QMP-Bench: A Research-Level Benchmark for Autonomous End-to-End Quantum Many-Body Simulations

Ken Deng [1] **Xiangfei Wang** [1] **Guijing Duan** [2] **Chen Mo** [3] **Junkun Huang** [4] **Runqing Zhang** [5] **Ling Qian** [5] **Zhiguo Huang** [5] **Jize Han** [5] **Di Luo** [1 6]

Abstract

Although LLM-driven systems exhibit remarkable potential for automated scientific discovery, the community severely lacks realistic, end-to-end benchmark datasets for actual research, particularly for scientific simulations. To address this gap, we introduce QMP-Bench, a research-level dataset comprising 100 end-to-end simulation tasks extracted directly from 21 high-impact articles in quantum many-body physics. Unlike human-curated problems, QMP-Bench challenges agents to autonomously reproduce physical results from scratch using domain-specific numerical libraries including ITensors, NetKet, Qiskit, and ORCA. Baseline evaluations of frontier LLMs on these tasks reveal severe struggles with domain-specific programming errors and physical hallucinations, whereas introducing a framework equipped with verification mechanisms significantly improves the performance. Consequently, QMP-Bench serves as a rigorous testbed for driving the development of reliable and physics-informed AI physicists.

1. Introduction

Fueled by the growing capabilities of large language models (Wei et al., 2022; Yao et al., 2023a; Guo et al., 2025) and sophisticated agent paradigms (Yao et al., 2023b; Gao et al., 2023; Madaan et al., 2023; Hou et al., 2025), the landscape of scientific research is being rapidly transformed by automated discovery systems. These AI-driven approaches have demonstrated immense potential across broad research frontiers, accelerating advancements in machine learning (Lu et al., 2024; Tang et al., 2025; Li et al., 2025), material synthesis (Szymanski et al., 2023; Mandal et al., 2025), and physical sciences (Desai et al., 2025; Cao et al., 2025; Sha et al., 2025; Song et al., 2025). More specifically, within the realm of physics, these systems find a highly natural application in scientific simulations, with recent works highlighting the strong performance of LLM-based agents (Miao et al., 2025; Gustin et al., 2026; Li et al., 2026).

Despite recent advancements, the community urgently needs physics-centric benchmark datasets that are grounded in realistic scientific environments and present research-level difficulty to truly assess AI scientists. Current physics benchmarks are roughly divided into two groups. The first group fundamentally relies on human-curated, well-defined problems, encompassing both exam- and Olympiad-level evaluations (Wang et al., 2024; Rein et al., 2024; Feng et al., 2025; Qiu et al., 2025), as well as benchmarks (Zhu et al., 2025; Chung et al., 2025) that attempt to simulate professional research tasks. They differ from genuine scientific exploration, limiting their ability to fully reflect the capacities of frontier LLMs and agents in realistic workflows. The second group comprises research-level benchmarks that derive tasks directly from research articles, thereby significantly closing the gap between AI systems and realistic physics research (Qiu et al., 2026; Miao et al., 2026). However, these benchmarks encompass a broad range of physical disciplines and lack a concentrated evaluation of LLM agents that perform sophisticated scientific simulations for quantum many-body problems.

In this work, we introduce QMP-Bench, an end-to-end benchmark for research-level **Q**uantum **M**any-Body **P**hysics, which contains 100 simulation tasks directly extracted from 21 high-impact articles. Quantum many-body simulations exhibit complex emergent phenomena and represent technically demanding frontier physics, making them a premier testbed for AI systems. QMP-Bench forces agents

[1]Department of Physics, Tsinghua University, Beijing 100084, China [2]Department of Physics and Beijing Key Laboratory of Optoelectronic Functional Materials & Micro-nano Devices, Renmin University of China, Beijing 100872, China [3]School of Physics, Sun Yat-sen University, Guangzhou 510275, China [4]School of Gifted Young, University of Science and Technology of China, Hefei 230026, China [5]China Mobile (Suzhou) Software Technology Co., Ltd., Suzhou 215163, China [6]Institute of Advanced Study, Tsinghua University, Beijing 100084, China. Correspondence to: Di Luo <diluo1000@gmail.com>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

to independently reproduce physical results from scratch via integrating specialized libraries such as ITensors, NetKet, Qiskit, and ORCA. Baseline evaluation of frontier LLMs indicates that despite their remarkable general capacities, they fall short when faced with domain-specific programming errors and physical hallucinations. Furthermore, evaluating these tasks with recently proposed tools like the PhysVEC framework (Deng et al., 2026) reveals that architectures equipped with self-verification mechanisms achieve significant improvements, establishing QMP-Bench as an effective testbed for cultivating future AI scientists.

2. Related Work

2.1. AI Physicists

Recently, representative works have begun to integrate agent systems directly into physics problem-solving and scientific research (Miao et al., 2025; Gustin et al., 2026; Li et al., 2026). These systems exhibit the impressive potential of LLM-based approaches, especially in the field of numerical physics computations. Naturally, there is a strong demand within the community for research-level benchmarks, such as QMP-Bench, that can accurately measure an LLM's capacity for end-to-end reproduction (including information integration, code implementation, and physical fact judgment) during practical physical research.

