A Comprehensive Survey on Scientific Large Language Models in Physics

Anonymous ACL submission

Abstract

Large language models (LLMs) have emerged as powerful tools for processing and generating human-like text, raising intriguing possibilities for their application in physics—a field characterized by complex mathematical formulations, abstract concepts, and precise reasoning. While recent studies have demonstrated LLMs' potential in physics applications, from automating simulations to enhancing physics education, the field lacks a systematic framework for understanding and advancing these efforts. This paper presents a comprehensive survey of LLM applications in physics, which examines four critical domains: physical simulation, knowledge discovery, physical reasoning, and physics education, revealing both promising advances and fundamental challenges. We introduce a systematic taxonomy that classifies approaches based on LLM utilization patterns: generic encoders, language generators, auxiliary modules, and autonomous agents. Our analysis uncovers common patterns across successful applications while identifying key limitations in current approaches. We also compile and analyze relevant benchmarks and datasets, providing a resource for evaluating LLM performance in physics tasks. Finally, we outline critical challenges and promising research directions, offering a roadmap for leveraging LLMs to advance both physics research and education.

1 Introduction

Physics, as the fundamental science of matter, energy, and their interactions, underpins our understanding of the universe from quantum phenomena to cosmological scales. Despite significant theoretical and experimental advances (Long et al., 2021), many challenging problems in physics remain unsolved, ranging from efficient simulation of quantum systems to discovering new physical laws from complex experimental data (Lu, 2024). Recent advances in artificial intelligence, particularly in machine learning, have introduced promising new approaches to tackle these challenges (Boehnlein et al., 2022; Karagiorgi et al., 2022; Willard et al., 2020). Graph neural networks (GNNs) (Ramakrishnan et al., 2025; Shen et al., 2025) have proven particularly effective at modeling particle dynamics by capturing complex interactions (Liang et al., 2024; Mayr et al., 2023), while physics-informed neural networks (PINNs) (Karniadakis et al., 2021) have successfully addressed challenging problems in fluid mechanics such as solving Navier-Stokes equations (Cho et al., 2023).

Meanwhile, large language models (LLMs) have introduced a powerful pipeline across various domains, ranging from natural language processing (Zhao et al., 2023) to artificial general intelligence (Zhang et al., 2024a,b; Luo et al., 2025b). LLMs contain billions of parameters by stacking Transformer architectures, which are trained in massive text corpora (Yang et al., 2024). There have been a range of popular commercial LLMs such as GPT (Achiam et al., 2023), Llama (Touvron et al., 2023), and PaLM series (Chowdhery et al., 2023). These models, built on Transformer architectures and trained on massive scientific corpora, have already demonstrated remarkable abilities in tasks ranging from deriving equations from physics principles to suggesting novel experimental designs. Their ability to combine vast knowledge representation (Lu et al., 2023; Shu et al., 2024; Meyer et al., 2023) with human-like reasoning patterns opens unprecedented possibilities for accelerating physics discovery and deepening our understanding of physical systems (Luo et al., 2025a).

In literature, research applying LLMs to physics primarily focuses on four main areas, i.e., simulation (Ali-Dib and Menou, 2024), knowledge discovery (Du et al., 2024), reasoning (Meadows et al., 2024), and education (West, 2023) (see Figure 1). In physical simulation, LLMs enhance traditional computational methods by providing inter-

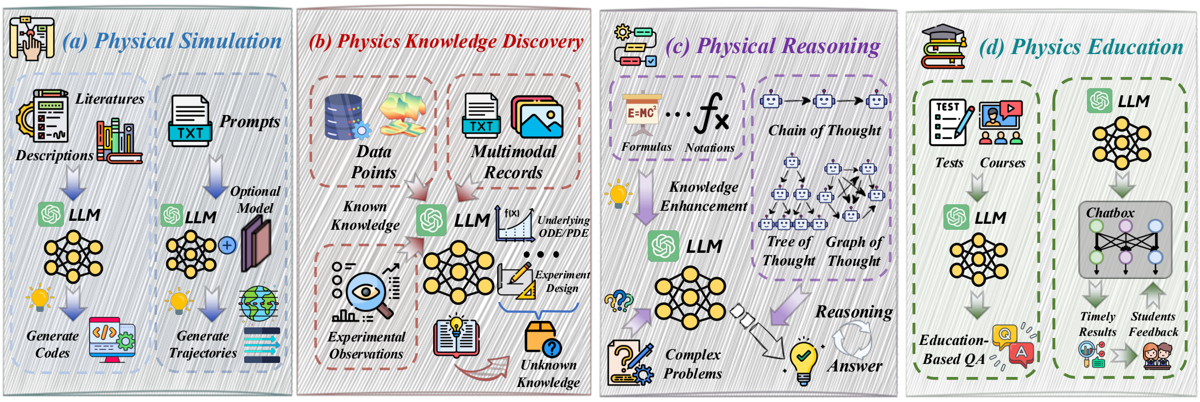

Figure 1: In this work, we focus on four mainstream areas of LLM applications in physics.

pretable guidance and improving accuracy. For knowledge discovery, LLMs act as domain experts, helping identify underlying physical laws and guiding experimental design. Physical reasoning leverages LLMs' ability to combine theoretical knowledge with logical inference, while physics education applications exploit their capacity for generating explanations and interactive learning environments (Gupta, 2023; Wang et al., 2024d).

The rapid growth of research in this field (Meadows et al., 2024; Kumar et al., 2023; Pan et al., 2025) has brought an urgent need for a comprehensive overview which summarizes existing literature and provides significant insights for future studies at the intersection of LLMs and physics. While there are a few surveys on LLMs for scientific research (Zhang et al., 2024a,b; Luo et al., 2025b; Yan et al., 2025), they have primarily focused on biology, chemistry, and mathematics, which leaves a significant gap in physics. This survey addresses that gap by providing a systematic overview of how LLMs can advance the physical domain.

In addition, this survey introduces a novel taxonomy based on how LLMs are utilized, which classifies existing works into four groups, i.e., as generic encoders, language generators, auxiliary modules, and autonomous agents. When used as generic encoders, LLMs leverage their strong representation learning capabilities to extract features for inputs (Ren et al., 2024). As language generators, these works produce responses following the given instructions (Zeng et al., 2023). When serving as auxiliary modules, LLMs work alongside traditional non-LLM systems to solve physics problems collaboratively (Yang et al., 2023). As autonomous agents, LLMs use their reasoning abilities to interact with external tools (Huang et al., 2024a) and solve physics problems automatically (Xu et al., 2024a). We organize our analysis by four areas in physics and summarize important insights of relevant research, followed by popular datasets and benchmarks. Lastly, we identify several key challenges when applying LLMs to physics and corresponding promising future directions.

In summary, the contribution of the paper is three-folds: (1) *Comprehensive Review*. We present the first comprehensive survey of LLMs in physics research, which provides a thorough overview of recent literature. (2) *Novel Taxonomy*. We introduce a novel taxonomy of current research based on how LLMs are utilized, which offers a clear framework for understanding this field. (3) *Future Guidance*. We present important challenges and opportunities in this field as a guidance for future research in LLM applications in physics.

2 Overview of Survey

2.1 Emerging LLM Capabilities for Physics Applications

The potential of LLMs in physics stems from their unique emerging capabilities (Wei et al., 2022):

Advanced Reasoning Framework (Huang and Chang, 2022; Wang et al., 2023). LLMs support various sophisticated reasoning techniques, including chain-of-thought and least-to-most prompting, enabling them to tackle complex physics problems through structured, multi-step approaches. This capability is essential for solving intricate physics problems that require careful consideration of multiple principles and constraints.

Instruction Following (Zeng et al., 2023; Yin et al., 2023). Their ability to accurately follow detailed instructions enables automated handling of various physics tasks, from generating computational

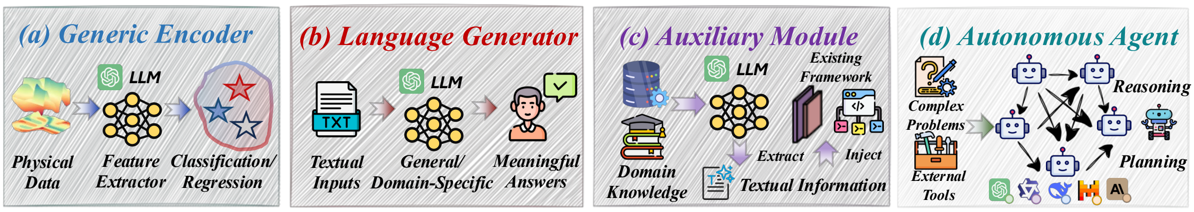

Figure 2: We propose a taxonomy of recent works on LLMs for physics, which divides them into four categories.

code to deriving mathematical formulas and analyzing experimental data. This capability streamlines many routine tasks of physics research while maintaining rigorous accuracy.

Knowledge Transfer and Generalization (Liu et al., 2024a). LLMs demonstrate exceptional ability to transfer knowledge across related domains, a crucial capability in physics where insights from one field often inform developments in another. This generalization ability can accelerate hypothesis generation and theoretical developments.

Autonomous Research Planning (Boiko et al., 2023). Through their capability to decompose complex problems and design experimental approaches, LLMs can function as autonomous research assistants. They can plan and execute sophisticated physics simulations, manage experimental workflows, and analyze results, particularly valuable in fields like fluid dynamics and particle physics.

2.2 Methodology Taxonomy

In this survey, we divide current applications of LLMs in physics into four main categories based on how they utilize LLMs (Figure 2):

LLM as Generic Encoder. LLMs demonstrate strong representation learning capabilities with strong generalization due to their massive parameters (Ren et al., 2024; Cai et al., 2024; Bogdanov et al., 2024). This category employs LLMs as feature extractors of generic input to generate outputs in the required formats for both classification and regression tasks in physics. These models are typically trained in an end-to-end manner.

LLM as Language Generator. The most direct application of LLMs is their ability to generate meaningful responses to textual inputs (Wei et al., 2022). This category either uses general-purpose commercial LLMs such as GPT and LLaMA series, or enhances their physics domain knowledge through fine-tuning on domain-specific corpus.

LLM as Auxiliary Module. A range of physics problems do not require text outputs, but they can still benefit from textual information such as knowledge databases and human guidance (Zhou et al., 2024a). This category uses LLMs to interpret and incorporate such information into existing computational frameworks.

LLM as Autonomous Agent. LLMs exhibit strong capabilities in autonomous planning for complicated tasks such as knowledge discovery (Mudur et al., 2024; Du et al., 2024). After planning, these approaches typically enable LLMs to interact with external tools such as servers and software, and autonomously analyze their outputs to generate final solutions. We classify LLM applications into these four categories in Appendix C.

2.3 Applications & Organization

In this survey, we focus on four principal physical areas including physical simulation (Sec. 3), physics knowledge discovery (Sec. 4), physical reasoning (Sec. 5), and physics education (Sec. 6), which aligns with our organization. For every area, we provide more detailed categorization for a clear understanding (see Figure 3). Afterwards, we will summarize the benchmark and datasets in the domain in Sec. 7. Finally, we will point out the challenges in this area and provide several future directions as a guidance in Sec. 8.

3 LLMs for Physical Simulation (Table 1)

Physical simulation aims to infer the behavior of physical systems according to established rules such as conservation laws (Leyli-Abadi et al., 2022). A direct solution is to use computational software (Dickinson et al., 2014) with numerical solvers to generate trajectories. Alternatively, data-driven approaches (Huang et al., 2024b) usually train deep networks such as graph neural networks to model system dynamics in the hidden space for trajectory generation. Accordingly, current works on LLMs for simulation can be roughly divided into code generation approaches and trajectory generation approaches (Wang et al., 2024a).

Code Generation Approaches. Following the first line, these approaches aim to achieve automatic

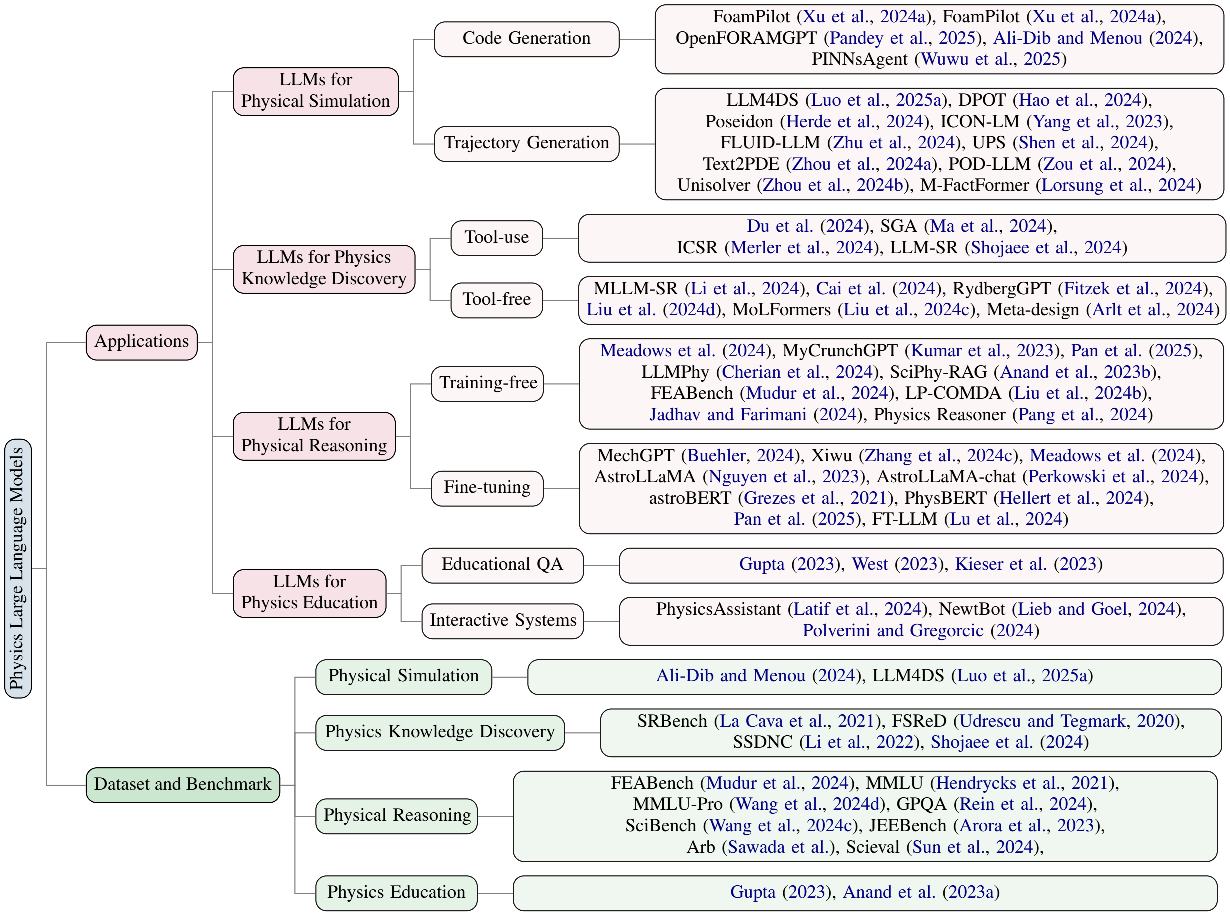

Figure 3: An overview of the taxonomy of LLM applications in physics.

code generation with human instruction (Pandey et al., 2025). For example, Ali-Dib and Menou (2024) have evaluated the performance of GPT-4 in generating simulation codes in physics-related domains. It has been found that in most cases, GPT-4 cannot solve the problem with their codes with extensive errors and unnecessary lines. LLM agents (Wang et al., 2024b) are popular in this line due to their ability to execute the code and analyze the feedback. FoamPilot (Xu et al., 2024a) utilizes retrieval-augmented generation (RAG) to build an agent framework for fire dynamics simulation, which enhances the understanding of the FireFOAM code and then generates proper configurations for source code based on users' requests. OpenFORAMGPT (Pandey et al., 2025) is also an LLM agent for fluid dynamics, which leverages the strong GPT model O1 for better performance. It also follows the RAG procedure to integrate domain knowledge, which can greatly facilitate engineering efforts. PINNsAgent (Wuwu et al., 2025) is another agent framework that utilizes LLMs to identify the best configuration for PDE solving. It incorporates both characteristics of PDEs and a tree-based search strategy for automatic PDE solving without human heuristics.

Trajectory Generation Approaches. The second line aims to directly generate the trajectories either in the text form (Luo et al., 2025a; Gruver et al., 2023; Xu et al., 2024b) or by collaborating with the other data-driven models (Zhou et al., 2024b,a; Lorsung et al., 2024; Zou et al., 2024). LLM4DS (Luo et al., 2025a) systematically utilizes prompt engineering to describe the states of dynamical systems with interactions considered and then generate the future prediction in an auto-regressive fashion. It builds a benchmark to demonstrate the potential of LLMs in dynamical system modeling. To enhance the performance, several works train Transformer-based models with massive data (Hao et al., 2024; Herde et al., 2024), resulting in general-purpose foundation models for PDEs. Another solution is to incorporate textual guidance into non-LLM data-driven models. ICON-LM (Yang et al., 2023) makes the attempt by incorporating extended text descriptions and trains the model to derive numerical predictions from both the input data and accompanying captions. Unisolver (Zhou et al., 2024b) leverages the strengths of both data-driven and physics-informed approaches, enhancing the generalization ability of LLMs across PDE scenarios by conditioning on comprehensive physical information. M-FactFormer (Lorsung et al., 2024) enhances the capabilities of LLMs in PDE surrogate modeling by integrating textual information into neural operators. FLUID-LLM (Zhu et al., 2024) projects multiple snapshots into spatio-temporal signals, which are then fed into pre-trained LLMs along with a decoder to output the future predictions. UPS (Shen et al., 2024) proposes an FNO-Transformer architecture that leverages pre-trained LLMs to warmup the Transformer model and employs explicit alignment strategies to mitigate the modality gap. Text2PDE (Zhou et al., 2024a) is a diffusion model for physics simulation, which includes a prompt of text-based instruction including physical phenomenon description to guide the generation process. POD-LLM (Zou et al., 2024) aligns spatio-temporal signals after orthogonal decomposition and text-based prompt data by patch reprogramming, and then adopts frozen LLMs and trainable head to generate future trajectories. However, dissenting voices still persist in recent works. For instance, DASHA (Xu et al., 2024b) argues that it is consistently possible to train simple supervised models that can match or even outperform the latest foundation models.

4 LLMs for Physics Knowledge Discovery (Table 2)

Physics knowledge discovery aims to identify unknown principles and laws such as PDEs based on experimental observations in physical science. Previous works usually incorporate physics-informed neural networks (PINN) (Stephany and Earls, 2024; Chen et al., 2021; Stephany and Earls, 2022) with regression methods to recover the underlying laws. However, these approaches usually require complicated optimization calculations and efforts of domain experts (Stephany and Earls, 2024). In contrast, LLMs can achieve autonomous knowledge discovery with strong reasoning and planning skills (Du et al., 2024). Recent approaches can be divided into two groups based on whether they interacted with external tools, i.e., tool-use approaches and tool-free approaches.

Tool-use Approaches. Physics knowledge discovery can be understood as a search problem with alternative proposals and evaluation in an LLM agentic framework. In particular, they usually utilize LLMs to generate several potential proposals based on prompts about domain knowledge and previous trajectories. Then, they execute external tools such as simulation software and source codes to validate the guess with feedback alternatively. For example, (Du et al., 2024) utilize the reasoning capacity of LLMs to achieve automatic equation discovery by combining genetic algorithms and score-based optimization, which accelerates the search process of PDEs and ODEs. SGA (Ma et al., 2024) adopts a bi-level framework where LLMs put forward hypotheses based on observation while simulation is done to provide feedback as guidance. SGA has been evaluated on both constitutive law discovery and molecule design. ICSR (Merler et al., 2024) utilizes LLMs to refine the skeletons based on the fitness scores for symbolic regression as an optimization loop and utilize optimization methods for coefficients. LLM-SR (Shojaee et al., 2024) further proposes to utilize LLMs to generate source codes based on the equation skeletons, which can evaluate the hypotheses automatically.

Tool-free Approaches. An alternative solution is to build a map between the input and target values in a learnable manner. Here, the input can be of any form and these approaches utilize the Transform-based architecture due to its effectiveness. (Cai et al., 2024) utilize the Transformer architecture to predict the integer coefficients with the consideration of highly complicated relationships across different terms in theoretical high-energy physics. RydbergGPT (Fitzek et al., 2024) also follows the Transformer architecture with the input of interacting Hamiltonian, which outputs the qubit measurement probabilities in quantum physics. Several approaches leverage LLMs to directly output source codes for knowledge discovery. Meta-design (Arlt et al., 2024) utilizes LLMs to generate Python codes for a wide range of quantum states, and it adopts abundant synthetic data to train the LLM for the generalized ability of scientific discovery in physics. (Liu et al., 2024d) generate simulation data using the fire simulation toolkit, which is utilized to fine-tune the popular chemical language model MolLFormer. After training, MoLFormers are adopted to predict the target properties with physical prior induced. (Liu et al., 2024d) directly input the observation of a dynamical system into the LLM, but leverage LLM's probabilistic output instead of the text output to discover the evolution laws based on the Markov processes. These approaches froze the parameters of LLMs and utilize the reasoning ability of LLMs with in-context learning for knowledge discovery. In comparison, MLLM-SR (Li et al., 2024) is a multi-modal framework, where a branch is involved in analyzing the observation data such as images and videos, and another branch is adopted to provide requirements. The whole framework is trained with instruction tuning (Zhang et al., 2023; Peng et al., 2023).

5 LLMs for Physical Reasoning (Table 3)

Physical reasoning (Meadows et al., 2024; Kumar et al., 2023) refers to solving complicated research-oriented physics tasks and answering questions with necessary calculations. As LLMs have a strong reasoning ability for providing accurate answers and solutions with analysis in various domains (Wei et al., 2022), they can be applied to the physical domain as well. Current research can be generally divided into training-free approaches and fine-tuning approaches according to whether LLMs are trained in the domain-specific corpus.

Training-free Approaches. The first line of research is close to evaluation, which prepares physics-related question-answering (QA) datasets and instructions and feeds them to commercial LLMs. For example, Meadows et al. (2024) build a carefully designed dataset with extensive notations in the domain of physics and have validated the limit of current LLMs in understanding physics content. MyCrunchGPT (Kumar et al., 2023) is a scientific machine learning platform, which is adopted from ChatGPT to enhance the applicabilities for users. The platform demonstrates strong ability in handling examples on fluid mechanics. Pan et al. (2025) evaluate the calculation performance of LLMs in quantum physics. With the enhancement of correction steps, GPT-4 can effectively obtain the final Hartree-Fock Hamiltonian in a range of cases, which demonstrates the strong potential of LLMs in quantum many-body physics. SciPhy-RAG (Anand et al., 2023b) leverages the retrieval-augmentation generation module to enhance the question-answering ability of LLMs. They also fine-tune the LLMs with instruction tuning, which achieves enhanced performance on physical reasoning benchmarks. Similarly, several models further utilize agent frameworks enhanced with external tools for complicated tasks such as code generation. LLMPhy (Cherian et al., 2024) iteratively generates source codes to infer the important physics attributes and layout parameters from a series of given observations and provide feedback using a physical simulator. After inferring the physics model, LLMPhy can widely solve a wide range of reasoning questions such as predicting steady-state poses. Mudur et al. (2024) build LLM agents which can interact with physics tools including simulation software COMSOL Multiphysics, and thus effectively solve physics reasoning problems. LP-COMDA (Liu et al., 2024b) is a physics-information LLM agent for power converter modulation design. The agent can analyze the requirements from humans and then interact with a physics-informed surrogate model for optimal results. Physics Reasoner (Pang et al., 2024) is an agentic framework for physical reasoning, which consists of three agents for problem analysis, formula retrieval and guided reasoning, respectively. With the enhanced reasoner focused on formula understanding, it can generate proper source codes for execution.

Fine-tuning Approaches. The second line is to fine-tune LLMs using domain-specific datasets for better understanding of physics fields. MechGPT (Buehler, 2024) is a fine-tuned LLM using a domain-specific mechanics and materials dataset of question-answer pairs, which can effectively solve a range of tasks including knowledge retrieval and creative applications. Xiwu (Zhang et al., 2024c) is an LLM applied to high energy physics, which is trained on a carefully designed dataset from effective collection and cleaning tools. This customized LLM can outperform the strong GPT-4 on code generation and question answering in the field of high energy physics. Grezes et al. introduce astroBERT (Grezes et al., 2021), which is trained on a huge dataset of astronomy papers in recent years. The authors have further developed an entity recognition tool to further enrich the astronomy dataset. PhysBERT (Hellert et al., 2024) is a text embedding model based on BERT, which is trained on a huge dataset of over 100,000 physics publications and has shown superior ability of physics-related problem solving. AstroLLaMA (Nguyen et al., 2023) is a foundation model trained on the corpus of astronomy containing over 300,000 abstracts of publications, which has shown state-of-the-art performance on paper summarization. AstroLLaMA-chat (Perkowski et al., 2024) is a chatbox version based on AstroLLaMA, which can greatly facilitate research in the domain of astronomy. Jadhav and Farimani (2024) combine LLMs and the finite element method to generate mechanical design automatically. The finite element method can provide the feedback of the current design while an LLM agent can improve the design based on the feedback. Lu et al. (2024) fine-tune the LLMs using a dataset of metasurface geometry. This work validates that LLMs can achieve lower error compared with traditional machine learning methods with the potential of detecting hidden patterns in the data.

6 LLMs for Physics Education (Table 4)

Compared with research-oriented physical reasoning, physics education primarily focuses on answering educational questions (Kieser et al., 2023; Lu et al., 2024) and developing interactive systems (Latif et al., 2024), which we will introduce separately. These approaches usually adopt the commercial LLMs including ChatGPT due to the interactive characteristics.

Educational Question Answering (QA). These works usually adopt commercial LLMs to solve the reasoning and concept problems in physics education. In addition to QA on force concept inventory, Kieser et al. (2023) use ChatGPT to simulate comprehension as well as preconceptions from different students. West (2023) demonstrate that GPT-4 can achieve promising grades in introductory physics courses, which bring in performance increment compared with GPT-3.5. Lu et al. (2024) fine-tune the LLMs using a dataset of metasurface geometry. This work validates that LLMs can achieve lower error compared with traditional machine learning methods with the potential of detecting hidden patterns in the data. Pranav Gupta (Gupta, 2023) has explored the performance of GPT-4 and GPT-3.5 on Physics GRE and found that LLMs have difficulty in generating accurate answers.

Interactive Systems. These works focus on enhancing the teaching experiment with advanced robot systems and chatboxs. For example, PhysicsAssistant (Latif et al., 2024) is a robot system built on YOLOv8 and GPT-3.5-turbo for K-12 physics education. Their experiments have found that the proposed system can achieve comparable performance compared with GPT-4 but with high efficiency. NewtBot (Lieb and Goel, 2024) is a physics education LLM-based chatbox that can serve as a personalized tutor to release the burden of secondary teachers. Students have been found to have a better experience than the standard GPT model with personalized feedback input. Polverini and Gregorcic (2024) demonstrate a series of examples to emphasize the importance of prompt techniques on LLMs and how to maximize the functionality of LLMs for physics education.

7 Datasets and Benchmarks (Table 5)

In this section, we briefly introduce the benchmarks and datasets in these aforementioned four tasks, which can provide a guidance to facilitate researchers in this area.

Physical Simulation. Ali-Dib and Menou (2024) propose a benchmark to evaluate the ability of LLMs to solve complex physics problems that require computational simulations. They test LLMs on PhD-level to research-level tasks in physics, using widely used simulation tools such as REBOUND (celestial mechanics), MESA (stellar physics), Dedalus (1D fluid dynamics), and SciPy (non-linear dynamics). They construct 50 original problems, avoiding common textbook examples to ensure that LLMs must generalize beyond memorized training data. The study evaluates the performance of LLMs based on correctness in coding, physics reasoning, necessity, and sufficiency of the generated solutions. Luo et al. (2025a) establishes a comprehensive benchmark *LLM4DS* to evaluate LLM's performance across nine datasets on dynamical system modeling. The benchmark includes two tasks, i.e., dynamic forecasting and relational reasoning.

Physics Knowledge Discovery. LLM applications in physics discovery (Grayeli et al., 2024) are evaluated on symbolic regression benchmarks including *SRBench* (La Cava et al., 2021) and *FSReD* (Udrescu and Tegmark, 2020). For example, *SSDNC* (Li et al., 2022) is a test set specifically designed to evaluate how well symbolic regression models handle variations in numerical constants while retaining the same underlying expression structure "skeleton". Shojaee et al. (2024) introduce a new benchmark for scientific equation discovery that spans multiple non-trivial domains, namely, nonlinear oscillators, bacterial growth models, and real-world material stress–strain data. These benchmark problems are deliberately designed so that LLMs cannot merely rely on memorized standard physics or biology equations such as textbook formulas. Instead, they require genuine reasoning and inference from the data.

Physical Reasoning. *FEABench* (Mudur et al., 2024) is a benchmark designed to evaluate LLMs and LLM agents on their ability to solve physics problems with finite element analysis (FEA). The study focuses on whether LLMs can reason through natural language descriptions of problems to generate API calls, and iteratively improve solutions. It includes two datasets: *FEABench Gold* consists of 15 manually verified solvable problems, that span across different physics domains such as heat transfer, electromagnetism, and quantum mechanics; *FEABench Large* consists of 200 algorithmically extracted problems. There are also several scientific LLM benchmarks which include physics reasoning subsets. For example, popular LLM evaluation benchmarks such as *MMLU* (Hendrycks et al., 2021), *MMLU-Pro* (Wang et al., 2024d), and *GPQA* (Rein et al., 2024) all contain a subsection of multiple choice physics questions that require physics knowledge and reasoning skills. *SciBench* (Wang et al., 2024c) includes a subset of free-response physics problems extracted from fundamentals of physics, statistical thermodynamics, and classical dynamics of particles and systems. Similarly, *JEEBench* (Arora et al., 2023) contains several free-response physics questions and their corresponding detailed solutions. *Arb* (Sawada et al.) and *Scieval* (Sun et al., 2024) are another two general scientific reasoning benchmarks that include a group of physics questions.

Physics Education. Gupta (2023) uses an actual physics GRE test consisting of 100 multiple choice questions across nine major physics topics including classical mechanics, electromagnetism, quantum mechanics. Each question is presented to an LLM as an image snippet without additional text or instructions, and the LLM responds with one of the five options. A penalized scoring is applied for incorrect answers. Anand et al. (2023a) introduces a novel dataset derived from NCERT exemplar solutions to explore the ability of LLMs in solving domain-specific high school physics problems. Initially containing 766 questions with LaTeX-based representations, the dataset was significantly expanded to 7,983 questions through advanced techniques, broadening its diversity and coverage.

8 Challenges and Future Directions

Despite the great progress, we summarize three important challenges and potential future directions in recent LLMs for physics:

Numerical Data. Due to the next-token prediction mechanism, LLMs could consider each digit separately (Requeima et al., 2024; Gruver et al., 2023; Wang et al., 2024e). Therefore, their ability of understanding complicated numbers is quite limited, which deteriorates their performance in physical simulations. Towards this end, a potential solution is to enhance the understanding of numerical data for effective physical calculation.

Generalization Across Multiple Domains. Commerical LLMs are usually trained on general knowledge datasets, which could limit the reasoning ability when it comes to specific domains (Hu et al., 2023). Note that physics consists of a range of subdomains including electromagnetism, astrophysics and quantum mechanics (Duque, 2024), which are quite infrequent in general corpus. Barman et al. (2025) point out that we should have actively fine-tuned LLMs in the physical domains rather than believing in universal LLMs such as GPT series. However, fine-tuning LLMs requires extensive efforts for data collection and computation. Therefore, an efficient framework for adapting LLMs to specific physical domains is highly expected.

Hallucination. LLMs could generate plausible but incorrect physics explanations during reasoning, especially when it comes to new domains (Ji et al., 2023; Yao et al., 2023). This hallucination comes from the fact that current LLMs follow the paradigm of pattern recognition instead of human-like understanding, which would damage the reliability of LLMs. In future works, researchers need to carefully build trustworthy LLMs, which could be achieved by introducing verification mechanisms with domain knowledge and external tools.

9 Conclusion

In this work, we present a comprehensive survey of LLMs for physics, which involves four mainstream physical tasks, i.e., physical simulation, physics knowledge discovery, physical reasoning and physical education. We further provide a novel taxonomy of current works based on how they leverage LLMs for physical problems. Besides, we introduce the current benchmark datasets to facilitate researchers. In the end, we provide challenges of current research and potential future directions. In summary, our work provides the first systematic review of current progress in LLMs for physics, which can serve as a roadmap for researchers in the fields of LLM applications and physics.

10 Limitations

This paper mainly covers LLM applications for physics. We also notice that there are several works of LLM for material discovery which is highly related to physics knowledge discovery, which we do not include in our survey. In the future, we will expand our survey with more advanced applications in these areas to provide more comprehensive insights for researchers in this domain.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mohamad Ali-Dib and Kristen Menou. 2024. Physics simulation capabilities of llms. *Physica Scripta*, 99(11):116003.

Avinash Anand, Krishnasai Addala, Kabir Baghel, Arnav Goel, Medha Hira, Rushali Gupta, and Rajiv Ratn Shah. 2023a. Revolutionizing high school physics education: A novel dataset. In *International Conference on Big Data Analytics*, pages 64–79. Springer.

Avinash Anand, Arnav Goel, Medha Hira, Snehal Buldeo, Jatin Kumar, Astha Verma, Rushali Gupta, and Rajiv Ratn Shah. 2023b. Sciphyrag-retrieval augmentation to improve llms on physics q &a. In *International Conference on Big Data Analytics*, pages 50–63. Springer.

Sören Arlt, Haonan Duan, Felix Li, Sang Michael Xie, Yuhuai Wu, and Mario Krenn. 2024. Meta-designing quantum experiments with language models. *arXiv preprint arXiv:2406.02470*.

Daman Arora, Himanshu Singh, and Mausam. 2023. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, Singapore. Association for Computational Linguistics.

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. 2025. Large physics models: Towards a collaborative approach with large language models and foundation models. *arXiv preprint arXiv:2501.05382*.

Amber Boehnlein, Markus Diefenthaler, Nobuo Sato, Malachi Schram, Veronique Ziegler, Cristiano Fanelli, Morten Hjorth-Jensen, Tanja Horn, Michelle P Kuchera, Dean Lee, et al. 2022. Colloquium: Machine learning in nuclear physics. *Reviews of modern physics*, 94(3):031003.

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. Nuner: Entity recognition encoder pretraining via llm-annotated data. *arXiv preprint arXiv:2402.15343*.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Markus J Buehler. 2024. Mechgpt, a language-based strategy for mechanics and materials modeling that connects knowledge across scales, disciplines, and modalities. *Applied Mechanics Reviews*, 76(2):021001.

Tianji Cai, Garrett W Merz, François Charton, Niklas Nolte, Matthias Wilhelm, Kyle Cranmer, and Lance J Dixon. 2024. Transforming the bootstrap: Using transformers to compute scattering amplitudes in planar n= 4 super yang-mills theory. *arXiv preprint arXiv:2405.06107*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Zhao Chen, Yang Liu, and Hao Sun. 2021. Physics-informed learning of governing equations from scarce data. *Nature communications*, 12(1):6136.

Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. 2024. Llmphy: Complex physical reasoning using large language models and world models. *arXiv preprint arXiv:2411.08027*.

Junwoo Cho, Seungtae Nam, Hyunmo Yang, Seok-Bae Yun, Youngjoon Hong, and Eunbyung Park. 2023. Separable physics-informed neural networks. *Advances in Neural Information Processing Systems*, 36:23761–23788.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Edmund JF Dickinson, Henrik Ekström, and Ed Fontes. 2014. Comsol multiphysics®: Finite element software for electrochemical analysis. a mini-review. *Electrochemistry communications*, 40:71–74.

Mengge Du, Yuntian Chen, Zhongzheng Wang, Longfeng Nie, and Dongxiao Zhang. 2024. Large language models for automatic equation discovery of nonlinear dynamics. *Physics of Fluids*, 36(9).

Erick I Duque. 2024. Emergent electromagnetism. *Physical Review D*, 110(12):125006.

David Fitzek, Yi Hong Teoh, Hin Pok Fung, Gebremedhin A Dagnew, Ejaaz Merali, M Schuyler Moss, Benjamin MacLellan, and Roger G Melko. 2024. Rydberggpt. *arXiv preprint arXiv:2405.21052*.

José-Enrique García-Ramos, Álvaro Sáiz, José M Arias, Lucas Lamata, and Pedro Pérez-Fernández. 2024. Nuclear physics in the era of quantum computing and quantum machine learning. *Advanced Quantum Technologies*, page 2300219.

Arya Grayeli, Atharva Sehgal, Omar Costilla-Reyes, Miles Cranmer, and Swarat Chaudhuri. 2024. Symbolic regression with a learned concept library. *arXiv preprint arXiv:2409.09359*.

Felix Grezes, Sergi Blanco-Cuaresma, Alberto Accomazzi, Michael J Kurtz, Golnaz Shapurian, Edwin Henneken, Carolyn S Grant, Donna M Thompson, Roman Chyla, Stephen McDonald, et al. 2021. Building astrobert, a language model for astronomy & astrophysics. *arXiv preprint arXiv:2112.00590*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36:19622–19635.

Wen Guan, Gabriel Perdue, Arthur Pesah, Maria Schuld, Koji Terashi, Sofia Vallecorsa, and Jean-Roch Vlimant. 2021. Quantum machine learning in high energy physics. *Machine Learning: Science and Technology*, 2(1):011003.

Pranav Gupta. 2023. Testing llm performance on the physics gre: some observations. *arXiv preprint arXiv:2312.04613*.

Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. 2024. Dpot: Auto-regressive denoising operator transformer for large-scale pde pre-training. *arXiv preprint arXiv:2403.03542*.

Thorsten Hellert, João Montenegro, and Andrea Pollastro. 2024. Physbert: A text embedding model for physics scientific literature. *APL Machine Learning*, 2(4).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. 2024. Poseidon: Efficient foundation models for pdes. *arXiv preprint arXiv:2405.19101*.

Tony Hey, Keith Butler, Sam Jackson, and Jeyarajan Thiyagalingam. 2020. Machine learning and big scientific data. *Philosophical Transactions of the Royal Society A*, 378(2166):20190054.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024a. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*.

Zijie Huang, Wanjia Zhao, Jingdong Gao, Ziniu Hu, Xiao Luo, Yadi Cao, Yuanzhou Chen, Yizhou Sun, and Wei Wang. 2024b. Physics-informed regularization for domain-agnostic dynamical system modeling. *arXiv preprint arXiv:2410.06366*.

