Advancing and Evaluating AI Agent for Frontier Physics Research

Tingjia Miao [1 2 3] **Wenkai Jin** [1] **Jinxin Tan** [4 5 6] **Muhua Zhang** [4 5 6] **Xianghe Pang** [1] **Zexi Liu** [1] **Yuwen Du** [1] **Tian Jin** [1] **Tu Guo** [4 5] **Yuelin Hu** [2 4] **Zhengliang Zhang** [2 4 5] **Jingkun Liu** [2] **Jiejun Zhang** [2 4] **Yunjie Huang** [1] **Yinuo Gao** [2 4] **Yuhan Wang** [2 4] **Wenbo Li** [4] **Shuo Chen** [7] **Bingyang Zheng** [1] **Fengyang Li** [1] **Rui Ye** [1] **Yuzhi Zhang** [8] **Linfeng Zhang** [8] **Kun Chen** [7] **Wei Wang** [4 5 6] **Weinan E** [1] **Siheng Chen** [1]

Abstract

Advances of LLMs in reasoning, domain knowledge, and tool use reveal the potential of agentic science, where AI conducts autonomous research and enables AI-driven scientific discovery. Yet such paradigm remains difficult to realize in theoretical and computational physics research, where research workflows require domain knowledge, long-horizon reasoning, and reliable numerical computation. Towards agentic physics research, we present PHYSMASTER, a research agent system integrating adaptive multi-trajectory exploration and trusted external knowledge to conduct reliable ultra-long-horizon exploration. Further, we construct PRL-BENCH, a research-oriented benchmark adapted from 100 *Physical Review Letters* papers. Evaluation shows that even frontier models face frequent conceptual errors, lack of advanced theoretical-physics knowledge, and unstable derivations over extended horizons, while PHYSMASTER substantially mitigates these limitations. Together, PHYSMASTER and PRL-BENCH provide a pioneering paradigm for advancing and evaluating autonomous AI scientists for frontier physics research.

1. Introduction

Artificial intelligence for science is moving toward the new paradigm of *agentic science*, where AI systems are expected not merely to assist isolated subtasks but to conducts end-to-end scientific workflows.

This paradigm shift has motivated series of AI scientist systems, including AI co-scientist, Robin, The AI Scientist, AI Scientist-v2, and Kosmos (Natarajan et al., 2025; Ghareeb et al., 2025; Lu et al., 2024; Yamada et al., 2025; Mitchener et al., 2025). These systems demonstrate the promise of iterative planning, tool use, and multi-agent collaboration, but they remain limited text-centric, well-defined, or short-horizon tasks. They still lack the robustness needed for research workflows and the capabilities to apply specialized knowledge at the right level of abstraction, execute reliable computation, and iteratively revise errors.

Theoretical and computational physics constitute a demanding yet suitable setting for agentic science. Frontier physics research features high knowledge barriers, long-horizon reasoning, stringent correctness requirements, and tight coupling between theoretical derivation and executable computation. These properties make it difficult for standalone models to maintain reliable progress through extended workflows. However, structured agentic design directly targets these bottlenecks: planning can stabilize long-horizon exploration, trusted knowledge can improve theoretical grounding, and tool integration can bring symbolic and numerical computation within the operational scope of the agent. This makes theoretical and computational physics a natural domain for developing autonomous scientific research agents.

Towards agentic physics research, we present **PhysMaster**, a research agent system for frontier theoretical and computational physics. PhysMaster integrates hierarchical agent collaboration, Monte Carlo Tree Search based multi-trajectory exploration, explicit memory over long-horizon search to support reliable ultra-long-horizon exploration via iterative planning, derivation, computation, critique revision.

As a core module of PhysMaster, we introduce **LANDAU**, short for Layered Academic Data Universe, named in honor of the eminent physicist Lev Landau. LANDAU is a **physics research specialized knowledge system** integrating literature, authoritative prior knowledge, research related agent skills, and validated research workflows to enhance the eff-

[1]School of Artificial Intelligence, Shanghai Jiao Tong University [2]Zhiyuan College, Shanghai Jiao Tong University [3]SciLand [4]School of Physics and Astronomy, Shanghai Jiao Tong University [5]Tsung-Dao Lee Institute, Shanghai Jiao Tong University [6]State Key Laboratory of Dark Matter Physics, Shanghai Jiao Tong University [7]Institute of Theoretical Physics, Chinese Academy of Sciences [8]DP Technology. Correspondence to: Siheng Chen <sihengc@sjtu.edu.cn>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

ciency reliability of physics research.

Despite the growing interest in scientific agents, existing evaluations only partially capture the challenges posed by long-horizon research workflows. Benchmarks such as OlympiadBench, OlympicArena, OlymMath, and Humanity's Last Exam substantially raise the level of scientific reasoning and domain knowledge required of models (He et al., 2024; Huang et al., 2025; Sun et al., 2025; Phan et al., 2025). Nevertheless, they remain largely confined to well-defined settings with explicit objectives and relatively clear solution pathways. Physics-specific evaluations such as TPBench, PHYSICS, and PhyBench improve domain alignment, yet most instances remain shorter-horizon than authentic research-style tasks (Chung et al., 2025; Feng et al., 2025; Qiu et al., 2025). FrontierScience and PRBench move closer to research-oriented assessment, but they still leave unresolved how to systematically evaluate whether an agent can plan, explore, revise, and verify over extended scientific workflows (Wang et al., 2026; Qiu et al., 2026).

To address this evaluation gap, we further construct **PRL-Bench**, a research-oriented benchmark adapted from 100 recent *Physical Review Letters* papers. PRL-Bench is designed to preserve key features of real physics research, including research-context-rich problem formulations, non-trivial theoretical targets, open solution paths, long-horizon workflow structure, and objectively verifiable outcomes. Evaluation on PRL-Bench shows that even frontier models face frequent conceptual errors, lack advanced theoretical-physics knowledge, and exhibit unstable derivations over extended horizons, while PhysMaster substantially mitigates these limitations.

We make three contributions in this paper:

- We present **PhysMaster**, a research agent framework for frontier physics that combines hierarchical collaboration, adaptive multi-trajectory exploration, and explicit memory for long-horizon scientific workflows.
- As a core module of **PhysMaster**, we introduce **LANDAU**—the Layered Academic Data Universe, a **physics research specialized knowledge system** integrating library, prior, skill, and workflow to enhance the reliability of physics research exploration.
- We construct **PRL-Bench**, a research-oriented benchmark adapted from 100 *Physical Review Letters* papers, and use it to evaluate both frontier-model baselines and PhysMaster-augmented systems under matched conditions.

Together, **PhysMaster** and **PRL-Bench** provide a concrete paradigm for advancing and evaluating autonomous AI scientists for frontier physics research.

2. Related Work

2.1. Scientific Benchmarks and Research-Oriented Evaluation

Recent scientific benchmarks have substantially increased the difficulty of model evaluation. General expert-level benchmarks, including OlympiadBench, OlympicArena, OlymMath, and Humanity's Last Exam, show that frontier models remain brittle on advanced science and mathematics tasks (He et al., 2024; Huang et al., 2025; Sun et al., 2025; Phan et al., 2025). However, these benchmarks mostly target well-defined problems with explicit objectives and relatively fixed solution structures, and therefore only partially assess whether an AI system can organize extended scientific workflows.

Physics-specific benchmarks further improve domain alignment. TPBench, PHYSICS, and PhyBench evaluate theoretical reasoning and physics problem solving in more specialized settings (Chung et al., 2025; Feng et al., 2025; Qiu et al., 2025), while FrontierScience and PRBench move toward expert-level and research-oriented evaluation (Wang et al., 2026; Qiu et al., 2026). Building on this trajectory, our work uses PRL-Bench as the main empirical testbed for PhysMaster and focuses on the complementary systems question: how a research agent should structure decomposition, exploration, execution, and memory over long-horizon physics workflows.

2.2. AI Scientist Systems and Scientific Agents

Agentic scientific systems have recently expanded from literature assistance to broader research automation. AI co-scientist emphasizes debate, proposal generation, and iterative hypothesis refinement for science-facing workflows (Natarajan et al., 2025). Robin, The AI Scientist, AI Scientist-v2, and Kosmos demonstrate different combinations of multi-agent orchestration, experimental iteration, literature grounding, and autonomous reporting (Ghareeb et al., 2025; Lu et al., 2024; Yamada et al., 2025; Mitchener et al., 2025). Collectively, these systems suggest that scientific competence increasingly depends on orchestration rather than raw next-token prediction alone.

Within physics, existing AI-for-physics efforts have achieved strong results on isolated parts of the research loop, but frontier theoretical and computational physics research demands the joint capabilities of reasoning, numerical computation, domain-specific priors, and long-horizon control. With such aim and inspired by agentic tree-search systems for AI research (Liu et al., 2025), we build PhysMaster, integrating adaptive multi-trajectory exploration strategy and trusted external knowledge to enhance the reliability and efficiency in physics research.

3. PhysMaster

3.1. Hierarchical Agent Collaboration

PhysMaster is organized as a hierarchical research workflow rather than a single-pass language interface. A Clarifier first reformulates the raw physics problem into an explicit task contract, separating the global objective from executable subgoals. A Supervisor then coordinates the overall search process, decides which subtask should be advanced next, and allocates computation across candidate paths. Theoretician workers execute local reasoning, code-based verification, and tool use, while Critic evaluates intermediate outputs and Summarizer distills the best trajectory into a stable final report.

Table 1. Core Agents of PhysMaster

Component	Role
Clarifier	Converts the raw query into a structured contract and optionally attaches a matched workflow template.
Supervisor	Maintains the search tree, schedules subtasks, dispatches node execution, and invokes critic evaluation.
Theoretician	Solves a selected subtask using language reasoning, Python execution, skills, and optional literature search.
Critic	Evaluates candidate outputs, assigns rewards, and determines whether nodes are completed or need further revision.
Summarizer	Distills best trajectory into formal report.

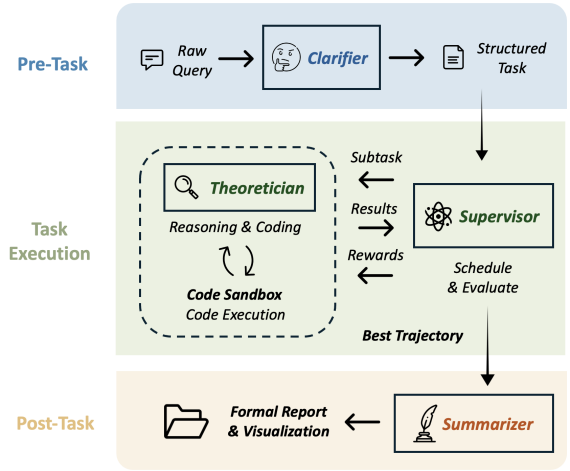

Figure 1. Hierarchical agent collaboration and core workflow of PhysMaster

This division of labor is useful for theoretical and computational physics because intermediate steps are heterogeneous: some require conceptual modeling, some require symbolic derivation, and others require executable numerical checks. A hierarchical agent structure therefore improves controllability and recoverability in long-horizon workflows by separating global coordination from local solving.

3.2. Exploration Strategy and Memory System

Long-horizon physics tasks rarely admit a single obvious route from problem statement to final answer. PhysMaster therefore adopts Monte Carlo Tree Search as an external control structure for multi-path exploration. Instead of committing to one linear chain of reasoning, the system represents each local attempt as a node in a search tree, with `draft` nodes opening new candidate paths and `revise` nodes locally correcting existing ones. Critic-guided rewards are then used to decide which branches should be further expanded.

Node selection follows a UCB-style objective,

$$\mathrm{UCB}(v) = \frac{Q_v}{N_v} + C\sqrt{\frac{\ln(N_{\text{parent}}+1)}{N_v}}, \tag{1}$$

where Q_v is accumulated reward, N_v is the visit count of node v, and C is the exploration constant. The first term encourages continued investment in branches with stronger historical returns, while the second preserves opportunities for less explored branches, preventing the search from collapsing too early onto a locally plausible but globally weak trajectory.

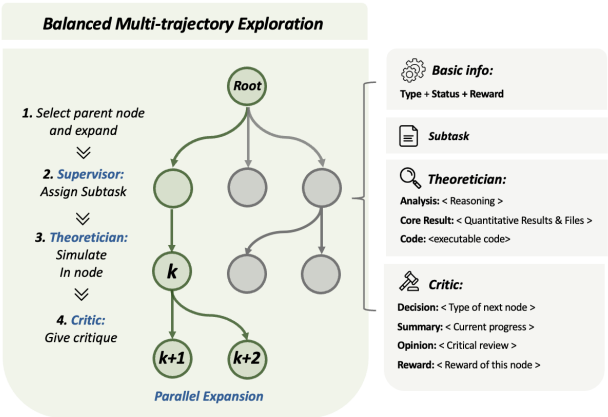

Figure 2. MCTS-based multi-trajectory exploration in PhysMaster

Operationally, one search round contains selection, expansion, critic evaluation, backpropagation, and optional pruning. The Supervisor chooses the most promising leaf node, determines whether the next step should be `draft` or `revise`, and dispatches Theoretician workers to execute local attempts. Critic then returns both a scalar reward and a structured decision, such as `complete`, `to_revise`, or `to_redraft`. Rewards are propagated back to ancestor nodes, so local progress can influence future computation allocation.

This exploration strategy is paired with an explicit memory system over the tree. Following the thesis structure, PhysMaster separates memory into three levels: node-level experience, compressed knowledge, and cross-task wisdom. The purpose of this hierarchy is not to preserve as much history as possible, but to let information at different granularities remain at the level where it is most useful.

Table 2. Memory hierarchy in the tree-based exploration process.

Scope	Description
Per-Node Experience	Full Theoretician output, include reasoning, tool calls, and code. Available to the Critic for evaluation, then compressed.
Compressed Knowledge	Distilled summary attached to each node. Ancestors and siblings share insights through tree context.
Cross-Task Wisdom	Best trajectory distilled post-task and stored for future RAG.

In this way, PhysMaster treats long-horizon scientific problem solving as a managed process of exploration, evaluation, compression, and revision, rather than as a single unbounded prompt trajectory.

3.3. LANDAU: Layered Academic Data Universe

To support reliable research-time reasoning, PhysMaster is further equipped with LANDAU, short for *Layered Academic Data Universe*, named in honor of the eminent physicist Lev Landau. LANDAU is a physics research specialized external knowledge system. Its role is to provide stable support when parameter memory alone is insufficient for frontier theoretical physics. We summarize its composition as

$$\text{LANDAU} = \mathcal{L} \cup \mathcal{P} \cup \mathcal{S} \cup \mathcal{W}, \tag{2}$$

where $\mathcal{L}$ denotes the library layer, $\mathcal{P}$ the prior layer, $\mathcal{S}$ the skill layer, and $\mathcal{W}$ the workflow layer.

Library layer $\mathcal{L}$. This layer provides access to scientific papers and local literature collections. Its main role is to supply research evidence and frontier background that may not be stably covered by pretraining alone.

Prior layer $\mathcal{P}$ This layer stores high-confidence knowledge distilled from textbooks, lecture notes, and curated references. It serves as a stable source of conceptual constraints and theoretical frameworks. A series of typical authoritative textbook in $\mathcal{P}$ are listed in appendix B.

Skill layer $\mathcal{S}$. This layer packages recurring analysis routines into reusable execution units. Its role is to reduce repeated trial-and-error in familiar subtasks and to make domain-specific operating habits available during task execution.

Workflow layer $\mathcal{W}$. This layer provides high-level scaffolds for recurrent task classes. Its role is to help the system begin from a reasonable research structure and reduce early-stage drift in long-horizon tasks.

In this sense, LANDAU is designed as a persistent external knowledge environment for the full workflow. It can support task clarification, local reasoning, and critic evaluation, while remaining distinct from the search-time memory used to retain intermediate trajectory state.

4. Evaluation

4.1. Benchmark Overview

We evaluate PhysMaster on PRL-Bench, a research-oriented benchmark constructed from 100 recent papers published in *Physical Review Letters*. The benchmark covers five theory- and computation-intensive subfields of modern physics: (i) Astronomy & Astrophysics, (ii) Condensed Matter Physics, (iii) High-Energy Physics, (iv) Quantum Information and Foundations,and (v) Statistical & Computational Physics. These source papers are selected from authoritative frontier research and emphasize theoretical derivation and numerical computation, while experimental studies or tasks requiring extraordinary datasets, specialized simulation platforms, or excessive computational resources are excluded.

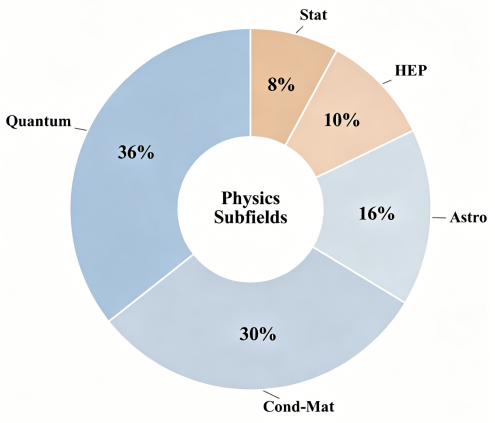

Figure 3. Subfield distribution of PRL-Bench

PRL-Bench is designed to approximate the structure of authentic theoretical and computational physics research while retaining objective evaluability. In collaboration with more than ten domain experts, each paper is converted into a research-oriented task by extracting its scientific background, workflow structure, and core results. The resulting tasks preserve motivation-rich problem statements, nontriv-

Table 3. PRL-Bench performance of standalone frontier models and corresponding PhysMaster-augmented settings. Scores are normalized to a 0–100 scale across five physics subfields and the global benchmark. Relative improvements with PhysMaster-augmentation are computed over the corresponding standalone model.

Setting	Astro	Cond-Mat	HEP	Quantum	Stat	Global
GPT-5.4	35.02	37.49	30.99	40.37	33.88	**37.20**
GPT-5.4 + PhysMaster	44.62	49.39	48.30	44.44	37.75	**45.84**
Relative improvement	27.41%↑	31.74%↑	55.86%↑	10.08%↑	11.42%↑	**23.23%**↑
Gemini-3.1-Pro	37.41	43.74	47.52	47.64	40.76	**44.27**
Gemini-3.1-Pro + PhysMaster	45.82	46.26	59.33	51.93	43.75	**49.31**
Relative improvement	22.48%↑	5.76%↑	24.85%↑	9.00%↑	7.34%↑	**11.38%**↑
Claude-Opus-4.6	28.75	39.36	40.46	39.98	32.10	**37.43**
Claude-Opus-4.6 + PhysMaster	35.69	50.58	42.20	43.44	33.50	**43.50**
Relative improvement	24.14%↑	28.51%↑	4.30%↑	8.65%↑	4.36%↑	**16.22%**↑

ial research targets, open solution paths, and long-horizon workflows. At the same time, each task is equipped with verifiable final answers and structured rubrics, including analytical formulas, numerical values, discrete judgments, and intermediate reasoning checkpoints. A sample task of PRL-Bench is shown in appendix A

4.2. Comparative Evaluation of Frontier Models and PhysMaster-Augmented Models

We report the PRL-Bench performance of standalone frontier models and the corresponding PhysMaster-augmented settings.

All evaluations follow the same protocol as PRL-Bench. Code execution is allowed, so that models can perform numerical computation and programmatic validation when required by the task. External search is disabled to prevent information leakage from retrieving the original papers or related solutions. For the PhysMaster-augmented settings, we constrain the exploration budget to at most 10 nodes for each task.

Each problem is executed five times per model to reduce stochastic variance, and the final score is averaged across runs. Evaluation follows an LLM-as-judge protocol, where GPT-5 is used as a strict judge to verify both final-answer correctness and satisfaction of intermediate rubric items. Scores are assigned according to the predefined answer and rubric points for each subtask and are normalized to a 0–100 scale.

Table 3 reports the PRL-BENCH baseline results for three representative frontier models and, in parallel, reserves the corresponding rows for model+PHYSMASTER evaluation. Among the standalone baselines shown here, Gemini-3.1-Pro achieves the strongest overall score at 44.27, while GPT-5.4 and Claude-Opus-4.6 remain at 37.4. These results indicate that research-oriented physics tasks remain highly challenging even for strong frontier models.

The comparative evaluation shows that the PhysMaster-augmented settings mproves performance across all five subfields, increasing the overall score by 10%-15%. The improvement is especially pronounced in astrophysics, condensed matter physics, and statistical physics, where the tasks often require heterogeneous modeling choices and long-range coordination across analytical and computational components. These results indicate that PhysMaster provides a substantial system-level gain: task decomposition, external knowledge support, and structured execution jointly improve the stability and effectiveness of frontier models on physics research tasks..

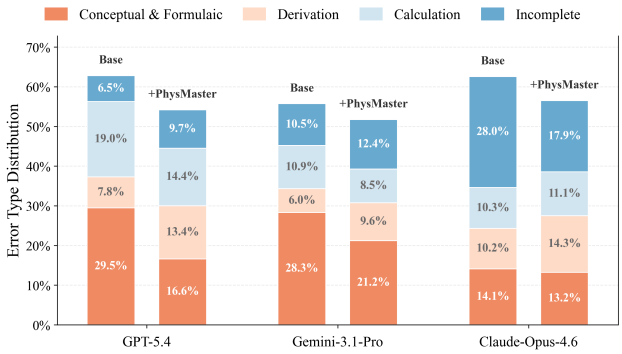

Figure 4. Comparative error type distribution of standalone frontier models and PhysMaster-augmented models

4.3. Error Analysis

To analyze failure modes more rigorously, we divide the errors into four categories. **Conceptual or formulaic errors** refer to failures caused by selecting an inappropriate theoretical model, formula, approximation regime, or boundary condition. These errors primarily reflect insufficient command of advanced physics knowledge and an inability to identify the correct theoretical starting point. **Derivation errors** arise when the reasoning chain contains spurious formulas, unjustified assumptions, or intermediate

steps that do not logically support subsequent conclusions. These errors reflect instability in symbolic reasoning and long-range argument construction. **Calculation errors** include algebraic mistakes, numerical inaccuracies, or incorrect program-execution results, and therefore measure the reliability of computation-assisted reasoning. **Incomplete** responses occur when the model omits key subtasks, provides only partial answers, or fails to produce a complete solution over a long research workflow.

Existing standalone-model results show that failures arise from a coupled limitation in domain knowledge, derivation stability, numerical reliability, and long-horizon task adaptation. In particular, **conceptual and formulaic errors** constitute the dominant failure mode for most frontier models, indicating that frontier models still lack advanced physics knowledge and tend to apply relevant but invalid theoretical template. **Derivation and calculation errors** further indicate that even when the high-level framework is plausible, models still struggle to maintain rigorous intermediate reasoning and numerically reliable validation over extended horizons.

The comparative evaluation allows us to examine whether PHYSMASTER changes both overall accuracy and the composition of remaining errors. The significant reduction in **conceptual and formulaic errors** directly reflects the effect of the trusted academic knowledge base and precise retrieval-augmented generation, which provide more stable domain knowledge support and reduce the probability of selecting an invalid theoretical framework, formula, or applicability condition. For frontier theoretical-physics tasks, this knowledge layer is not merely a source of background information; it constrains the initial research trajectory and lowers the risk of pursuing an incorrect physical line of reasoning.

A reduction in **calculation errors** reflects the effect of task scheduling and reliable knowledge. By decomposing a complex research problem into more controllable subtasks, PHYSMASTER reduces the context burden on each THEORETICIAN agent responsible for concrete derivations or numerical computations. At the same time, the knowledge base and retrieval mechanism provide additional constraints during reasoning and programmatic validation, making both symbolic and numerical execution more stable.

However, the relatively minor derivation errors do not decrease and instead show a slight increase. We conjecture that this is because multi-agent decomposition can reduce local context pressure, but cannot fully eliminate information loss across agents. Meanwhile, even after agentic augmentation, the intrinsic reasoning capacity of the base model is likely to remain a central limiting factor.

For the similarly minor incomplete errors, we also observe a slight increase. This is mainly attributable to our controlled exploration budget and possible information loss across the agentic workflow.

Notably, Claude-Opus-4.6 exhibits a distinct failure pattern, with incomplete error. We attribute this to Claude's engineering-oriented tendency, which leads to over-attention to details, repeated derivation attempts and iterative self-corrections. These behaviors can exhaust the context window or introduce information loss over long trajectories, thereby producing incomplete final answers. This phenomenon is substantially mitigated by PhysMaster augmentation.

This pattern would suggest that PHYSMASTER mitigates the most knowledge- and computation-sensitive failure modes, while the remaining errors reveal deeper constraints imposed by long-range reasoning and base-model capability.

5. Conclusion

In this work, we presented PhysMaster, an autonomous research agent system for theoretical physics, designed to support the closed-loop workflow from physical-picture understanding to theoretical modeling and numerical computation. Rather than acting as a single-pass language interface, PhysMaster organizes problem solving through hierarchical agent collaboration, MCTS-based multi-path exploration, explicit search-time memory, and a compact external knowledge base through LANDAU.

To evaluate this system scientifically, we construct PRL-Bench, a research-oriented benchmark built from 100 *Physical Review Letters* papers across five theory- and computation-intensive subfields. The benchmark preserves research context, workflow structure, and verifiable targets, making it a stringent testbed for long-horizon scientific agents rather than only for isolated reasoning ability. In the current version of this paper, we retain the published frontier-model baselines and reserve paired result slots for model+PhysMaster evaluation under matched conditions.

Overall, this work argues that building a useful AI physicist requires more than stronger answer generation. It requires explicit control over decomposition, exploration, execution, memory, and knowledge grounding. PhysMaster is an initial systems answer to this requirement, and PRL-Bench provides a benchmark that exposes why such structure is necessary. Together, they clarify both the promise and the current limits of AI systems for frontier physics research, and point to future progress in deeper domain knowledge integration, more stable long-horizon exploration, stronger numerical reliability, and more rigorous evaluation of scientific research agents.

Impact Statement

This work aims to advance reliable agentic AI systems for frontier theoretical and computational physics. By integrating long-horizon exploration, trusted external knowledge, and code-assisted verification, PHYSMASTER may help improve the reliability and efficiency of AI-driven physics research, while PRL-BENCH provides reliable evaluations in end-to-end scientific tasks. Potential risks include conceptual errors, lack of advanced theoretical-physics knowledge, and unstable derivations over extended horizons. We partially mitigate these risks through explicit workflow decomposition, critic-guided evaluation, execution traces, and verifiable rubrics. We therefore view PHYSMASTER as a autonomous research system to accelerate physics research, rather than a replacement for human scientific judgment.

References

Chung, D. J., Gao, Z., Kvasiuk, Y., Li, T., M"unchmeyer, M., Rudolph, M., Sala, F., and Tadepalli, S. C. Theoretical physics benchmark (tpbench)—a dataset and study of ai reasoning capabilities in theoretical physics. *Machine Learning: Science and Technology*, 6(3):030505, 2025.

Feng, K., Zhao, Y., Liu, Y., Yang, T., Zhao, C., Sous, J., and Cohan, A. Physics: Benchmarking foundation models on university-level physics problem solving. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11717–11743, 2025.

Ghareeb, A. E., Chang, B., Mitchener, L., Yiu, A., Szostkiewicz, C. J., Laurent, J. M., Razzak, M. T., White, A. D., Hinks, M. M., and Rodriques, S. G. Robin: A multi-agent system for automating scientific discovery. *arXiv preprint arXiv:2505.13400*, 2025.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

Huang, Z., Wang, Z., Xia, S., Li, X., Zou, H., Xu, R., Fan, R.-Z., Ye, L., Chern, E., Ye, Y., Zhang, Y., Yang, Y., Wu, T., Wang, B., Sun, S., Xiao, Y., Li, Y., Zhou, F., Chern, S., Qin, Y., Ma, Y., Su, J., Liu, Y., Zheng, Y., Zhang, S., Lin, D., Qiao, Y., and Liu, P. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai, 2025. URL https://arxiv.org/abs/2406.12753.

Liu, Z., Cai, Y., Zhu, X., Zheng, Y., Chen, R., Wen, Y., Wang, Y., Chen, S., et al. Ml-master: Towards ai-for-ai via integration of exploration and reasoning. *arXiv preprint arXiv:2506.16499*, 2025.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Mitchener, L., Yiu, A., Chang, B., Bourdenx, M., Nadolski, T., Sulovari, A., Landsness, E. C., Barabasi, D. L., Narayanan, S., Evans, N., et al. Kosmos: An ai scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025.

Natarajan, V. et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Qiu, S., Guo, S., Song, Z.-Y., Sun, Y., Cai, Z., Wei, J., Luo, T., Yin, Y., Zhang, H., Hu, Y., et al. Phybench: Holistic evaluation of physical perception and reasoning in large language models. *arXiv preprint arXiv:2504.16074*, 2025.

Qiu, S., Deng, J., Deng, Y., Dong, H., Fu, J., Li, M., Li, Z., Zhang, Z., Zheng, H., Bao, L., Lv, A., Mo, Z., Niu, Y., Peng, Y., Tian, Y., Wang, Y., Wang, Z., Wang, Z.-Y., Wei, J., Wu, L., Xue, A., Yang, L., Yuan, G., Zhan, X., Zhang, J., Zheng, Z., Liu, P., Zhen, L., Li, K., Li, Q., Zhou, Z., Nian, G.-E., Xiao, Y., Cao, Q.-H., Dai, L., Feng, X., Gao, P., Gu, Y., Liu, C., Liu, J., Luo, M.-x., Ma, Y.-Q., Peng, L.-Y., Song, H., Wang, S., Wang, C., Wang, T., Wang, Y.-N., Wu, C., Zhao, P., and Zhu, H. X. Prbench: End-to-end paper reproduction in physics research. 2026. URL https://arxiv.org/abs/2603.27646.

Sun, H., Min, Y., Chen, Z., Zhao, W. X., Fang, L., Liu, Z., Wang, Z., and Wen, J.-R. Challenging the boundaries of reasoning: An olympiad-level math benchmark for large language models, 2025. URL https://arxiv.org/abs/2503.21380.

Wang, M., Lin, R., Hu, K., Jiao, J., Chowdhury, N., Chang, E., and Patwardhan, T. Frontierscience: Evaluating ai's ability to perform expert-level scientific tasks. *arXiv preprint arXiv:2601.21165*, 2026.

Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., and Ha, D. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

A. Sample PRL-Bench-Style Task

Tensor-network simulation in (2+1)d lattice gauge theory

Introduction

Traditional Monte Carlo sampling for lattice gauge fields is effective for equilibrium properties, but faces major obstacles for real-time dynamics and sign problems. Tensor-network approaches provide a classical simulation strategy that avoids the sign problem. While (1+1)d systems are well controlled, extending to (2+1)d with PEPS remains challenging due to higher-dimensional entanglement and exact gauge constraints.

We consider an Abelian lattice gauge theory on a square lattice with Hamiltonian

$$H = H_M + H_B + H_E,$$

with gauge invariance enforced by Gauss-law constraints. A gauge-invariant PEPS ansatz is adopted, with vertex tensors A and link tensors B. Consider open boundary conditions.

Core tasks

1. (2pt) **Pure $\mathbb{Z}_3$ gauge theory.** For a 5×5 lattice with $(D, D_k, h) = (6, 2, 1)$, characterize the local response of the ground-state energy to the coupling g at $g = 0.375$ by determining its first and second derivatives.

2. (3pt) **Odd $\mathbb{Z}_2$ theory.** On a 6×6 lattice with static charges $Q_{\mathbf{x}} = 1$, evaluate the squared expectation values of the staggered observables D_x and D_y at $(h, g) = (1, 0.4)$, and assess their symmetry relation.

3. (1pt) **$\mathbb{Z}_2$ gauge theory with hard-core bosons.** For $Q_{\mathbf{x}} = 0$ and $(m, h, g, J) = (0, 1, 0.33, 0.5)$ on a half-filled 4×4 lattice, determine the ground-state expectation value of the Hamiltonian restricted to the central 3×3 bulk.

4. (5pt) **Real-time dynamics.** Starting from the matter-free $\mathbb{Z}_2$ ground state on a 6×6 lattice, create a localized vison excitation and evolve it in real time with $\Delta t = 0.005$ at $(g, h) = (0.1, 1.0)$. Evaluate:
   - At $T = 4.5$, the plaquette $\langle P_{0,0} \rangle$ and total energy $\langle H \rangle$;
   - At $T = 6.0$, the plaquette $\langle P_{0,1} \rangle$ and total energy $\langle H \rangle$;
   - At $T = 7.5$, the plaquette $\langle P_{1,1} \rangle$ and total energy $\langle H \rangle$.

Answers & Rubrics

1. Answer: (1pt) $\frac{\partial \langle H \rangle}{\partial g} = 29.2609913710$, (1pt) $\frac{\partial^2 \langle H \rangle}{\partial g^2} = -179.0345184105$

2. Answer: (1pt) $\langle D_x \rangle^2 = 0.2421075221119777$, (1pt) $\langle D_y \rangle^2 = 0.2421075221119777$

   Rubrics: (1pt) Identifies that C_4 symmetry enforces $\langle D_x \rangle^2 = \langle D_y \rangle^2$.

3. Answer: (1pt) $\langle h_c \rangle = -4.3252682631$

4. Answer:
   - (1pt) $\langle P_{0,0} \rangle = 0.56388$, (1pt) $\langle H \rangle = -34.64316$
   - (1pt) $\langle P_{0,1} \rangle = 0.80091$, (1pt) $\langle H \rangle = -34.64316$
   - (1pt) $\langle P_{1,1} \rangle = 0.65918$, (1pt) $\langle H \rangle = -34.64316$

   Rubrics:
   - (1pt) Recognizes conservation of total energy during real-time evolution.
   - (1pt) Infers vison propagation from spatiotemporal plaquette dynamics.

B. Typical Authoritative Textbook Priors in LANDAU

Quantum Field Theory, Renormalization, and High-Energy Theory

- Steven Weinberg, *The Quantum Theory of Fields*, Vol. 1–3
- Ashok Das, *Field Theory: A Path Integral Approach*
- J. C. Collins, *Renormalization*
- L. D. Faddeev, *Gauge Fields: An Introduction to Quantum Theory*
- Joseph Polchinski, *String Theory*, Vol. I–II
- Ralph Blumenhagen et al., *Introduction to Conformal Field Theory*
- Pierre Deligne et al., *Quantum Fields and Strings: A Course for Mathematicians*

Particle Physics, Standard Model, and Perturbative QCD

- Ta-Pei Cheng and Ling-Fong Li, *Gauge Theory of Elementary Particle Physics*
- John C. Collins, *Foundations of Perturbative QCD*
- Yuri V. Kovchegov and Eugene Levin, *Quantum Chromodynamics at High Energy*

Condensed Matter Field Theory and Many-Body Theory

- Alexander Altland and Ben Simons, *Condensed Matter Field Theory*
- Piers Coleman, *Introduction to Many-Body Physics*
- Henrik Bruus and Karsten Flensberg, *Many-Body Quantum Theory in Condensed Matter Physics*
- Nicolas Dupuis, *Field Theory of Condensed Matter and Ultracold Gases*
- Subir Sachdev, *Quantum Phases of Matter*

Statistical Physics and Non-equilibrium Physics

- Mehran Kardar, *Statistical Physics of Particles*
- S. R. de Groot and P. Mazur, *Non-equilibrium Thermodynamics*

General Relativity, Black Holes, and Cosmology

- Steven Weinberg, *Gravitation and Cosmology*
- S. Chandrasekhar, *The Mathematical Theory of Black Holes*
- Mike Guidry, *Modern General Relativity: Black Holes, Gravitational Waves, and Cosmology*
- Daniel Baumann, *Cosmology*

Astrophysics, Galactic Dynamics, Accretion, and Compact Objects

- James Binney and Scott Tremaine, *Galactic Dynamics*
- Juhan Frank, Andrew King and Derek Raine, *Accretion Power in Astrophysics*
- Stuart L. Shapiro and Saul A. Teukolsky, *Black Holes, White Dwarfs and Neutron Stars*
- Norman K. Glendenning, *Compact Stars: Nuclear Physics, Particle Physics and General Relativity*