Can Theoretical Physics Research Benefit from Language Agents?

Anonymous Author(s)

Affiliation Address email

Abstract

Large Language Models (LLMs) are rapidly advancing across diverse domains, yet their application in theoretical physics research is not yet mature. This position paper argues that LLM agents can potentially help accelerate theoretical, computational, and applied physics when properly integrated with domain knowledge and toolbox. We analyze current LLM capabilities for physics—from mathematical reasoning to code generation—identifying critical gaps in physical intuition, constraint satisfaction, and reliable reasoning. We envision future physics-specialized LLMs that could handle multimodal data, propose testable hypotheses, and design experiments. Realizing this vision requires addressing fundamental challenges: ensuring physical consistency, and developing robust verification methods. We call for collaborative efforts between physics and AI communities to help advance scientific discovery in physics.

1 Introduction

2

3

8

9

10

11 12

13

15

16

17

18

19

20

21

23 24

25

26

27

28

29

30 31

32

33

34

35

Large Language Models (LLMs) represent a major advance at the forefront of artificial intelligence (AI), exhibiting remarkable proficiency in understanding natural language and performing increasingly complex reasoning tasks [1, 2, 3, 4, 5, 6]. While impacting various sectors, their potential in fundamental scientific research is only beginning to be systematically explored [7]. Physics, with its complex blend of abstract theory, demanding computation, rigorous experimentation, and reliance on approximations and physical intuition, presents both unique challenges and fertile ground for LLM applications. Position: We argue that LLM agents, when appropriately adapted and integrated with domain-specific knowledge and toolbox, could potentially serve as a promising technology with the capacity to accelerate discovery in theoretical physics, with broader implications for computational and applied physics, provided their current limitations in rigorous reasoning, physical grounding, and reliability are systematically addressed through targeted interdisci**plinary research.** This position challenges the current paradigm where LLMs serve primarily as assistants for information retrieval. We contend that LLMs may evolve into autonomous collaborators for physicists, augmenting capabilities from literature review and conceptual exploration, to computational simulation and data interpretation. However, realizing this potential requires acknowledging current limitations and undertaking dedicated, physics-informed research efforts. Supporting this cautious optimism is recent progress in LLM architecture, scale, and particularly advances in training reasoning models [3, 6, 4, 5, 8, 9, 10] that demonstrate growing agency in multistep problem-solving needed for physics research. **Overview of this work** This position paper is structured as follows. Section 2 introduces the taxonomy used in this paper, outlining a typical physics research workflow with an overview of the subskills LLMs might assist with. Section 3 provides an in-depth analysis of LLM capabilities for physics reasoning, categorized into mathematical skills, physics-specific reasoning beyond mathematics, code generation & execution, and general research skills. Subsequently, Section 4 discusses common LLM engineering techniques relevant to physics applications. Section 5 explores open directions and desirable future capabilities for next-generation LLM-powered systems

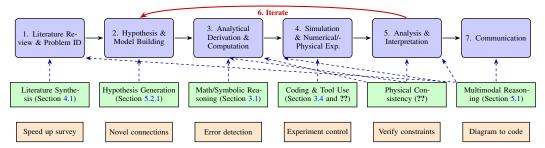


Figure 1: A schematic workflow of theoretical physics research (top row, blue), potential LLM capabilities (middle row, green), and key opportunities (bottom row, orange). Tool use capability connects with experimental research through automated instrument control and data analysis.

to better assist physics research. Appendix A discusses key risks and challenges. Finally, Section 6
 summarizes our position and offers a concluding perspective. Sections on Limitations and Ethical
 Considerations are included in the appendix.