2.2. Scientific Benchmarks for AI Scientists

Existing studies, such as SciBench (Wang et al., 2024), GPQA (Rein et al., 2024), PHYSICS (Feng et al., 2025) and Physics Supernova (Qiu et al., 2025), have provided excellent benchmarks for evaluating the scientific reasoning capabilities at the exam-level or Olympiad-level tasks. Works such as CriPt (Zhu et al., 2025) and TPBench (Chung et al., 2025) have introduced benchmarks that step closer to professional physics research. However, these remain human-curated tasks, leaving a gap with realistic research scenarios.

Recently, works like PRBench (Qiu et al., 2026) and PRL-Bench (Miao et al., 2026) have proposed physical research tasks directly based on research articles, sharing a similar objective with QMP-Bench. Nevertheless, QMP-Bench distinguishes itself by concentrating specifically on quantum many-body simulations, and by demanding the utilization of community-recognized third-party libraries to accomplish these tasks.

3. Benchmark

Quantum many-body physics studies systems of interacting quantum particles whose collective behavior gives rise to emergent phenomena such as superconductivity and topological phases, and whose simulation is inherently challenging due to exponential complexity. Therefore, quantum many-body simulations are both scientifically important and practically challenging to perform in research, making this domain a valuable testbed for AI-driven research. More specifically, in this work, we focus on problems that can be addressed using tensor networks, neural network ansatz, quantum circuit simulation, and density-functional theory calculation.

3.1. Task Design

In QMP-Bench, we define the tasks as the direct reproduction of figures and subplots from high-impact research articles in the field of quantum many-body simulations. Unlike benchmarks built around human-curated intermediate artifacts, QMP-Bench provides end-to-end tasks that require reproducing results directly from the original article.

As illustrated in Fig. 1 (b), each task in the dataset consists of two components: (i) The .tex source file of the corresponding article collected from arXiv. (ii) A short target description, in which we explicitly specify the figure in the article to be reproduced. If a figure involves several parameter choices or extensive scans, we restrict the task to reproducing only a few representative cases. The LLM-driven agent then starts to construct the simulation script based on the article's .tex file and the corresponding target.

3.2. Dataset Composition

Fig. 1 (a) demonstrates the composition of this dataset. QMP-Bench includes 36 tensor-network problems (denoted `dmrg`) with ITensors (Fishman et al., 2022), 18 neural-network ansatz problems (`nqs`) with NetKet (Carleo et al., 2019; Vicentini et al., 2022), 20 quantum circuit problems (`qsim`) with Qiskit (Aleksandrowicz et al., 2019), and 26 density-functional theory problems (`dft`) with ORCA (Neese, 2012). These tasks are drawn from 21 published high-impact articles, including Physical Review Letters, Science, Nature communications, and JACS.

As summarized in Fig. 1 (b), QMP-Bench covers a range of canonical quantum many-body simulation methods and tasks. To be specific, `dmrg` tasks contain physical models such as Bosons and Electrons; `nqs` incorporates methods like RBM and tVMC; `qsim` includes algorithms like Trotter and IQPE; and `dft` addresses problems like MOs and XAS.

3.3. Evaluation Pipelines

To comprehensively evaluate the performance of current frontier LLMs on the QMP-Bench dataset, we construct two distinct evaluation pipelines (Fig. 1 (c)):

- `Conventional ReAct Baselines`: The first

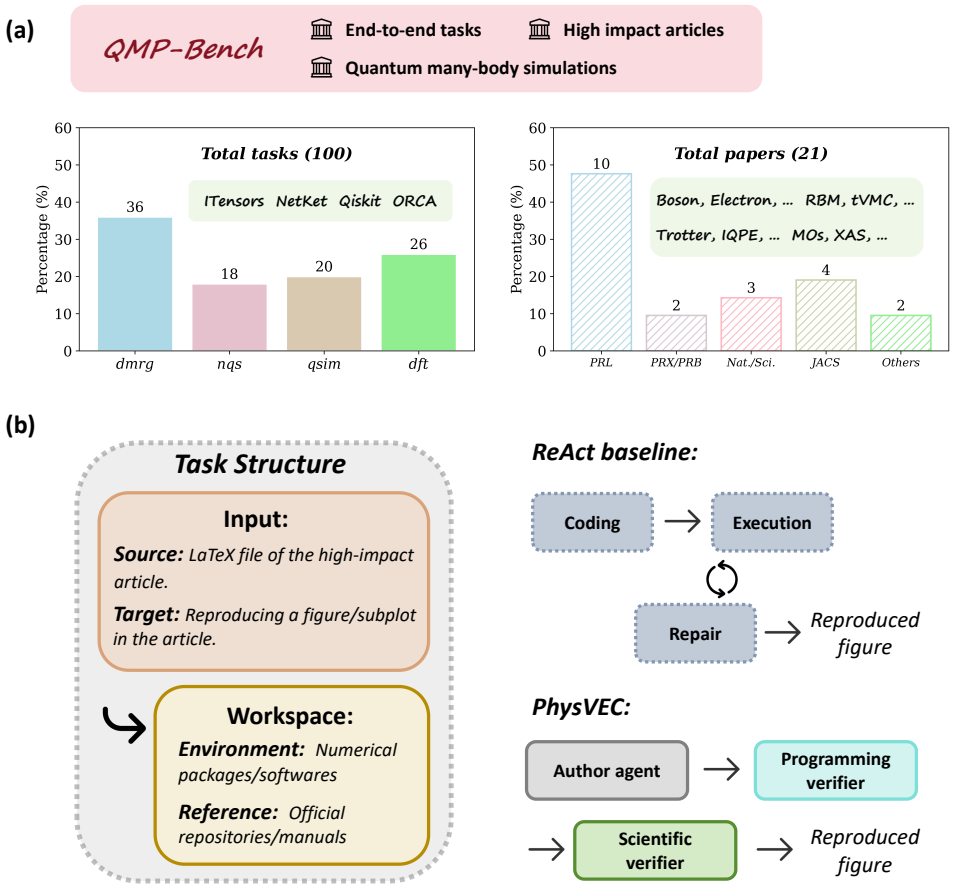

Figure 1. QMP-Bench dataset and evaluation pipelines. (a) QMP-Bench: a benchmark dataset comprising 100 figures drawn from 21 high-impact articles, covering numerical studies that can be implemented with ITensors, NetKet, Qiskit, and ORCA. (b) Evaluation pipelines. We evaluate both conventional ReAct-style iteration workflows and an existing verification-driven approach, PhysVEC.

pipeline follows a standard ReAct-style architecture. The LLM reads the article, handles the coding of the calculation script, and then enters an iterative execution-repair loop to refine code quality (ReAct). We also include a variant, ReAct-RAG, which allows the LLM to actively consult local references (e.g., numerical library manuals and documents) during the tuning loop to aid the repair phase.

- `The PhysVEC Framework`: The second pipeline adopts the PhysVEC framework as introduced in Deng et al. (2026), which integrates self-verification through refinement. After the Author agent proposes the initial script, the pipeline relies on two dedicated verification modules. The Programming verifier uses unit tests and integration tests to strictly enforce programming correctness. Subsequently, the Scientific verifier steps in to ensure physical validity by utilizing physics-informed checks like rubric tests and physical assertion tests.

In the subsequent sections, we will demonstrate that QMP-Bench poses a significant challenge for conventional ReAct architectures, where frontier LLMs generally exhibit poor performance. However, when evaluated under the PhysVEC framework, its specific agent architecture enables the same LLMs to accomplish more tasks. These comparative results highlight the critical importance of introducing effective verification mechanisms in automated scientific research systems.

4. Results

4.1. Programming Correctness

Scientific simulations using domain-specific libraries are often plagued by subtle programming errors, particularly those involving third-party library calls, interface compatibility, and complex data structures. To systematically investigate these challenges, we evaluate several frontier LLMs, including `GPT-5.1`, `Gemini 2.5 Flash`, `Qwen3-Max`, and `Claude Sonnet 4`, on our proposed QMP-Bench

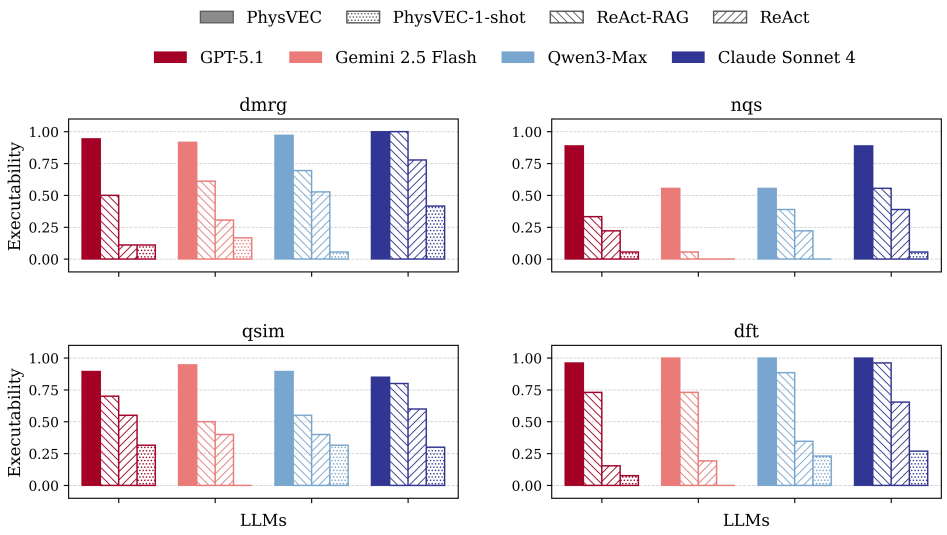

Figure 2. Executability of all evaluation pipelines. PhysVEC-1-shot denotes the results generated directly from the Author agent without any further verification and error correction.

dataset. In this evaluation, we explicitly compare conventional ReAct baselines against PhysVEC framework with self-verification design across the generated scripts.

For each pair of LLMs and topics, we define executability S as the fraction of scripts that successfully run without errors. In our evaluation, the maximum number of refinement iterations for all frameworks is set to 8. The quantitative results shown in Fig. 2 highlight that the end-to-end simulation tasks in QMP-Bench pose a challenge to conventional agent architectures. Specifically, the standard ReAct framework struggles with its success rate generally falling below 50%. Furthermore, even when external knowledge retrieval is incorporated during the refinement loops (ReAct-RAG), the executability of the generated scripts remains far from reliable, particularly pronounced in `nqs` tasks. In contrast, frameworks with verification modules like PhysVEC demonstrate better robustness against these bottlenecks. Through systematic diagnosis and error correction governed by the Programming Verifier, PhysVEC substantially improves the final script executability across all topics, bringing the success rate close to 100% in most of cases.

Fig. 3 reveals the step-by-step improvement of script executability across refinement iterations. The panels show the trajectories of `GPT-5.1` on `dmrg` and `nqs` tasks as representative examples, with results of other model-topic pairs demonstrating similar trends. Crucially, these curves highlight the difficulty of QMP-Bench: simple ReAct architectures fail to meaningfully boost programming correctness through refinement loops, yielding flat curves that plateau at low success rates. Conversely, applying the PhysVEC approach yields higher programming correctness in fewer iterations and significantly raises the overall upper bound of the refinement process, achieving reliable end-to-end script generation.

This severe challenge posed by QMP-Bench generally arises from two main complexities. First, these end-to-end tasks require LLMs to build complete research-level computational pipelines, demanding extensive reasoning chains. For instance, a `dmrg` task involves everything from lattice initialization and Hamiltonian construction to solver configuration and observable definitions. Second, these numeric simulations rely heavily on domain-specific libraries (e.g. ITensors, NetKet, etc.), introducing dense grammatical and structural intricacies during function calls. In conventional refinement iterations, executing such scripts typically crashes at the first encountered error, which restricts the environmental feedback to a single bug per step and leads to a sluggish and localized refinement trajectory. The PhysVEC framework deploys a Programming verifier that utilizes unit tests and integration tests to systematically diagnose and correct errors across the entire script, resulting in better performance.

4.2. Physical Validity

In the context of scientific simulations, generating executable and bug-free code is only the initial step. To be functionally reliable in real-world research scenarios, the generated scripts must also achieve strict physical validity by faithfully reflecting the fundamental definitions, assumptions, and constraints of the target problem. Therefore, to evaluate the capability of AI-driven systems to achieve this essential physical validity in actual research settings, we assess the performance of conventional ReAct baselines and the PhysVEC framework on QMP-Bench.

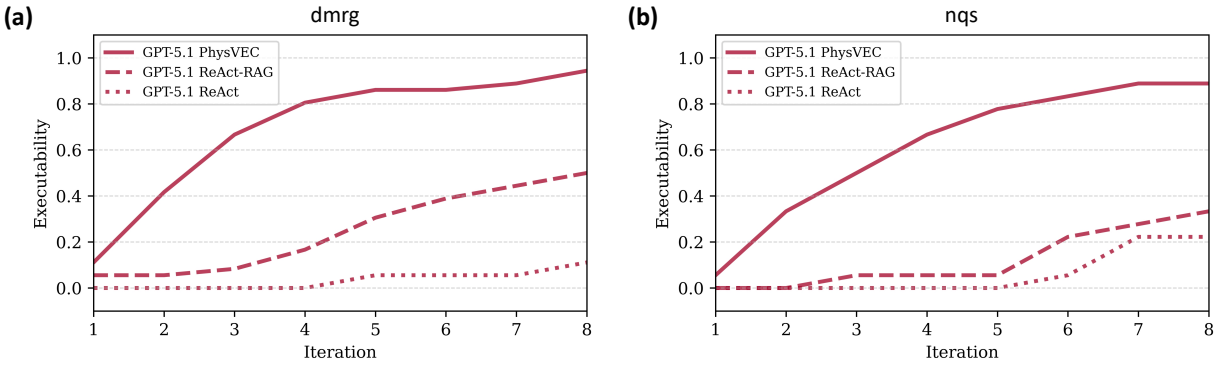

Figure 3. Executability over iterations for `GPT-5.1` on (a) `dmrg` tasks and (b) `nqs` tasks. These panels illustrate the comparison between standard ReAct baselines and the PhysVEC framework. Other model-topic combinations show similar trajectories.

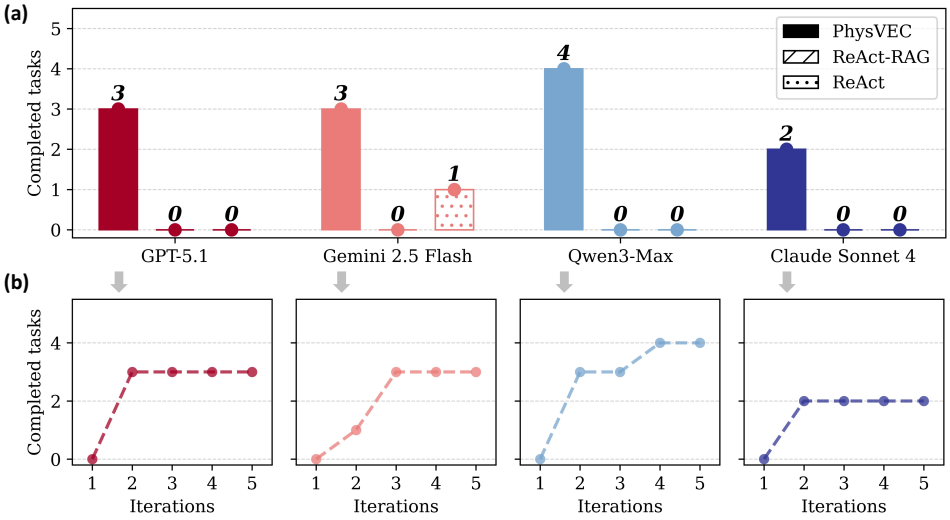

Figure 4. Performance in physical validity on QMP-Bench subset. (a) Number of completed tasks of all evaluation pipelines on the QMP-Bench subset. (b) Cumulative number of completed tasks as a function of iteration under the PhysVEC framework.

Due to computational constraints, we restrict the evaluation of script physical validity to a five-task subset (`dmrg` tasks) of QMP-Bench. Success is determined by comparing the final execution outputs against the ground truth of the original article results, as summarized in Fig.4 (a). The results highlight a critical failure set in conventional ReAct architectures: standard ReAct and ReAct-RAG baselines are virtually incapable of reproducing accurate physical results. Across all evaluated LLMs, these baselines fail entirely, except for one isolated success by `Gemini 2.5 Flash`. In contrast, the PhysVEC framework yields a substantial improvement, results in significantly more faithful physical results. Furthermore, Fig. 4 (b) presents the cumulative number of successfully completed tasks as a function of refinement iterations for PhysVEC.

The poor performance of the baselines primarily is due to the realistic, end-to-end nature of QMP-Bench. In QMP-Bench tasks, LLMs are forced to operate directly from original scientific articles, demanding a precise comprehension and implementation of both the underlying physical systems and the numerical solver configurations. With this demanding requirement, conventional baselines frequently exhibit severe physical error modes, including misconstructed Hamiltonians, incorrect implementations of observables (e.g. correlation functions), and unconverged simulation results. These failures can be largely mitigated by frameworks like PhysVEC, which rely on a Scientific verifier to perform physics-informed checks (e.g., comparing against rubrics, automatically executing limiting case tests, and performing convergence judgments) that ensure strict physical alignment with the original text.

5. Conclusion

In conclusion, we constructed QMP-Bench, an end-to-end evaluation dataset comprising 100 quantum many-body

physics tasks extracted from 21 research articles. Moving beyond simplified human-curated questions, QMP-Bench uniquely challenges agents to independently reproduce research findings using specialized libraries like ITensors, NetKet, Qiskit, and ORCA. Baseline evaluations using frontier LLMs (`GPT-5.1`, `Gemini 2.5 Flash`, `Qwen3-Max`, `Claude Sonnet 4`) indicated that these models struggle significantly with domain-specific errors in authentic workflows. Nevertheless, evaluations using the PhysVEC framework showed that self-verification designs dramatically improve performance, yielding physically faithful results. This solidifies QMP-Bench as a vital and rigorous testbed for guiding the future development of physics-informed AI agents.

We acknowledge certain limitations in the current design of QMP-Bench. Primarily, validating scientific tasks within the dataset still requires human-crafted rubrics and physical assertions. To efficiently scale realistic benchmarks, developing methods to automatically synthesize these evaluation criteria remains a crucial open question. Additionally, the scope of QMP-Bench is currently confined to the reproduction of established findings in the existing literature. To fully realize the potential of AI physicists, the community still requires advanced evaluation protocols designed to measure the generation of novel physical hypotheses and predictive theories—the ultimate milestone for end-to-end autonomous scientific discovery.

Software and Data

An extended version of this work is available on arXiv at https://arxiv.org/abs/2604.00149. The complete code and data will be open-sourced upon the publication of the extended manuscript.

Acknowledgements

DL acknowledges support from Beijing Municipal Science and Technology Commission and Zhongguancun Science Park Administrative Committee (No. 20251090054). JH acknowledges support from the Natural Science Foundation of Jiangsu Province (No. BK20250404), the Youth Science and Technology Talent Support Project of Jiangsu Province (No. JSTJ-2025-600).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning and its application to autonomous scientific discovery. By providing a rigorous, research-level testbed for evaluating LLM-generated scientific code and physical results, QMP-Bench has the potential to guide the development of reliable AI agents and lower barriers to computational research across scientific disciplines. We encourage the community to deploy such AI systems as tools that augment, rather than replace, human scientific judgment. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Aleksandrowicz, G., Alexander, T., Barkoutsos, P., Bello, L., Ben-Haim, Y., Bucher, D., et al. Qiskit: An open-source framework for quantum computing, January 2019. URL https://doi.org/10.5281/zenodo.2562111.

Cao, S., Zhang, Z., Alghadeer, M., Fasciati, S. D., Piscitelli, M., Bakr, M., Leek, P., and Aspuru-Guzik, A. Automating quantum computing laboratory experiments with an agent-based AI framework. *Patterns*, 6(10):101372, 2025.

Carleo, G., Choo, K., Hofmann, D., Smith, J. E., Westerhout, T., Alet, F., Davis, E. J., Efthymiou, S., Glasser, I., Lin, S.-H., Mauri, M., Mazzola, G., Mendl, C. B., van Nieuwenburg, E., O'Reilly, O., Théveniaut, H., Torlai, G., Vicentini, F., and Wietek, A. NetKet: A machine learning toolkit for many-body quantum systems. *SoftwareX*, 10:100311, 2019. ISSN 2352-7110. doi: https://doi.org/10.1016/j.softx.2019.100311. URL https://www.sciencedirect.com/science/article/pii/S2352711019300974.

Chung, D. J. H., Gao, Z., Kvasiuk, Y., Li, T., Münchmeyer, M., Rudolph, M., Sala, F., and Tadepalli, S. C. Theoretical physics benchmark (TPBench) – a dataset and study of AI reasoning capabilities in theoretical physics. *arXiv preprint arXiv:2502.15815*, 2025.

Deng, K., Wang, X., Duan, G., Mo, C., Huang, J., Zhang, R., Qian, L., Huang, Z., Han, J., and Luo, D. Towards verifiable and self-correcting AI physicists for quantum many-body simulations. *arXiv preprint arXiv:2604.00149*, 2026.

Desai, S., Addamane, S., Tsao, J. Y., Brener, I., Dingreville, R., and Iyer, P. P. Self-driving lab discovers principles for steering spontaneous emission beyond conventional Fourier optics. *Nature Communications*, 17(1):204, Dec 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. *arXiv preprint arXiv:2503.21821*, 2025.

Fishman, M., White, S. R., and Stoudenmire, E. M. The ITensor software library for tensor network calculations. *SciPost Phys. Codebases*, pp. 4, 2022. doi: 10.21468/

SciPostPhysCodeb.4. URL https://scipost.org/10.21468/SciPostPhysCodeb.4.

Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., and Neubig, G. PAL: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2023.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, Sep 2025.

Gustin, I., Calderón, L. M., Pérez-Sánchez, J. B., Gonthier, J. F., Nakamura, Y., Panicker, K., Ramprasad, M., Zhang, Z., Zou, Y., Bernales, V., and Aspuru-Guzik, A. El Agente Cuántico: Automating quantum simulations. *arXiv preprint arXiv:2512.18847*, 2026.

Hou, X., Zhao, Y., Wang, S., and Wang, H. Model Context Protocol (MCP): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*, 2025.

Li, W., Ren, J., Cheng, L., and Gong, C. Autonomous quantum simulation through large language model agents. *arXiv preprint arXiv:2601.10194*, 2026.

Li, Z., Li, Z., Guo, Z., Ren, X., and Huang, C. DeepCode: Open agentic coding. *arXiv preprint arXiv:2512.07921*, 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.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. Self-Refine: Iterative refinement with self-feedback. In *The Thirty-seventh Annual Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=43rnkOcpI1.

Mandal, I., Soni, J., Zaki, M., Smedskjaer, M. M., Wondraczek, K., Wondraczek, L., Gosvami, N. N., and Krishnan, N. M. A. Evaluating large language model agents for automation of atomic force microscopy. *Nature Communications*, 16(1):9104, Oct 2025.

Miao, T., Dai, J., Liu, J., Tan, J., Zhang, M., Jin, W., Du, Y., Jin, T., Pang, X., Liu, Z., et al. Physmaster: Building an autonomous AI physicist for theoretical and computational physics research. *arXiv preprint arXiv:2512.19799*, 2025.

Miao, T., Jin, W., Zhang, M., Tan, J., Hu, Y., Guo, T., Zhang, J., Wang, Y., Li, W., Gao, Y., et al. PRL-Bench: A comprehensive benchmark evaluating LLMs' capabilities in frontier physics research. *arXiv preprint arXiv:2604.15411*, 2026. URL https://arxiv.org/abs/2604.15411.

Neese, F. The ORCA program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2(1):73–78, 2012. doi: 10.1002/wcms.81.

Qiu, J., Shi, J., Juan, X., Zhao, Z., Geng, J., Liu, S., Wang, H., Wu, S., and Wang, M. Physics supernova: AI agent matches elite gold medalists at IPhO 2025. *arXiv preprint arXiv:2509.01659*, 2025.

Qiu, S., Deng, J., Deng, Y., Dong, H., Fu, J., Li, M., Li, Z., Zhang, Z., Zheng, H., Bao, L., et al. PRBench: End-to-end paper reproduction in physics research. *arXiv preprint arXiv:2603.27646*, 2026.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Sha, R., Wang, B., Yang, J., Ma, X., Wu, C., Yan, L., Zhou, C., Liu, J., Wang, G., Yan, S., and Zhu, L. LLM-based multi-agent copilot for quantum sensor. *arXiv preprint arXiv:2508.05421*, 2025.

Song, Z., Zhou, Q., Ren, C., Ling, C., Ju, M., and Wang, J. LLM-Feynman: Leveraging large language models for universal scientific formula and theory discovery. *arXiv preprint arXiv:2503.06512*, 2025.

Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624 (7990):86–91, Dec 2023.

Tang, J., Xia, L., Li, Z., and Huang, C. AI-Researcher: Autonomous scientific innovation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=kQWyOYUAC4.

Vicentini, F., Hofmann, D., Szabó, A., Wu, D., Roth, C., Giuliani, C., Pescia, G., Nys, J., Vargas-Calderón, V., Astrakhantsev, N., and Carleo, G. NetKet 3: Machine learning toolbox for many-body quantum systems. *SciPost Phys. Codebases*, pp. 7, 2022. doi: 10.21468/SciPostPhysCodeb.7. URL https://scipost.org/10.21468/SciPostPhysCodeb.7.

Wang, X., Hu, Z., Lu, P., Zhu, Y., Zhang, J., Subramaniam, S., Loomba, A. R., Zhang, S., Sun, Y., and Wang, W. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, 2023a. URL https://openreview.net/forum?id=1hflw0tjM8.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=WE_vluYUL-X.

Zhu, M., Tian, M., Yang, X., Zhou, T., Yuan, L., Zhu, P., Chertkov, E., Liu, S., Du, Y., Ji, Z., et al. Probing the critical point (CritPt) of AI reasoning: a frontier physics research benchmark. *arXiv preprint arXiv:2509.26574*, 2025.

A. Tensor Network

Tensor network methods represent quantum states as networks of low-rank tensors, providing a compact description of the entanglement structure in many-body systems. In this work, agents are equipped with the ITensors package (Fishman et al., 2022) in Julia to implement tensor-network-based computations.

B. Neural Network Ansatz

Quantum many-body wavefunctions can also be approximated by neural network architectures, which provides a flexible variational ansatz that can efficiently capture complex correlations and entanglement patterns combined with various variational optimization methods. In this work, agents employ the NetKet package (Carleo et al., 2019; Vicentini et al., 2022) to construct neural-network wavefunctions and perform the corresponding variational calculations.

C. Quantum Circuit Simulation

Quantum many-body systems can also be simulated within the quantum circuit framework. Algorithms such as Trotterized time evolution, Iterative Quantum Phase Estimation (IQPE), and Variational Quantum Eigensolvers (VQE) are frequently used for various systems. In this work, the agents use the Qiskit package (Aleksandrowicz et al., 2019) to simulate the quantum circuit and the corresponding algorithms on a classical computer.

D. Density-Functional Theory Calculation

Density-functional theory (DFT) uses the particle density to represent many-electron systems and is widely used for calculating electronic structure and spectral properties. In this work, the agents use the ORCA (Neese, 2012) to perform DFT calculations and analyze the corresponding properties. In ORCA, integration tests are not applicable because the program does not accept incomplete input files.

E. Token and Tool Usage

Figs. 5 (a) and (b) report the token consumption of all evaluation pipelines in the programming test, averaged over all tasks for a given model and topic. The external PhysVEC framework consumes substantially more tokens than the conventional ReAct baselines, due to the more complex multi-agent architecture and repeated verification operations. However, the input token cost typically remains below 1 million, while the output token cost is mostly under 0.3 million, corresponding to only a few dollars per task on average. This cost is practically acceptable for automatic reproduction of scientific codes. Fig. 5 (c) shows the number of tool calls (mainly retrieve operations) in the tests.

Fig. 6 reports the token consumption of PhysVEC during the scientific test stage. Both input and output token costs remain mostly below 0.5 million per task.

F. Dataset Articles

We list the articles included in our benchmark dataset below. The dataset covers tensor network (labeled as 'dmrg'), neural network ansatz ('nqs'), quantum circuit simulation for many-body systems ('qsim') and density-functional theory calculation for quantum chemistry ('dft'). The number of tasks extracted from each article is also listed.

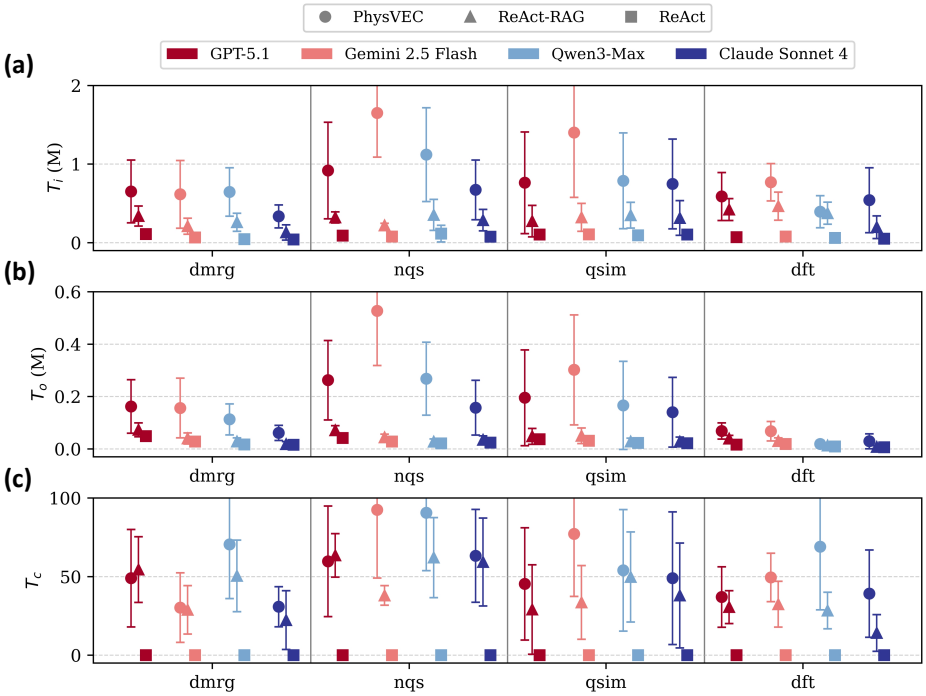

Figure 5. Token usage and tool calls in programming tests. (a) Input tokens T_i. (b) Output tokens T_o. (c) Tool calls T_c.

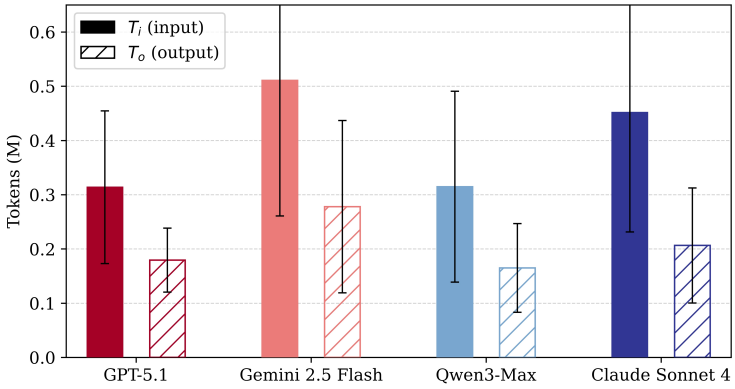

Figure 6. Token usage in scientific tests. Solid bars represent input tokens and hatched bars represent output tokens.

Table 1. Article list for `dmrg` tasks

Journal	Title	Number of Tasks
PRL	Phase diagram of spin-1 bosons on one-dimensional lattices	5
PRL	Correlations in a two–chain Hubbard model	5
PRL	Disorder induced quantum phase transition in random-exchange spin-1/2 chains	6
PRL	Plaquette ordered phase and quantum phase diagram in the spin-1/2 J1-J2 square Heisenberg model	3
PRL	Hidden order in one dimensional Bose insulators	2
PRL	Topological edge states in the one-dimensional super-lattice Bose-Hubbard model	5
PRL	Density-matrix renormalization group study of trapped imbalanced Fermi condensates	5
PRL	One hole in the two-leg t-J ladder and adiabatic continuity to the non-interacting limit	4
PRL	Evidence for a superfluid density in t–J ladders	1

Table 2. Article list for `nqs` tasks

Journal	Title	Number of Tasks
PRL	Symmetries and many-body excited states with neural-network quantum states	4
PRX	Quantum entanglement in neural network states	4
PRB	Quantum skyrmion dynamics studied by neural network quantum states	3
Science	Solving the quantum many-body problem with artificial neural networks	7

Table 3. Article list for `qsim` tasks

Journal	Title	Number of Tasks
npj Quantum Information	Simulating quantum many-body dynamics on a current digital quantum computer	9
EPJ Quantum Technology	Quantum simulation of the Hubbard model on a graphene hexagon: Strengths of IQPE and noise constraints	5
Nature communications	Observing ground-state properties of the Fermi-Hubbard model using a scalable algorithm on a quantum computer	6

Table 4. Article list for `dft` tasks

Journal	Title	Number of Tasks
JACS	Spectroscopic evidence for a 3d10 ground state electronic configuration and ligand field inversion in [Cu(CF3)4]1–	6
JACS	Nature of S-states in the Oxygen-evolving complex resolved by high-energy resolution fluorescence detected X-ray absorption spectroscopy	6
JACS	Combining valence-to-core X-ray emission and Cu K-edge X-ray absorption spectroscopies to experimentally assess Oxidation state in organometallic Cu(I)/(II)/(III) complexes	5
Nature communications	Quantitative prediction of rate constants and its application to organic emitters	4
JACS	Valence-to-core X-ray emission spectroscopy: A sensitive probe of the nature of a bound ligand	5