Yayati Jadhav and Amir Barati Farimani. 2024. Large language model agent as a mechanical designer. *arXiv preprint arXiv:2404.17525*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Georgia Karagiorgi, Gregor Kasieczka, Scott Kravitz, Benjamin Nachman, and David Shih. 2022. Machine learning in the search for new fundamental physics. *Nature Reviews Physics*, 4(6):399–412.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440.

Karthik Kashinath, M Mustafa, Adrian Albert, JL Wu, C Jiang, Soheil Esmaeilzadeh, Kamyar Azizzadenesheli, R Wang, Ashesh Chattopadhyay, A Singh, et al. 2021. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194):20200093.

Fabian Kieser, Peter Wulff, Jochen Kuhn, and Stefan Küchemann. 2023. Educational data augmentation in physics education research using chatgpt. *Physical Review Physics Education Research*, 19(2):020150.

Varun Kumar, Leonard Gleyzer, Adar Kahana, Khemraj Shukla, and George Em Karniadakis. 2023. Mycrunchgpt: A llm assisted framework for scientific machine learning. *Journal of Machine Learning for Modeling and Computing*, 4(4).

William La Cava, Bogdan Burlacu, Marco Virgolin, Michael Kommenda, Patryk Orzechowski, Fabrício Olivetti de França, Ying Jin, and Jason H Moore. 2021. Contemporary symbolic regression methods and their relative performance. *Advances in neural information processing systems*, 2021(DB1):1.

Ehsan Latif, Ramviyas Parasuraman, and Xiaoming Zhai. 2024. Physicsassistant: An llm-powered interactive learning robot for physics lab investigations. *arXiv preprint arXiv:2403.18721*.

Milad Leyli-Abadi, Antoine Marot, Jérôme Picault, David Danan, Mouadh Yagoubi, Benjamin Donnot, Seif Attoui, Pavel Dimitrov, Asma Farjallah, and Clement Etienam. 2022. Lips-learning industrial physical simulation benchmark suite. *Advances in Neural Information Processing Systems*, 35:28095–28109.

Wenqiang Li, Weijun Li, Linjun Sun, Min Wu, Lina Yu, Jingyi Liu, Yanjie Li, and Songsong Tian. 2022. Transformer-based model for symbolic regression via joint supervised learning. In *The Eleventh International Conference on Learning Representations*.

Yanjie Li, Weijun Li, Lina Yu, Min Wu, Jingyi Liu, Wenqiang Li, Shu Wei, and Yusong Deng. 2024. Mllm-sr: Conversational symbolic regression base multi-modal large language models. *arXiv preprint arXiv:2406.05410*.

Guojun Liang, Prayag Tiwari, Sławomir Nowaczyk, and Stefan Byttner. 2024. Higher-order spatio-temporal physics-incorporated graph neural network for multivariate time series imputation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1356–1366.

Anna Lieb and Toshali Goel. 2024. Student interaction with newtbot: An llm-as-tutor chatbot for secondary physics education. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Chaoqun Liu, Qin Chao, Wenxuan Zhang, Xiaobao Wu, Boyang Li, Anh Tuan Luu, and Lidong Bing. 2024a. Zero-to-strong generalization: Eliciting strong capabilities of large language models iteratively without gold labels. *arXiv preprint arXiv:2409.12425*.

Junhua Liu, Fanfan Lin, Xinze Li, Kwan Hui Lim, and Shuai Zhao. 2024b. Physics-informed llm-agent for automated modulation design in power electronics systems. *arXiv preprint arXiv:2411.14214*.

Ning Liu, Siavash Jafarzadeh, Brian Y Lattimer, Shuna Ni, Jim Lua, and Yue Yu. 2024c. Large language models, physics-based modeling, experimental measurements: the trinity of data-scarce learning of polymer properties. *arXiv preprint arXiv:2407.02770*.

Tianyu Liu, Tianqi Chen, Wangjie Zheng, Xiao Luo, and Hongyu Zhao. 2023. scelmo: Embeddings from language models are good learners for single-cell data analysis. *bioRxiv*, pages 2023–12.

Toni JB Liu, Nicolas Boullé, Raphaël Sarfati, and Christopher J Earls. 2024d. Llms learn governing principles of dynamical systems, revealing an in-context neural scaling law. *arXiv preprint arXiv:2402.00795*.

Kenneth R Long, Donatella Lucchesi, Mark A Palmer, Nadia Pastrone, Daniel Schulte, and V Shiltsev. 2021. Muon colliders to expand frontiers of particle physics. *Nature Physics*, 17(3):289–292.

Cooper Lorsung et al. 2024. Explain like i'm five: Using LLMs to improve PDE surrogate models with text.

Darui Lu, Yang Deng, Jordan M Malof, and Willie J Padilla. 2024. Can large language models learn the physics of metamaterials? an empirical study with chatgpt. *arXiv preprint arXiv:2404.15458*.

Haibao Lu. 2024. When physics meets chemistry at the dynamic glass transition. *Reports on Progress in Physics*, 87(3):032601.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning? *arXiv preprint arXiv:2309.01809*.

Xiao Luo, Binqi Chen, Haixin Wang, Zhiping Xiao, Ming Zhang, and Yizhou Sun. 2025a. How do large language models perform in dynamical system modeling. In *NAACL Findings*.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025b. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*.

Pingchuan Ma, Tsun-Hsuan Wang, Minghao Guo, Zhiqing Sun, Joshua B Tenenbaum, Daniela Rus, Chuang Gan, and Wojciech Matusik. 2024. Llm and simulation as bilevel optimizers: A new paradigm to advance physical scientific discovery. In *Forty-first International Conference on Machine Learning*.

Andreas Mayr, Sebastian Lehner, Arno Mayrhofer, Christoph Kloss, Sepp Hochreiter, and Johannes Brandstetter. 2023. Boundary graph neural networks for 3d simulations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9099–9107.

Jordan Meadows, Tamsin James, and Andre Freitas. 2024. Exploring the limits of fine-grained llm-based physics inference via premise removal interventions. *arXiv preprint arXiv:2404.18384*.

Matteo Merler, Katsiaryna Haitsiukevich, Nicola Dainese, and Pekka Marttinen. 2024. In-context symbolic regression: Leveraging large language models for function discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 589–606.

Lars-Peter Meyer, Claus Stadler, Johannes Frey, Norman Radtke, Kurt Junghanns, Roy Meissner, Gordian Dziwis, Kirill Bulert, and Michael Martin. 2023. Llm-assisted knowledge graph engineering: Experiments with chatgpt. In *Working conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow*, pages 103–115. Springer Fachmedien Wiesbaden Wiesbaden.

Spandan Mondal and Luca Mastrolorenzo. 2024. Machine learning in high energy physics: a review of heavy-flavor jet tagging at the lhc. *The European Physical Journal Special Topics*, 233(15):2657–2686.

Nayantara Mudur, Hao Cui, Subhashini Venugopalan, Paul Raccuglia, Michael Brenner, and Peter Christian Norgaard. 2024. Feabench: Evaluating language models on real world physics reasoning ability. In *NeurIPS 2024 Workshop on Open-World Agents*.

Tuan Dung Nguyen, Yuan-Sen Ting, Ioana Ciucă, Charlie O'Neill, Ze-Chang Sun, Maja Jabłońska, Sandor Kruk, Ernest Perkowski, Jack Miller, Jason Li, et al. 2023. Astrollama: Towards specialized foundation models in astronomy. *arXiv preprint arXiv:2309.06126*.

Haining Pan, Nayantara Mudur, William Taranto, Maria Tikhanovskaya, Subhashini Venugopalan, Yasaman Bahri, Michael P Brenner, and Eun-Ah Kim. 2025. Quantum many-body physics calculations with large language models. *Communications Physics*, 8(1):49.

Sandeep Pandey, Ran Xu, Wenkang Wang, and Xu Chu. 2025. Openfoamgpt: a rag-augmented llm agent for openfoam-based computational fluid dynamics. *arXiv preprint arXiv:2501.06327*.

Xinyu Pang, Ruixin Hong, Zhanke Zhou, Fangrui Lv, Xinwei Yang, Zhilong Liang, Bo Han, and Changshui Zhang. 2024. Physics reasoner: Knowledge-augmented reasoning for solving physics problems with large language models. *arXiv preprint arXiv:2412.13791*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

David Peral-García, Juan Cruz-Benito, and Francisco José García-Peñalvo. 2024. Systematic literature review: Quantum machine learning and its applications. *Computer Science Review*, 51:100619.

Ernest Perkowski, Rui Pan, Tuan Dung Nguyen, Yuan-Sen Ting, Sandor Kruk, Tong Zhang, Charlie O'Neill, Maja Jablonska, Zechang Sun, Michael J Smith, et al. 2024. Astrollama-chat: Scaling astrollama with conversational and diverse datasets. *Research Notes of the AAS*, 8(1):7.

Giulia Polverini and Bor Gregorcic. 2024. How understanding large language models can inform the use of chatgpt in physics education. *European Journal of Physics*, 45(2):025701.

Aravind Ramakrishnan, David IW Levin, and Alec Jacobson. 2025. Rigid body adversarial attacks. *arXiv preprint arXiv:2502.05669*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM on Web Conference 2024*, pages 3464–3475.

James Requeima, John Bronskill, Dami Choi, Richard E Turner, and David Duvenaud. 2024. Llm processes: Numerical predictive distributions conditioned on natural language. *arXiv preprint arXiv:2405.12856*.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John Nay, Kshitij Gupta, and Aran Komatsuzaki. Arb: Advanced reasoning benchmark for large language models. In *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*.

Junhong Shen, Tanya Marwah, and Ameet Talwalkar. 2024. Ups: Towards foundation models for pde solving via cross-modal adaptation. *arXiv preprint arXiv:2403.07187*.

Siqi Shen, Yu Liu, Daniel Biggs, Omar Hafez, Jiandong Yu, Wentao Zhang, Bin Cui, and Jiulong Shan. 2025. Transfer learning in scalable graph neural network for improved physical simulation. *arXiv preprint arXiv:2502.06848*.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K Reddy. 2024. Llm-sr: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*.

Dong Shu, Tianle Chen, Mingyu Jin, Chong Zhang, Mengnan Du, and Yongfeng Zhang. 2024. Knowledge graph large language model (kg-llm) for link prediction. *arXiv preprint arXiv:2403.07311*.

Robert Stephany and Christopher Earls. 2022. Pde-read: Human-readable partial differential equation discovery using deep learning. *Neural Networks*, 154:360–382.

Robert Stephany and Christopher Earls. 2024. Pde-learn: Using deep learning to discover partial differential equations from noisy, limited data. *Neural Networks*, 174:106242.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. 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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Silviu-Marian Udrescu and Max Tegmark. 2020. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631.

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881.

Haixin Wang, Yadi Cao, Zijie Huang, Yuxuan Liu, Peiyan Hu, Xiao Luo, Zezheng Song, Wanjia Zhao, Jilin Liu, Jinan Sun, et al. 2024a. Recent advances on machine learning for computational fluid dynamics: A survey. *arXiv preprint arXiv:2408.12171*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024b. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024c. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 50622–50649. PMLR.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024d. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Zhenhua Wang, Guang Xu, and Ming Ren. 2024e. Llm-generated natural language meets scaling laws: New explorations and data augmentation methods. *arXiv preprint arXiv:2407.00322*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Colin G West. 2023. Advances in apparent conceptual physics reasoning in gpt-4. *arXiv preprint arXiv:2303.17012*.

Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. 2020. Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*, 1(1):1–34.

Qingpo Wuwu, Chonghan Gao, Tianyu Chen, Yihang Huang, Yuekai Zhang, Jianing Wang, Jianxin Li, Haoyi Zhou, and Shanghang Zhang. 2025. Pinnsagent: Automated pde surrogation with large language models. *arXiv preprint arXiv:2501.12053*.

Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, et al. 2024. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *bioRxiv*, pages 2024–05.

Leidong Xu, Danyal Mohaddes, and Yi Wang. 2024a. Llm agent for fire dynamics simulations. *arXiv preprint arXiv:2412.17146*.

Zongzhe Xu, Ritvik Gupta, Wenduo Cheng, Alexander Shen, Junhong Shen, Ameet Talwalkar, and Mikhail Khodak. 2024b. Specialized foundation models struggle to beat supervised baselines. *arXiv preprint arXiv:2411.02796*.

Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. 2025. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Liu Yang, Siting Liu, and Stanley J Osher. 2023. Fine-tune language models as multi-modal differential equation solvers. *arXiv preprint arXiv:2308.05061*.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Wenpeng Yin, Qinyuan Ye, Pengfei Liu, Xiang Ren, and Hinrich Schütze. 2023. Llm-driven instruction following: Progresses and concerns. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 19–25.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. 2024a. Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv preprint arXiv:2406.10833*.

Zhengde Zhang, Yiyu Zhang, Haodong Yao, Jianwen Luo, Rui Zhao, Bo Huang, Jiameng Zhao, Yipu Liao, Ke Li, Lina Zhao, et al. 2024c. Xiwu: A basis flexible and learnable llm for high energy physics. *arXiv preprint arXiv:2404.08001*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Anthony Zhou, Zijie Li, Michael Schneier, John R Buchanan Jr, and Amir Barati Farimani. 2024a. Text2pde: Latent diffusion models for accessible physics simulation. *arXiv preprint arXiv:2410.01153*.

Hang Zhou, Yuezhou Ma, Haixu Wu, Haowen Wang, and Mingsheng Long. 2024b. Unisolver: Pde-conditional transformers are universal pde solvers. *arXiv preprint arXiv:2405.17527*.

Max Zhu, Adrián Bazaga, and Pietro Liò. 2024. Fluid-llm: Learning computational fluid dynamics with spatiotemporal-aware large language models. *arXiv preprint arXiv:2406.04501*.

Weihao Zou, Weibing Feng, Pin Wu, and Jiangnan Wu. 2024. Method for rapid prediction of flow fields based on large language models. *Available at SSRN 5019019*.

A Background

A.1 Large Language Models for Scientific Research

With the rise of scientific machine learning (Hey et al., 2020; Chang et al., 2024), LLMs have received increasing attention in scientific research including biology, chemistry, and medicine (Zhang et al., 2024a,b; Luo et al., 2025b; Yan et al., 2025). For example, in the biological domain, LLMs have been leveraged for single-cell (Xiao et al., 2024; Liu et al., 2023) and protein analysis, which help researchers answer complex questions and extract deep embeddings from textual inputs. LLMs have also made significant contributions to the physical domain, especially in simulation and education (Huang et al., 2024a; Latif et al., 2024). Despite their growing impact, there has not yet been a comprehensive survey of LLMs for physics. This paper addresses this gap by providing the first systematic overview of the field.

A.2 Machine Learning for Physics Research

Machine learning has achieved great progress in physics research across various areas including fluid mechanics (Liang et al., 2024; Mayr et al., 2023; Kashinath et al., 2021), high-energy physics (Guan et al., 2021; Mondal and Mastrolorenzo, 2024), and quantum physics (Peral-García et al., 2024; García-Ramos et al., 2024). In fluid mechanics, data-driven approaches have accelerated simulation processes, while in high-energy physics, existing approaches have been developed to analyze particle collisions using collider data. As a powerful tool, LLMs have wide applications across four key areas, i.e., simulation (Ali-Dib and Menou, 2024), knowledge discovery (Du et al., 2024), reasoning (Meadows et al., 2024), and physics education (West, 2023). Our survey provides a comprehensive overview of existing literature in these four areas to guide future research.

B Difference from Existing Surveys

There have been several surveys related to our work on scientific LLMs (Zhang et al., 2024a,b). In particular, Zhang et al. (2024a) summarize the current progress in scientific LLMs on biological and chemical domains, which includes textual scientific models, molecular models, protein models, genomic models, and multi-modal models. Zhang et al. (2024b) provide an overview of LLMs for scientific discovery, which is mostly focused on chemistry, biology, and medicine, while only including seven works on LLMs for physics. Luo et al. (2025b) focus on how to leverage LLMs to facilitate scientific research at different stages, i.e., hypothesis, planning, writing, and reviewing. Yan et al. (2025) points out that multimodal LLMs can benefit reasoning tasks in general science domains. In summary, they almost focus on general science with an emphasis on biological and chemical domains while failing to provide an overview from the physics perspective. Compared with these surveys, we provide the first comprehensive overview of LLMs targeting at the physical domains.

C Summary of LLM Applications for Physics

We provide a detailed summary of LLM applications for physics in four areas, i.e., physical simula-

Table 1: LLMs for physical simulation.

Model	Fndn. LLM	Category	Different Modalities	Finetuned
ICON-LM (Yang et al., 2023)	ChatGPT	Auxiliary Module	✓	Small Model
Ali-Dib and Menou (2024)	GPT-4	Language Generator		
FoamPilot (Xu et al., 2024a)	GPT-4o	Autonomous Agent	✓	
FLUID-LLM (Zhu et al., 2024)	OPT-125m, OPT-2.7b	Auxiliary Module		Small Model
UPS (Shen et al., 2024)	RoBERTa	Auxiliary Module	✓	Small Model
DPOT (Hao et al., 2024)	Fourier Transformer	Generic Encoder		✓
Poseidon (Herde et al., 2024)	Transformer	Generic Encoder		✓
Text2PDE (Zhou et al., 2024a)	Claude 3.5 Sonnet	Auxiliary Module	✓	Small Model
POD-LLM (Zou et al., 2024)	Undiscovered	Auxiliary Module	✓	Small Model
Unisolver (Zhou et al., 2024b)	Llama3-8B	Auxiliary Module	✓	Small Model
M-FactFormer (Lorsung et al., 2024)	Llama3-8B	Auxiliary Module	✓	Small Model
LLM4DS (Luo et al., 2025a)	GPT-3.5, Llama3-70B	Language Generator		
OpenFORAMGPT (Pandey et al., 2025)	O1	Autonomous Agent		
PINNsAgent (Wuwu et al., 2025)	GPT-4	Autonomous Agent		

Table 2: LLMs for physics knowledge discovery.

Model	Fndn. LLM	Category	Different Modalities	Finetuned
Du et al. (2024)	GPT-3.5, GPT-4, Llama2-7B	Autonomous Agent	✓	
SGA (Ma et al., 2024)	GPT-4	Autonomous Agent	✓	
ICSR (Merler et al., 2024)	Llama3-7B	Autonomous Agent	✓	
MLLM-SR (Li et al., 2024)	Vicuna	Auxiliary Module	✓	Small Model
Cai et al. (2024)	Transformer	Generic Encoder		✓
RydbergGPT (Fitzek et al., 2024)	Transformer	Generic Encoder		✓
LLM-SR (Shojaee et al., 2024)	GPT-3.5, Mixtral-8x7B	Autonomous Agent	✓	
Meta-design (Arlt et al., 2024)	Transformer	Language Generator		✓
Liu et al. (2024d)	Llama2-70B	Generic Encoder		
MoLFormers (Liu et al., 2024c)	MolLFormer	Generic Encoder		✓

tion (Table 1), physics knowledge discovery (Table 2), physical reasoning (Table 3), and physics education (Table 4). We summarize the categories of these models based on our new taxonomy, whether models involve different modalities, and whether the model is fine-tuned.

D Summary of Datasets and Benchmarks

We provide a detailed summary of LLM for physics datasets and benchmarks in Table 5, and believe our work can serve as an important guidance for researchers in both fields of LLM applications and physics.

Table 3: LLMs for physical reasoning.

Model	Fndn. LLM	Category	Different Modalities	Finetuned
LLMPhy (Cherian et al., 2024)	GPT-o1-mini, GPT-4o VLM	Autonomous Agent	✓	
FEABench (Mudur et al., 2024)	Benchmark	Autonomous Agent	✓	✓
MechGPT (Buehler, 2024)	OpenOrca-Platypus2-13B	Language Generator		✓
Xiwu (Zhang et al., 2024c)	Vicuna-1.5	Language Generator		✓
Meadows et al. (2024)	Benchmark	Language Generator		
MyCrunchGPT (Kumar et al., 2023)	ChatGPT	Language Generator		
LP-COMDA (Liu et al., 2024b)	GPT-3.5	Autonomous Agent	✓	
AstroLLaMA (Nguyen et al., 2023)	Llama2-7B	Language Generator		✓
AstroLLaMA-chat (Perkowski et al., 2024)	Llama2-7B	Language Generator		✓
astroBERT (Grezes et al., 2021)	BERT	Language Generator		✓
PhysBERT (Hellert et al., 2024)	BERT	Language Generator		✓
(Jadhav and Farimani, 2024)	GPT-4	Autonomous Agent		
(Pan et al., 2025)	GPT-4	Language Generator		
FT-LLM (Lu et al., 2024)	GPT-3.5	Language Generator		✓
Physics Reasoner (Pang et al., 2024)	GPT-3.5, GPT-4, Llama3-70B	Autonomous Agent		

Table 4: LLMs for physics education.

Model	Fndn. LLM	Category	Different Modalities	Finetuned
Gupta (2023)	GPT-3.5, GPT-4	Language Generator		
PhysicsAssistant (Latif et al., 2024)	GPT-3.5	Autonomous Agent		
NewtBot (Lieb and Goel, 2024)	GPT-3.5	Language Generator		
Polverini and Gregorcic (2024)	ChatGPT-4	Language Generator		
Kieser et al. (2023)	ChatGPT-4	Language Generator		
(West, 2023)	GPT-4	Language Generator		
SciPhy-RAG (Anand et al., 2023b)	Vicuna-7B	Language Generator		✓

Table 5: An overview of datasets and benchmarks.

Task	Benchmark	Short Description
Physical Reasoning	FEABench (Mudur et al., 2024)	15 physics problems requiring numerical solutions via finite element analysis
	MMLU (Hendrycks et al., 2021)	Containing sub-sections on conceptual physics, high school physics, and college physics problems
	MMLU-Pro (Wang et al., 2024d)	10.8% of MMLU-Pro problems are under physics domain
	GPQA (Rein et al., 2024)	Containing 227 graduate-level multiple-choice physics problems
	SciBench (Wang et al., 2024c)	Containing 291 problems from 3 different physics textbooks
	JEEBench (Arora et al., 2023)	Containing 123 college-level physics problems
	Arb (Sawada et al.)	Containing 98 numerical physics problems and 31 symbolic physics problems
	Scieval (Sun et al., 2024)	Containing 1657 physics problem where 1165 of them are scientific calculation questions
Physical Simulation	Ali-Dib and Menou (2024)	47 problems on computational physics simulations, including celestial mechanics, stellar evolution, 1D fluid dynamics, and non-linear dynamics
	LLM4DS (Luo et al., 2025a)	9 datasets focusing on dynamic forecasting and relational reasoning
Physics Knowledge Discovery	SRBench (La Cava et al., 2021)	252 symbolic regression problems, including 122 black-box real-world problems and 130 synthetic known-form problems with ground-truth equations
	FSReD (Udrescu and Tegmark, 2020)	A symbolic regression database containing 100 "basic" physics equations and 20 additional "bonus" equations chosen for higher complexity
	SSDNC (Li et al., 2022)	A synthetic dataset with 100 symbolic expression skeletons and 10 re-sampled numeric coefficients for each skeleton
	Shojaee et al. (2024)	Modeling nonlinear oscillators, bacterial growth, and material stress behavior
Physics Education	Gupta (2023)	100 multiple choice GRE physics questions
	Anand et al. (2023a)	7,983 questions augmented from 766 NCERT school physics problems