2 Physics Research: An Overview

43 2.1 Research Stages & Skills

45

46

47

48

49

50

51 52

53

54

55

56

57

58

59

60

61

62

63 64

65 66

67

68

69

70

71

72

73

74

75

76

77

78

A typical workflow in physics research often involves several stages, as depicted in Figure 1 (top row). These stages are generally iterative and involve collaboration among various researchers with different backgrounds and skill sets. Scientific inquiry typically proceeds through an iterative workflow that begins with literature review and problem identification, where existing work is surveyed to assess the state of the art and uncover open questions or inconsistencies. Based on this foundation, researchers engage in hypothesis formulation and model building, proposing new ideas, constructing models to capture physical phenomena, and defining the assumptions that frame their scope. These models are then subjected to analytical derivation, involving mathematical analysis, symbolic reasoning, and numerical calculations to extract predictions. Complementing this, simulation and computational experiments are employed to test model behavior and guide the design of physical experiments, for instance simulating the BKT transition in a quantum XY model [11] before performing quantum optical experiments. The resulting data undergo analysis and interpretation of results, where findings are compared with prior work to generate physical insights. This process is inherently cyclical, requiring iteration of the above stages until the problem is satisfactorily addressed. Finally, the outcomes are consolidated through *communication*, including the preparation of papers and presentations to disseminate the results.

2.2 Opportunities and Challenges for LLMs in Physics Research

The intersection of AI and physics is not new [12], but the advent of powerful LLMs introduces the potential to help address persistent bottlenecks in physics research—especially for tasks demanding enormous time investment or the processing of vast information streams. LLMs might assist physics research in at least two primary modes: (1) automating repetitive tasks such as literature review and well-defined calculations (see Sections 3.1, 3.4 and 4.1), and (2) sparking new ideas through human-AI collaboration, where AI agents might provide alternative perspectives (see Sections 3.3 and 5.2.1). While LLMs have shown remarkable growth in assisting formal theorem proving [13, 14, 15] and augmenting biochemical research [16, 17], physics poses unique challenges. Unlike formal mathematics with its focus on rigorous axiom-based proof [13], theoretical physics centrally involves constructing models, making justified approximations (e.g., when to approximate $\sin \theta \approx \theta$ for small angles or apply perturbation theory where a Hamiltonian $H = H_0 + \lambda V$ is expanded in powers of a small parameter λ), seeking validation against experiments, and employing physical intuition. These represent the highly challenging task of connecting abstractions to physical reality [18, 19, 20]. Physics often involves mathematical problems that, while formally straightforward, gain complexity and nuance from their physical context, where mathematical rigor alone is insufficient. For example, diagonalizing a 2×2 matrix is a standard linear algebra task. However, in topological condensed matter physics, such a matrix might represent the Bloch Hamiltonian of a Chern insulator, $H(\mathbf{k})$ $d_x(\mathbf{k})\sigma_x + d_y(\mathbf{k})\sigma_y + d_z(\mathbf{k})\sigma_z$, where σ_i are Pauli matrices and $\mathbf{d}(\mathbf{k})$ is a vector function of momentum ${\bf k}$. The mathematical task is to find eigenvalues $E_{\pm}({\bf k})=\pm |{\bf d}({\bf k})|$ and eigenvectors. The

physics, however, lies in understanding how the winding number of the vector **d**(**k**) over the Brillouin zone determines topological invariants like the Chern number [21], which dictates phenomena like quantized Hall conductivity [22]. LLM agents assisting must go beyond mere diagonalization and try to connect the mathematical procedure with the underlying physical interpretation.

3 Skill Analysis for Physics Reasoning

Despite the emerging ecosystem of scientific reasoning benchmarks from general scientific knowledge [23, 24] to physics-specific reasoning [25, 26, 27, 28], they focus primarily on exam-like problems with one definitive answer for verification. These do not capture the full complexity of physics research involving tasks such as deriving properties of new physical models, modifying simulation code based on a paper, and interacting with experts to explore open-ended problems. We gq need more benchmarks analogous to SWE-Bench [29] on a full-cycle research workflow to gauge how LLMs perform in tackling open-ended research in a real-world scenario [30]. Furthermore, developing benchmarks from frontier research questions such as FrontierMath [31] or Humanity's Last Exam [32], within an ecosystem of domain experts [33], is key to probing the limits of AI reasoning for scientific discovery beyond solving close-ended Olympiad exam questions. In this section, we discuss concrete skills needed for physics research where limitations of current models call for focused improvement before LLMs can be reliable research partners. We categorize these skills to better understand the multifaceted potential and challenges.

3.1 Mathematical and Symbolic Reasoning

Skill Performing algebraic manipulation, calculus (differentiation, integration), linear algebra (matrix operations, tensor contractions like $T^{ijk}S_{jlm}=R^{ik}_{lm}$), and solving differential equations essential for theoretical physics. Analysis Next-token prediction inherent to LLMs can lead to cascading errors in complex mathematical operations. Despite saturation on legacy benchmarks like MATH [23], errors are frequently observed in algebra and calculus [34, 35]. They also struggle with unit consistency (e.g., mixing SI and natural units where $\hbar=c=1$), thereby raising questions about their reliability for research-level derivations (e.g., evaluating path integrals $\int \mathcal{D}\phi e^{iS[\phi]/\hbar}$) [36].

3.2 Beyond Math: Physics-Specific Reasoning Skills

We outline skills unique to understanding physical context, principles, and common practices, ordered roughly from currently more reliable to less reliable (or more complex) for LLMs.

3.2.1 Conceptual Framework, Formula Retrieval, and Application

Skill Articulating physics concepts, principles, and theories in natural language, adapting to specific notations; identifying and applying general physics formulas to well-structured problems. Analysis LLMs can generate textbook-style explanations through summarization, yet this apparent understanding can be superficially derived from statistical correlations rather than causal models of physical laws [18, 19]. This is evident in explanations that seem correct but contain subtle physical inaccuracies or miss crucial assumptions (e.g., in the context of perturbation theory, failing to state the conditions for its validity) [37]. LLMs have shown promising progress in applying formulas to well-defined problems mirroring textbook examples [27, 38, 26], but they may resort to memorized solutions rather than reasoning from first principles when confronted with novel variants of the same problems. Recent work [39, 20, 40] shows that perturbations to problem statements can cause significant performance decay, revealing models' fragile understanding of systematic solution strategies. Their ability to choose appropriate approximations or understand the domain of validity for a given formula remains limited.

3.2.2 Mathematical Deduction and Reasoning by Special Cases and Analogies

Skill Applying mathematical tools adaptively to respect the physical constraints and interpretations of variables and operations; simplifying complex problems by considering special or limiting cases, or by drawing analogies to simpler, well-understood physical systems. **Analysis** This involves applying mathematical tools correctly while respecting the physical constraints and interpretations of variables and operations. For instance, correctly applying vector calculus to electromagnetic fields requires not just knowing the formulas for divergence or curl, but understanding what these operations mean for fields, sources, and boundaries in a physical system. LLMs are improving but can still falter in maintaining this contextual awareness through complex derivations. A challenge here is the potential for LLMs to exhibit overcomplication bias. Furthermore, behavioral tuning (e.g., for verbosity or specific output formats like Markdown) might inadvertently reduce their core reasoning capabilities,

an effect sometimes termed an "alignment tax" [41]. The default system prompts of general-purpose LLMs may also not elicit the concise, formal style of mathematical physics, potentially hiding their performance on complex derivations. For shorter calculations, some LLMs have struggled with tasks like counting the number of 'r's in the word "strawberry" or computing '9.9-9.11'. In physics, there are many notations whose rules differ dramatically from other fields, and LLMs should understand the context and apply the correct rule. For example, the normal ordering notation with: ·: from quantum many-body physics.

142

145

146

147

148

149

150

151

152

153 154

155

156

157

158

159

160

162

163

164

165

166 167

168

170

171

172

173

174

177 178

179

180

181

182

183 184

185 186

187

188

A common strategy in physics research is to gain intuition about a complex problem by analyzing simpler, solvable special cases (e.g., zero temperature limit $T \to 0$, one-dimensional version of a 2D problem, specific symmetry points in parameter space) or by relating it to analogous systems (e.g., mapping a quantum spin system to a classical statistical mechanics model). LLMs show some ability to follow instructions to analyze special cases if explicitly prompted. For example, given a general expression for the magnetic susceptibility $\chi(T)$, an LLM might be able to evaluate its behavior as $T \to 0$ (e.g., Curie's law $\chi \propto 1/T$ for paramagnets [42]) or $T \to \infty$. However, spontaneously identifying fruitful special cases or insightful analogies that can simplify a problem or suggest a solution path is a more advanced reasoning skill that remains underdeveloped. **Example: Analyzing Interacting Systems** Consider a complex interacting quantum system described by a Hamiltonian $H = H_{\rm kin} + H_{\rm int}$. A physicist might first analyze the noninteracting limit (setting interaction strength U = 0 in $H_{\rm int}$), or a mean-field approximation, to build intuition. An LLM could be guided to do this, but proactively suggesting "Let's first consider the case where U = 0" or "This problem, under certain limits, is analogous to the Ising model $H_{\rm Ising} = -J \sum_{\langle i,j\rangle} \sigma_i^z \sigma_j^z - h \sum_i \sigma_i^x$ if we make X approximation" demonstrates a higher level of scientific reasoning.

3.2.3 Physical Consistency, Constraint Satisfaction, and Navigating Ambiguity

Skill Ensuring solutions respect fundamental physical principles (e.g., conservation laws like dE/dt = 0, $d\mathbf{P}/dt = 0$, dimensional consistency, causality, symmetries) and problem-specific constraints; recognizing ambiguity in problem statements or scientific texts, making justified assumptions to resolve ambiguity, or querying for clarification. Analysis A critical aspect of physics reasoning is ensuring solutions are physically sensible. LLMs must learn to self-check outputs against fundamental physical laws (e.g., conservation of energy, momentum, charge) and problem-specific constraints (e.g., boundary conditions like $\psi(x=\pm L/2)=0$ for a particle in a box [43], symmetries of the Hamiltonian such as [H, P] = 0 if parity P is conserved). This includes ensuring dimensional consistency of equations (e.g., verifying that terms being added have the same physical units, like Joules for energy) and respecting fundamental symmetries. Developing this "physical common sense" is needed. Current LLMs may generate solutions that are mathematically plausible but physically violate such principles if not carefully guided or checked. Self-correction techniques [44, 45, 46] must be adapted to evaluate physical plausibility alongside logical consistency. **Example: System-Bath Modeling.** In a coding task for modeling the system-bath interactions of quantum many-body spin systems, described by a Hamiltonian like $H = H_S(\{\sigma_i\}) + H_B(\{\tau_i\}) + H_{SB}(\{\sigma_i\}, \{\tau_i\})$, an LLM (e.g., Cursor integrated with Claude Sonnet) might erroneously place the system spins $\{\sigma_i\}$ and the bath spins $\{\tau_i\}$ on the same lattice sites if not explicitly prohibited. This configuration is physically nonsensical for typical models where system and bath are distinct objects with their own degrees of freedom, but might not be directly contradicted by a vague prompt. This type of error, stemming from a lack of "common sense" physical intuition about distinct subsystems, is something a human physicist would typically avoid. Such errors show the current gap in LLMs' physical intuition. Physics research often involves nuanced statements or different notations relying on implicit context. When faced with choices that lead to different solution paths, a human scientist typically seeks clarification with more context, yet LLMs tend to randomly pick one path without justification. This extends to interpreting under-specified problems common in physics, akin to Fermi problems (order-of-magnitude estimations often based on ambiguous information) [47], where making justified assumptions is essential. **Example: Notational Ambiguity** A research note might define a spin Hamiltonian $H_s = -J \sum_{\langle i,j \rangle} \sigma_i^z \sigma_j^z$, and then describe a Jordan-Wigner transformation to map it to a fermionic Hamiltonian, $H_f=-J\sum_{\langle i,j\rangle}(2c_i^\dagger c_i-1)(2c_j^\dagger c_j-1)+\ldots$ For brevity, an author might informally refer to both H_s and H_f as 'H' in different parts of the text. LLM agents often confuse properties or operations valid for H_s (acting on spin Hilbert space) with those for H_f (acting on Fock space). They might attempt to 'correct' the notation by consistently using a new symbol like $H_{\text{transformed}}$, or worse, attempt to apply operations valid for the original H_s to the

transformed H_f if they fail to track the change in underlying variables and the Hilbert space, leading to cascading errors in explanation.

3.2.4 Making Justified Physical Approximations

192

193

194

195

196

197

198

199

200

201

202

203

204

205 206 207

208 209

210

211

212

213

214

215

216

218

219

220

221

222

223

224

226 227

228

229

230

231

232

233

234

235

238

239

240

Skill Selecting appropriate levels of approximation based on physical context, stating assumptions explicitly, and understanding the domain of validity. Analysis Exact solutions are rare; progress often hinges on making well-justified approximations. LLMs need to select appropriate approximation levels (e.g., classical vs. quantum, relativistic vs. nonrelativistic, perturbative expansions, mean-field theory). They may default to standard textbook approximations (like the ideal gas law PV = nRT [48] or the harmonic oscillator potential $V(x) = kx^2/2$) without critically evaluating their validity for the specific problem context or stating the conditions under which they hold. This includes complex expansions like those in stochastic calculus or advanced quantum field theory, where the choice of approximation scheme is nontrivial. Example: Perturbation Theory Consider a quantum system with a Hamiltonian $H = H_0 + \lambda V$, where H_0 is exactly solvable (e.g., a free particle or harmonic oscillator), λ is a small dimensionless perturbation parameter, and V is the perturbation potential. An LLM might be asked for the first-order correction to the ground state energy $E_0^{(0)}$ of H_0 . It should retrieve the standard formula from time-independent perturbation theory: $E_0^{(1)} = \lambda \langle \psi_0^{(0)} | V | \psi_0^{(0)} \rangle$ (see, e.g., [49]), where $\psi_0^{(0)}$ is the ground state eigenfunction of H_0 . However, a crucial aspect is understanding the conditions for the validity of perturbation theory, such as $|\lambda \langle \psi_m^{(0)} | V | \psi_n^{(0)} \rangle | \ll |E_m^{(0)} - E_n^{(0)}|$ for $m \neq n$. An LLM might apply the formula without checking or stating this crucial assumption, or struggle to identify the appropriate H_0 and V if the problem is not explicitly presented in this standard perturbative form (a Taylor expansion in λ).

3.3 Being A Good AI Physicist: Developing Taste and Gracefulness

Skill Exhibiting good research "taste", such as resorting to mathematically elegant explanations by Occam's Razor and avoiding unnecessary complexity. Analysis While solving a problem is hard, solving it elegantly or finding the most insightful approach is much harder. A "good" physicist would not be satisfied with a brute-force answer but would strive for solutions that are simple, generalizable, and offer deeper understanding. This relates to developing a form of "research taste". Current LLMs may sometimes opt for overly complex or brute-force approaches if not guided. Training LLMs to recognize and prefer elegant or simpler solutions, perhaps through reinforcement learning from human feedback that rewards such qualities, could be an important direction [50]. Interpretability studies can also help understand how LLMs arrive at solutions and whether they are employing physical reasoning or relying on superficial pattern matching [51]. Example: Exploiting **Symmetry** Consider calculating the expectation value of the position operator \hat{x} for a particle in a onedimensional potential V(x) that is symmetric, i.e., V(x) = V(-x), such as the harmonic oscillator $V(x) = m\omega^2 x^2/2$ or an infinite square well centered at the origin. If the particle is in an energy eigenstate $|\psi_n\rangle$, its wavefunction $\psi_n(x)$ will have definite parity: either even $(\psi_n(-x) = \psi_n(x))$ or odd $(\psi_n(-x) = -\psi_n(x))$. Consequently, the probability density $|\psi_n(x)|^2$ is always an even function. The expectation value is $\langle \hat{x} \rangle_n = \int_{-\infty}^{\infty} \psi_n^*(x) x \psi_n(x) dx = \int_{-\infty}^{\infty} x |\psi_n(x)|^2 dx$. Since $x = \int_{-\infty}^{\infty} x |\psi_n(x)|^2 dx$. is an odd function and $|\psi_n(x)|^2$ is an even function, their product is odd. The integral of an odd function over a symmetric interval $(-\infty, \infty)$ is zero. Thus, $\langle \hat{x} \rangle_n = 0$ (a standard result discussed in, e.g., [52]). An LLM might attempt a brute-force approach: find the explicit form of $\psi_n(x)$ (e.g., Hermite polynomials for the harmonic oscillator), then perform the integration symbolically or numerically, potentially making calculation mistakes. A 'good AI physicist', however, would recognize the symmetry of the potential and the parity of the integrand to immediately conclude $\langle \hat{x} \rangle_n = 0$ without detailed calculation. Training LLMs to identify and use such symmetries reflects a deeper physical understanding and leads to more elegant and efficient problem-solving. This symmetry principle extends profoundly in physics [53], from continuous symmetries (e.g., