.2	20 318.4	22 182.9	20 201.1	15 469.8	18 977.4
	Zeroshot	15 936.0	17 851.0	18 357.8	15 954.6	21 768.8	20 952.5	20 672.1	18 951.2	14 550.9	15 786.8
O4-Mini-High	Hint	5345.0	6086.5	3813.3	1545.5	13 739.4	10 854.6	2566.3	8714.3	4469.3	8173.7
	No Comp	4575.9	3702.1	3336.9	1624.8	15 863.4	9539.9	2130.9	8756.7	4530.0	8640.5
	Zeroshot	6536.9	5555.7	4028.6	2661.5	13 835.6	13 012.4	3477.8	8909.5	5576.7	11 134.5
Gemini 2.5 Pro (Preview)	Hint	20 051.2	21 261.7	22 041.2	19 477.1	26 954.0	26 402.8	15 640.6	26 444.4	22 526.0	21 781.8
	No Comp	17 567.4	18 813.4	21 876.5	17 009.9	26 479.1	24 328.0	16 612.9	23 438.2	22 199.9	22 420.3
	Zeroshot	20 182.7	20 409.2	20 523.5	19 394.4	26 385.5	26 236.2	16 700.3	23 850.2	21 939.2	21 351.4

Model	**Prompt**	DimLS	GpEdg	GSDeg	Solub	2DEF	2DEFL	OthGrd	2DBF	InfRes	Avg
DeepSeek R1	Hint	6586.6	5079.4	11 792.4	2220.4	8191.7	10 516.3	3230.9	10 219.9	10 018.2	8331.7
	No Comp	5322.9	5178.9	10 077.8	2745.5	7489.6	9865.0	2512.0	10 132.3	6919.4	7664.8
	Zeroshot	5014.1	5365.7	11 397.6	2220.3	7621.5	10 446.0	3458.6	11 323.9	8649.0	8084.9
Claude 3.7 Sonnet (Thinking)	Hint	14 052.6	17 293.1	29 709.8	8725.1	16 925.4	25 224.0	12 846.5	19 499.4	16 712.8	18 220.0
	No Comp	14 565.5	19 226.3	22 652.3	11 181.5	17 803.3	24 001.0	13 721.2	18 717.2	14 371.0	17 879.2
	Zeroshot	11 997.8	17 201.9	26 542.0	10 072.3	16 644.4	23 130.3	12 099.9	20 960.1	14 012.4	17 549.6
O4-Mini-High	Hint	3911.7	3958.7	9079.1	1754.2	7551.6	14 260.4	4648.6	12 895.1	9460.3	6990.9
	No Comp	3673.2	3792.0	6693.0	2427.8	6766.7	12 709.1	4341.5	9266.6	9333.4	6405.5
	Zeroshot	4134.4	5193.5	8080.3	1912.1	8318.4	13 527.0	6437.0	12 987.5	16 460.9	7988.4
Gemini 2.5 Pro (Preview)	Hint	12 844.7	18 129.4	24 595.4	8205.2	22 265.6	24 617.9	18 649.2	28 798.5	15 979.4	20 877.1
	No Comp	12 147.8	17 457.0	24 887.4	8073.3	21 379.0	23 503.0	19 128.6	30 710.4	17 622.1	20 297.6
	Zeroshot	13 888.2	18 554.4	25 820.9	8387.2	22 079.7	24 866.8	18 291.4	29 737.3	18 759.5	20 913.5

5.3 LLM Failures in Applying Physical Principles

While many LLMs can state physical principles, they often struggle to apply them correctly or comprehensively, particularly the principle of symmetry. Models frequently fail to identify relevant symmetries or incorrectly assume symmetries that do not exist. This weakness is apparent in problems where symmetry provides a significant shortcut.

Example 2: 2D Electric Field (2DEF)

There is a uniformly charged square plane in space with corners at $(x, y, z) = (\pm 1, \pm 1, 0)$. At which of the following locations is the x-direction electric field strength equal to the y-direction electric field strength ($E_x = E_y$)?
a) $(0, 0, 1)$; e) $(1, 1, 1)$; i) $(-1, -1, 1)$; j) $(0, 0, -1)$; n) $(1, 1, -1)$; r) $(-1, -1, -1)$; v) $(2, 2, 0)$; ... (other options omitted)

For a physicist, the solution is straightforward: by symmetry, any location where $x = y$ will have $E_x = E_y$. This insight immediately identifies the correct answers (a, e, i, j, n, r, v). In contrast, one LLM attempted to solve the problem by setting up complex 2D integrals, ultimately arriving at an incorrect answer. This shows a failure to recognize and leverage the fundamental symmetry of the setup.

The capability to reason with such principles varies across models. To illustrate this, we compare responses from a reasoning and a non-reasoning model on a quantum mechanics problem.

Example 3: Quantum dynamics (DynCon)

Consider a $L = 100$ quantum spin chain prepared as the ground state of $H = -\sum_j X_j - 0.9 \sum_j Z_j Z_{j+1}$. Time-evolve this state under $H(t) = \sum_j Y_j X_{j+1} X_{j+2} Y_{j+3}$ from $t = 0$ to $t = 100$. Which of the following is true in the final state?
a) $\langle Z_{60} \rangle = 0$; b) $\langle Z_{39} Y_{40} \rangle = \langle Y_{90} Z_{91} \rangle$; c) $\langle Z_{39} X_{40} \rangle = \langle X_{61} Z_{62} \rangle$; d) None of the above is true.

In this problem, choices (a), (b), and (c) are all correct due to spin-flip, time-reversal, and reflection symmetries, respectively. The reasoning model correctly identified spin-flip and reflection symmetries but failed to apply time-reversal symmetry. The non-reasoning model also mentioned spin-flip and reflection but showed a superficial grasp by failing to see that choice (c) follows from them. When prompted with a hint, it incorrectly invoked translational symmetry (which is absent) and, like the reasoning model, showed no awareness of time-reversal symmetry (see Appendix A.2 for details).

Overall, while reasoning models are more effective at applying physical principles, they are still imperfect. Non-reasoning models demonstrate a shallower understanding, often using terminology without true comprehension.

6 Conclusion

We introduce *PhySense*, a comprehensive, novel, human-curated principle-based physics reasoning benchmark for evaluating large language models on scientific problem-solving across diverse physics domains. *PhySense* comprises 380 carefully designed problems spanning symmetry reasoning, dimensional analysis, renormalization-group analysis, topology, quantum dynamics, and more, together with three prompting strategies ("Zero shot", "Hint", and "No-computation"). Our extensive evaluation of seven state-of-the-art LLMs, including reasoning and non-reasoning models, reveals that while reasoning-focused LLMs outperform their non-reasoning counterparts, all models remain substantially below expert human performance. We observe consistent deficits in token efficiency, principled application of physical laws, and generalization across topics. Moreover, auxiliary prompting strategies (e.g., hints or "no-computation" directives) yield only marginal improvements, indicating the need for deeper integration of principle-based thinking to LLMs. For future directions, it will be important to try improving LLM's principle-based reasoning via supervised fine tuning or reinforcement learning. Our study provides valuable insights and guidance for developing LLMs with efficient, robust and interpretable principle-based reasoning, which are crucial for scientific collaborations and discoveries.

References

[1] Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery, 2024. URL `https://arxiv.org/abs/2406.10833`.

[2] B. Romera-Paredes, M. Barekatain, A. Novikov, and et al. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, Jan 2024. doi: 10.1038/s41586-023-06924-6. URL `https://doi.org/10.1038/s41586-023-06924-6`.

[3] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL `https://arxiv.org/abs/2408.06292`.

[4] Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. Llm-sr: Scientific equation discovery via programming with large language models, 2025. URL `https://arxiv.org/abs/2404.18400`.

[5] Haining Pan, Nayantara Mudur, William Taranto, Maria Tikhanovskaya, Subhashini Venugopalan, Yasaman Bahri, Michael P. Brenner, and Eun-Ah Kim. Quantum many-body physics calculations with large language models. *Communications Physics*, 8(1):49, December 2025. doi: 10.1038/s42005-025-01956-y. URL `https://doi.org/10.1038/s42005-025-01956-y`.

[6] Zhilong Song, Minggang Ju, Chunjin Ren, Qiang Li, Chongyi Li, Qionghua Zhou, and Jinlan Wang. Llm-feynman: Leveraging large language models for universal scientific formula and theory discovery. *arXiv preprint arXiv:2503.06512*, 2025.

[7] Kristian G Barman, Sascha Caron, Emily Sullivan, Henk W de Regt, Roberto Ruiz de Austri, Mieke Boon, Michael Färber, Stefan Fröse, Faegheh Hasibi, Andreas Ipp, et al. Large physics models: Towards a collaborative approach with large language models and foundation models. *arXiv preprint arXiv:2501.05382*, 2025.

[8] Yinggan Xu, Hana Kimlee, Yijia Xiao, and Di Luo. Advancing ai-scientist understanding: Making llm think like a physicist with interpretable reasoning. *arXiv preprint arXiv:2504.01911*, 2025.

[9] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michaely, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini robotics: Bringing ai into the physical world, 2025. URL `https://arxiv.org/abs/2503.20020`.

[10] Daniel JH Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. Theoretical physics benchmark (tpbench)–a dataset and study of ai reasoning capabilities in theoretical physics. *arXiv preprint arXiv:2502.15815*, 2025.

[11] Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, et al. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*, 2025.

[12] Jill Larkin, John McDermott, Dorothea P Simon, and Herbert A Simon. Expert and novice performance in solving physics problems. *Science*, 208(4450):1335–1342, 1980. ISSN 0036-8075.

[13] Michelene T. H. Chi, Paul J. Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2):121–152, apr 1981. doi: 10.1207/s15516709cog0502_2.

[14] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, apr 1988. doi: 10.1207/s15516709cog1202_4.

[15] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[17] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.

[18] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35: 2507–2521, 2022.

[19] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

[20] Daman Arora, Himanshu Singh, and Mausam. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.468. URL `https://aclanthology.org/2023.emnlp-main.468/`.

[21] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

[22] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[23] Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061, 2024.

[24] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*, 2023.

[25] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[26] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

[27] Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, et al. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *Advances in Neural Information Processing Systems*, 37:19209–19253, 2024.

[28] Jingzhe Ding, Yan Cen, and Xinyuan Wei. Using large language model to solve and explain physics word problems approaching human level. *arXiv preprint arXiv:2309.08182*, 2023.

[29] Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. *arXiv preprint arXiv:2502.00334*, 2025.

[30] Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, Haoxu Zhang, Yi Hu, et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. *arXiv preprint arXiv:2504.16074*, 2025.

[31] Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning. *arXiv preprint arXiv:2502.12054*, 2025.

[32] Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Raccuglia, et al. Curie: Evaluating llms on multitask scientific long context understanding and reasoning. *arXiv preprint arXiv:2503.13517*, 2025.

[33] Avinash Anand, Janak Kapuriya, Apoorv Singh, Jay Saraf, Naman Lal, Astha Verma, Rushali Gupta, and Rajiv Shah. Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 53–64. Springer, 2024.

[34] Nayantara Mudur, Hao Cui, Subhashini Venugopalan, Paul Raccuglia, Michael P Brenner, and Peter Norgaard. Feabench: Evaluating language models on multiphysics reasoning ability. *arXiv preprint arXiv:2504.06260*, 2025.

[35] OpenAI. Learning to reason with llms, September 2024. URL `https://openai.com/index/learning-to-reason-with-llms/`. Accessed: 2025-05-12.

[36] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[37] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[38] Anthropic. Claude 3.7 sonnet and extended thinking mode. `https://www.anthropic.com/news/claude-3-7-sonnet`, February 2025. Accessed: 2025-05-10.

[39] Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, November 2024. URL `https://qwenlm.github.io/blog/qwq-32b-preview/`. Accessed: 2025-05-12.

[40] OpenAI. Openai o3 and o4-mini system card. https://openai.com/index/o3-o4-mini-system-card/, April 2025. Accessed: 2025-05-10.

[41] Google DeepMind, March 2025. URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/.

[42] OpenAI. Introducing gpt-4.1 in the api, April 2025. URL https://openai.com/index/gpt-4-1/.

[43] Anthropic. Claude 3.7 sonnet and claude code, February 2025. URL https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-05-12.

[44] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

A Analysis of Several Examples

A.1 An example of LLM failing to apply principles

For some problems, even if the LLM is forced to use principle, it fails to figure out the correct way to use it. In the following example, Gemini-2.5 Pro fails to find the correct symmetry of the system. In general, the LLMs we test have better performance with symmetry group of a square lattice, but for other cases like triangular or honeycomb lattices, the LLMs have difficulty finding out the symmetry group to consider (see below).

Example 4: 2D Electric Field Lattice (2DEFL)

There are point charges on the infinite x-y plane in space. The coordinates of the charges are $\{(x, y, z) = m \cdot (\sqrt{3}/2, -1/2, 0) + n \cdot (0, 1, 0) | (m + n) \not\equiv 0 \mod 3, m, n \in Z\}$. i.e. forming a 2D infinite honeycomb lattice on the x-y plane. Each point has charge of equal amplitude. The charges with n being even integers are positive, and the charges with n being odd integers are negative. Which of the following locations in space have the z-direction electric field strength E_z being zero?
a) $(x = 0, y = 0, z = 1)$; b) $(x = 0, y = 1/2, z = 1)$; c) $(x = 0, y = -1/2, z = 1)$; d) $(x = 0, y = 1, z = 1)$; e) $(x = 0, y = -1, z = 1)$; f) $(x = \sqrt{3}/2, y = 0, z = 1)$; g) $(x = \sqrt{3}/2, y = 1/2, z = 1)$; h) $(x = \sqrt{3}/2, y = -1/2, z = 1)$; i) $(x = -\sqrt{3}/2, y = 0, z = 1)$; j) $(x = -\sqrt{3}/2, y = 1/2, z = 1)$; k) $(x = -\sqrt{3}/2, y = -1/2, z = 1)$; l) $(x = \sqrt{3}/4, y = 1/4, z = 1)$; m) $(x = \sqrt{3}/4, y = -1/4, z = 1)$; n) $(x = \sqrt{3}/4, y = 3/4, z = 1)$; o) $(x = \sqrt{3}/4, y = -3/4, z = 1)$; p) $(x = -\sqrt{3}/4, y = 1/4, z = 1)$; q) $(x = -\sqrt{3}/4, y = -1/4, z = 1)$; r) $(x = -\sqrt{3}/4, y = 3/4, z = 1)$; s) $(x = -\sqrt{3}/4, y = -3/4, z = 1)$;

Answer 2: "Hint" prompt

This problem requires a detailed analysis of the symmetries of the charge distribution. Given the complexity and the lack of an obvious symmetry that leads to $E_z = 0$ for the listed points without extensive calculation, and the peculiarity that some listed solutions are locations of charges themselves, it's possible that the problem requires advanced techniques or there is a specific property of this honeycomb lattice with alternating charges that is not immediately apparent ...... Due to the difficulty and potential complexity, i am unable to provide a justified answer. However, if this is from a known source, one might find existing literature on a similar problem. The final answer is $\boxed{b, e, g, i, l, o, p, r}$.

The LLM does not even specify any symmetry to consider and the answer above is wrong. In fact, the correct use of symmetry is: $E_z = 0$ can be fixed by a reflection + charge-conjugation combined symmetry, where the reflection plane includes a z-direction straight line, or by a rotation + charge-conjugation combined symmetry, where the rotation axis is a z-direction straight line. The charge configuration has such rotation + charge-conjugation combined symmetry at the middle points of the edges that connect a pair of positive and negative charges. Among the choices, those points are (f,i,o,r), which is our answer.

A.2 Answers of Example 3

Answer 3: o4-mini-high, "Zero-shot" prompt

Solution Outline 1. Symmetry 1: Global spin - flip in the X - basis
Define $\mathrm{Q} = \Pi_1{}^{00}\mathrm{X_j}$. $-QX_jQ = X_j, QZ_jQ = -Z_j, QY_jQ = -Y_j$ - Both $\mathrm{H_0} = -\Sigma \mathrm{X_j} - 0.9\Sigma \mathrm{Z_j Z_{j+1}}$ and $\mathrm{H_1} = \Sigma \mathrm{Y_j X_{j+1} X_{j+2} Y_{j+3}}$ commute with Q $\Rightarrow$ The state at all times is Q-symmetric $\Rightarrow \langle Z_j \rangle = 0 \Rightarrow$ statement (a) is true. 2. Symmetry 2: Reflection about the chain center
Define $\mathrm{R} : \mathrm{j} \rightarrow 101 - \mathrm{j}$. One checks $\mathrm{RH_0R} = \mathrm{H_0}$ and $\mathrm{RH_1R} = \mathrm{H_1}$. Thus the time - evolved state is R - symmetric. Under R: $Z_{39}X_{40} \rightarrow Z_{62}X_{61} = X_{61}Z_{62}$ (different - site Pauli commute) $\Rightarrow \langle Z_{39}X_{40} \rangle = \langle X_{61}Z_{62} \rangle \Rightarrow$ statement (c) is true. 3. Translation is not an exact symmetry of the open - chain $\mathrm{H_0}$ or $\mathrm{H_1}$, so (b) need not hold.

Final answer: $\boxed{\text{a, c}}$

491

Answer 4: o4-mini-high, "Hint" prompt

Solution Outline: 1. Symmetry of the initial Hamiltonian H_0 : - $\mathrm{H}_0 = -\Sigma_1{}^{00}\mathrm{X\_j} - 0.9\Sigma_1{}^{99}\mathrm{Z\_jZ\_\{j+1\}}$ commutes with the global spin - flip P= $\prod$ X_j. - Hence its (unique) ground state $|\psi_0\rangle$ is a P - eigenstate and $\langle \mathrm{Z\_j}\rangle_0 = 0$ for all j . 2. Time - evolution under $\mathrm{H(t)} = \Sigma_1{}^{97}\mathrm{Y\_jX\_\{j+1\}X\_\{j+2\}Y\_\{j+3\}}$: - H(t) also commutes with P $\Rightarrow \langle \mathrm{Z\_j}\rangle\mathrm{\_\ t}$ remains zero $\Rightarrow$ statement (a) is true. 3. Reflection symmetry $R(\mathrm{j} \to 101 - \mathrm{j})$: - Both H_0 and $\mathrm{H(t)}$ are invariant under R and $|\psi_0\rangle$ can be chosen R - symmetric. - Hence $\langle A_i B_{\mathrm{i}+1}\rangle_t = \langle RA_iB_{\mathrm{i}+1}R\rangle_t = \langle A_{101-\mathrm{i}}B_{100-\mathrm{i}}\rangle_t$. - For (c): i=39 gives (101-39,100-39)=(62,61), so $\langle Z_{39}X_{40}\rangle = \langle Z_{62}X_{61}\rangle = \langle X_{61}Z_{62}\rangle$ (operators on different sites commute) $\Rightarrow$ (c) is true. 4. Statement (b) involves $(39, 40)$ vs $(90, 91)$, which are neither related by P nor by $R \Rightarrow$ false.
Final Answer: $\boxed{a, c}$

492

Answer 5: o4-mini-high, "No Comp" prompt

Step-by-step: 1. Define the global "spin - flip" operator

$$G = \prod_{-}\{\mathrm{j} = 1\}\hat{}\{100\}X\_\mathrm{j}.$$

Note $GZ_jG = -Z_j$ and $GX_jG = +X_j$, so G flips the sign of Z and Y and leaves X invariant. 2. Check symmetries: - $H_{initial} = -\Sigma X\_\mathrm{j} - 0.9\Sigma Z\_\mathrm{j}Z\_\{\mathrm{j}+1\}$ commutes with G (the usual Z_2 symmetry of the transverse - field Ising model). - $H_{drive} = \sum\{j = 1\}\hat{}\{97\}Y\_jX\{j+1\}X\_\{j+2\}Y\_\{j+3\}$ also commutes with G (each term has two Y's). $\Rightarrow$ The time - evolved state remains a +1 eigenstate of G for all t. 3. Consequences for expectation values: - Any operator O with GOG $= -$O has $\langle \mathrm{O}\rangle = 0$ at all times.
In particular Z_j is G-odd $\Rightarrow \langle Z\_j\rangle = 0 \Rightarrow (a)$ is true. $-Z\_jX\_\{j+1\}$ is G-odd (one Z) $\Rightarrow \langle Z\_jX\_\{j+1\}\rangle = 0$, and likewise $\langle X\_jZ\_\{j+1\}\rangle = 0 \Rightarrow (c)$ holds $(0 = 0)$.
- $Z\_jY\_\{j+1\}$ and $Y\_jZ\_\{j+1\}$ are each G-even, but there is no symmetry forcing $\langle Z\_jY\_\{j+1\}\rangle = \langle Y\_kZ\_\{k+1\}\rangle$ for different string orderings $\Rightarrow$ (b) is not guaranteed.
Answer: $\boxed{a, c}$

493

Answer 6: GPT-4.1, "Zero-shot" prompt

Let's analyze each statement:
Hamiltonians: - Initial: $H_0 = -\sum_{j=1}^{100} X_j - 0.9\sum_{j=1}^{99} Z_jZ_{j+1}$ (Transverse field Ising). - Time-evolution: $H(t) = \sum_{j=1}^{97} Y_jX_{j+1}X_{j+2}Y_{j+3}$
Statement a: $\langle Z_{60}\rangle = 0$ of the initial Hamiltonian has $\langle Z_j\rangle = 0$ due to the symmetry. Time evolution with a symmetric Hamiltonian preserves this. So, this statement IS true.
Statement b: $\langle Z_{39}Y_{40}\rangle = \langle Y_{90}Z_{91}\rangle$
Statement c: $\langle Z_{39}X_{40}\rangle = \langle X_{61}Z_{62}\rangle$ Similarly, unless there is special symmetry (e.g., reflection through the center of the chain), these correlators will NOT be equal. Sites (39,40) and (61,62) are not mirror images either, so this is also FALSE. Statement d: None of above is true. Since (a) is true, this is FALSE.
Final Answer: $\boxed{a}$

494

Answer 7: GPT-4.1, "Hint" prompt

Let's analyze the situation step-by-step using symmetries: 1. Initial Hamiltonian:

$$H = -\sum_{j=1}^{100} X_j - 0.9\sum_{j=1}^{99} Z_jZ_{j+1}$$

The ground state of this Hamiltonian is unique and respects all the symmetries of the Hamiltonian (e.g., parity symmetry). 2. Time-evolution Hamiltonian:

495

$$H(t) = \sum_{j=1}^{97} Y_j X_{j+1} X_{j+2} Y_{j+3}$$

This Hamiltonian is uniform and translation-invariant by 1 site. Let's check for parity symmetries and other symmetries. a) $\langle Z_{60} \rangle = 0$
Let's check if the evolved state is invariant under some symmetry that would force this expectation value to vanish. - The time evolution Hamiltonian is even in Y and X but is odd under $Z \to -Z$ (since it contains only Y and X operators). - The initial state is unique and symmetric; under the global spin-flip $Z \to -Z$, the initial state's expectation of any single Z_j vanishes, i.e., $\langle Z_j \rangle = 0$. - The time-evolution Hamiltonian does not break this symmetry, so for all t, $\langle Z_j \rangle = 0$.
Thus, statement a) is TRUE. b) $\langle Z_{39} Y_{40} \rangle = \langle Y_{90} Z_{91} \rangle$ - The time-evolution Hamiltonian is translation-invariant. - The initial state is translation-invariant. - However, after time evolution, the system is still translation-invariant, so for any displacement k, $\langle Z_j Y_{j+1} \rangle$ should be the same for all valid j. - However, notice the indices: $(39, 40)$ and $(90, 91)$. Both are in the bulk (well away from the boundaries), where translation invariance definitely holds.
So, statement b) is TRUE. c) $\langle Z_{39} X_{40} \rangle = \langle X_{61} Z_{62} \rangle$ - Similarly, translation invariance applies. - However, under certain reflection symmetries or if the operator is odd under any symmetry not respected by the Hamiltonian, the mean value could vanish or be related differently. - But since both correlators are related by exactly the same relative placement, and the evolution/initial state are symmetric under translation, these expectation values should be equal.
Thus, statement c) is TRUE. d) None of above is true.
At least one statement is true, so d) is FALSE. Final Answer $\boxed{a, b, c}$

Answer 8: GPT-4.1, "No comp" prompt

Let's analyze each statement step-by-step:
Hamiltonians: - Ground state prepared with: $H = -\sum_{j=1}^{100} X_j - 0.9 \sum_{j=1}^{99} Z_j Z_{j+1}$ - Time evolution under: $H(t) = \sum_{j=1}^{97} Y_j X_{j+1} X_{j+2} Y_{j+3}$ - Time range: $t = 0$ to $t = 100$, so this is a long time evolution, likely causing full thermalization or strong scrambling. (a) $\langle Z_{60} \rangle = 0$? - The evolution Hamiltonian does not conserve Z_{60}. - Quantum thermalization in such a system would drive a single-site expectation like $\langle Z_{60} \rangle$ to zero in the absence of conserved quantities or symmetries. - The initial state is even under spin flip, the evolution is also even under global Y flip, but there is no protection for Z-no field acts to bias it.
So (a) is likely true.
(b) $\langle Z_{39} Y_{40} \rangle = \langle Y_{90} Z_{91} \rangle$? this is crucial). - The Hamiltonians are translationally invariant in the bulk, but 39-40 and 90-91 are separated by 51 sites, so are not related by any symmetry. - No argument for equality of those two-point functions by symmetry.
(b) is not necessarily true.
(c) $\langle Z_{39} X_{40} \rangle = \langle X_{61} Z_{62} \rangle$?
Look at the operators: - Sites $(39, 40)$ and $(61, 62)$ are separated by 22 sites. Again, in an open chain, these are not related by any symmetry. - No evident symmetry or conservation relates these two correlators.
(c) is not necessarily true.
(d) None of the above is true?
Since (a) is true (generic local observable under thermalizing evolution approaches zero expectation in the absence of conservation laws), (d) is not true.
Final Answer: $\boxed{a}$

B Full Experiment Results

Table 3: LLM accuracy scores (as percentages). The first table shows accuracy for the first 10 problem sets, and the second table shows accuracy for the remaining 9 problem sets and the overall average (AVG).

Model	**Prompt**	RGPow	SqGrd	QSpin	CFTOp	3DBF	GSDGen	WrdH	Ferm	DynCon	3DEF
Claude 3.7 Sonnet	Hint	15.0	10.0	15.0	20.0	10.0	0.0	10.0	10.0	5.0	30.0
	No Comp	10.0	5.0	10.0	10.0	15.0	0.0	20.0	25.0	5.0	20.0
	Zeroshot	15.0	5.0	5.0	15.0	15.0	0.0	0.0	15.0	10.0	25.0
DeepSeek R1	Hint	**25.0**	5.0	0.0	0.0	5.0	0.0	55.0	0.0	0.0	5.0
	No Comp	15.0	5.0	5.0	0.0	5.0	0.0	40.0	5.0	0.0	0.0
	Zeroshot	10.0	0.0	10.0	5.0	10.0	0.0	30.0	5.0	0.0	5.0
Claude 3.7 Sonnet (Thinking)	Hint	5.0	30.0	35.0	30.0	40.0	0.0	65.0	20.0	5.0	50.0
	No Comp	10.0	25.0	30.0	20.0	45.0	0.0	25.0	10.0	0.0	30.0
	Zeroshot	10.0	45.0	35.0	**35.0**	35.0	0.0	30.0	25.0	5.0	30.0
DeepSeek Chat V3	Hint	10.0	0.0	30.0	5.0	15.0	0.0	15.0	15.0	15.0	10.0
	No Comp	20.0	0.0	25.0	20.0	15.0	0.0	35.0	25.0	15.0	10.0
	Zeroshot	10.0	0.0	30.0	15.0	15.0	0.0	10.0	25.0	20.0	20.0
GPT-4.1	Hint	10.0	10.0	15.0	25.0	30.0	5.0	75.0	25.0	10.0	20.0
	No Comp	0.0	0.0	25.0	20.0	20.0	0.0	30.0	25.0	10.0	35.0
	Zeroshot	0.0	0.0	10.0	10.0	30.0	0.0	55.0	**30.0**	10.0	25.0
O4-Mini-High	Hint	5.0	20.0	45.0	25.0	35.0	15.0	70.0	50.0	0.0	40.0
	No Comp	**25.0**	10.0	45.0	20.0	45.0	20.0	50.0	35.0	5.0	50.0
	Zeroshot	15.0	15.0	35.0	15.0	45.0	20.0	80.0	15.0	10.0	65.0
Gemini 2.5 Pro (Preview)	Hint	10.0	**50.0**	65.0	25.0	**50.0**	5.0	**100.0**	**30.0**	25.0	65.0
	No Comp	20.0	40.0	65.0	25.0	**50.0**	**25.0**	**100.0**	25.0	20.0	**70.0**
	Zeroshot	10.0	35.0	**70.0**	25.0	40.0	15.0	95.0	25.0	**30.0**	50.0

Model	**Prompt**	DimLS	GpEdg	GSDeg	Solub	2DEF	2DEFL	OthGrd	2DBF	InfRes	AVG
Claude 3.7 Sonnet	Hint	15.0	10.0	25.0	70.0	10.0	5.0	55.0	5.0	25.0	18.2
	No Comp	25.0	10.0	25.0	50.0	15.0	10.0	40.0	5.0	20.0	16.8
	Zeroshot	25.0	0.0	20.0	55.0	15.0	10.0	50.0	0.0	25.0	16.1
DeepSeek R1	Hint	15.0	5.0	30.0	0.0	10.0	5.0	10.0	10.0	40.0	11.6
	No Comp	0.0	0.0	20.0	0.0	5.0	0.0	15.0	5.0	70.0	10.0
	Zeroshot	5.0	10.0	30.0	0.0	5.0	0.0	0.0	0.0	45.0	8.9
Claude 3.7 Sonnet (Thinking)	Hint	45.0	5.0	15.0	60.0	40.0	15.0	35.0	15.0	50.0	29.5
	No Comp	45.0	5.0	20.0	65.0	45.0	15.0	40.0	5.0	35.0	24.7
	Zeroshot	50.0	5.0	30.0	65.0	35.0	15.0	40.0	15.0	35.0	28.4
DeepSeek Chat V3	Hint	25.0	10.0	10.0	50.0	0.0	10.0	45.0	0.0	50.0	16.6
	No Comp	15.0	5.0	15.0	40.0	15.0	0.0	25.0	5.0	35.0	16.8
	Zeroshot	30.0	10.0	20.0	25.0	5.0	0.0	35.0	5.0	35.0	16.3
GPT-4.1	Hint	35.0	10.0	20.0	45.0	10.0	10.0	50.0	0.0	30.0	22.9
	No Comp	30.0	5.0	10.0	40.0	10.0	10.0	50.0	5.0	30.0	18.7
	Zeroshot	35.0	0.0	10.0	35.0	15.0	5.0	25.0	0.0	10.0	16.1
O4-Mini-High	Hint	**55.0**	10.0	**45.0**	45.0	45.0	45.0	65.0	30.0	35.0	35.8
	No Comp	40.0	5.0	25.0	40.0	35.0	40.0	**75.0**	35.0	15.0	32.4
	Zeroshot	50.0	10.0	15.0	40.0	45.0	40.0	60.0	**65.0**	30.0	35.3
Gemini 2.5 Pro (Preview)	Hint	40.0	10.0	25.0	75.0	**65.0**	**55.0**	30.0	30.0	**85.0**	44.2
	No Comp	**55.0**	**15.0**	30.0	**80.0**	60.0	40.0	35.0	30.0	**85.0**	**45.8**
	Zeroshot	**55.0**	**15.0**	40.0	70.0	**65.0**	**55.0**	30.0	35.0	**85.0**	44.5

Table 4: LLM performance scores across problem sets (H: Hint, N: No Comp, Z: Zeroshot). Table (a) shows results for the first 10 problem sets, and Table (b) shows results for the remaining 9 problem sets and the overall average score for each model configuration.

Model	Prompt	RGPow	SqGrd	QSpin	CFTOp	3DBF	GSDGen	WrdH	Ferm	DynCon	3DEF
Claude 3.7 Sonnet	Hint	512.8	349.1	468.8	343.6	392.3	354.3	768.6	471.1	454.9	447.9
	No Comp	469.6	403.3	474.3	401.3	445.7	398.5	724.9	507.7	530.0	485.2
	Zeroshot	400.7	382.7	462.5	341.8	405.9	376.0	718.5	459.9	510.3	469.6
DeepSeek R1	Hint	7748.4	6085.5	7632.0	6711.2	11 371.5	12 606.7	10 490.3	9499.3	8516.8	9784.7
	No Comp	6183.9	3830.3	8215.8	7438.9	10 062.0	10 215.3	10 626.1	9688.8	9623.3	9504.5
	Zeroshot	7317.8	3968.2	6483.7	7647.8	11 879.9	12 193.4	10 163.2	9927.0	8581.3	9954.6
Claude 3.7 Sonnet (Thinking)	Hint	17 735.1	15 298.6	18 616.3	18 987.8	19 144.3	20 767.3	20 272.2	21 516.2	13 818.2	19 036.2
	No Comp	16 001.6	15 081.4	17 857.8	17 108.1	20 267.2	20 318.4	22 182.9	20 201.1	15 469.8	18 977.4
	Zeroshot	15 936.0	17 851.0	18 357.8	15 954.6	21 768.8	20 952.5	20 672.1	18 951.2	14 550.9	15 786.8
DeepSeek Chat V3	Hint	665.1	475.6	623.8	382.8	1064.0	665.0	1274.8	717.8	1048.8	1144.8
	No Comp	623.1	475.8	562.2	407.7	930.8	756.7	1325.4	668.3	1008.4	1437.7
	Zeroshot	627.5	670.2	716.3	411.3	850.1	604.2	1291.4	742.2	823.2	1489.7
GPT-4.1	Hint	633.6	678.4	671.2	480.4	869.6	1091.4	1300.4	715.6	709.1	1112.2
	No Comp	674.9	733.0	725.9	587.4	1058.3	1024.2	1188.9	839.8	913.5	1278.7
	Zeroshot	516.8	577.8	541.9	466.5	878.4	914.3	1234.8	684.2	678.1	1044.8
O4-Mini-High	Hint	5345.0	6086.5	3813.3	1545.5	13 739.4	10 854.6	2566.3	8714.3	4469.3	8173.7
	No Comp	4575.9	3702.1	3336.9	1624.8	15 863.4	9539.9	2130.9	8756.7	4530.0	8640.5
	Zeroshot	6536.9	5555.7	4028.6	2661.5	13 835.6	13 012.4	3477.8	8909.5	5576.7	11 134.5
Gemini 2.5 Pro (Preview)	Hint	20 051.2	21 261.7	22 041.2	19 477.1	26 954.0	26 402.8	15 640.6	26 444.4	22 526.0	21 781.8
	No Comp	17 567.4	18 813.4	21 876.5	17 009.9	26 479.1	24 328.0	16 612.9	23 438.2	22 199.9	22 420.3
	Zeroshot	20 182.7	20 409.2	20 523.5	19 394.4	26 385.5	26 236.2	16 700.3	23 850.2	21 939.2	21 351.4

Model	Prompt	DimLS	GpEdg	GSDeg	Solub	2DEF	2DEFL	OthGrd	2DBF	InfRes	Avg
Claude 3.7 Sonnet	Hint	528.8	435.2	404.1	334.6	374.5	368.1	351.8	404.5	340.9	426.6
	No Comp	617.8	462.3	397.9	392.2	427.1	422.4	422.5	403.8	357.7	460.2
	Zeroshot	509.7	426.3	351.1	316.9	400.8	375.9	363.9	382.3	322.7	419.9
DeepSeek R1	Hint	6586.6	5079.4	11 792.4	2220.4	8191.7	10 516.3	3230.9	10 219.9	10 018.2	8331.7
	No Comp	5322.9	5178.9	10 077.8	2745.5	7489.6	9865.0	2512.0	10 132.3	6919.4	7664.8
	Zeroshot	5014.1	5365.7	11 397.6	2220.3	7621.5	10 446.0	3458.6	11 323.9	8649.0	8084.9
Claude 3.7 Sonnet (Thinking)	Hint	14 052.6	17 293.1	29 709.8	8725.1	16 925.4	25 224.0	12 846.5	19 499.4	16 712.8	18 220.0
	No Comp	14 565.5	19 226.3	22 652.3	11 181.5	17 803.3	24 001.0	13 721.2	18 717.2	14 371.0	17 879.2
	Zeroshot	11 997.8	17 201.9	26 542.0	10 072.3	16 644.4	23 130.3	12 099.9	20 960.1	14 012.4	17 549.6
DeepSeek Chat V3	Hint	1140.4	482.6	928.0	369.7	975.3	832.9	625.2	866.8	863.9	797.2
	No Comp	955.9	433.8	1094.7	337.0	714.1	744.6	541.3	827.2	770.2	769.2
	Zeroshot	987.7	456.4	729.3	292.7	839.4	831.5	597.1	839.3	424.4	748.6
GPT-4.1	Hint	1340.6	467.9	1269.8	262.1	692.8	941.3	548.5	677.0	348.3	779.5
	No Comp	1378.8	646.7	880.8	305.6	831.0	1144.4	594.1	833.7	438.7	846.2
	Zeroshot	1209.9	482.3	686.0	193.0	643.8	922.7	432.8	752.0	352.8	695.4
O4-Mini-High	Hint	3911.7	3958.7	9079.1	1754.2	7551.6	14 260.4	4648.6	12 895.1	9460.3	6990.9
	No Comp	3673.2	3792.0	6693.0	2427.8	6766.7	12 709.1	4341.5	9266.6	9333.4	6405.5
	Zeroshot	4134.4	5193.5	8080.3	1912.1	8318.4	13 527.0	6437.0	12 987.5	16 460.9	7988.4
Gemini 2.5 Pro (Preview)	Hint	12 844.7	18 129.4	24 595.4	8205.2	22 265.6	24 617.9	18 649.2	28 798.5	15 979.4	20 877.1
	No Comp	12 147.8	17 457.0	24 887.4	8073.3	21 379.0	23 503.0	19 128.6	30 710.4	17 622.1	20 297.6
	Zeroshot	13 888.2	18 554.4	25 820.9	8387.2	22 079.7	24 866.8	18 291.4	29 737.3	18 759.5	20 913.5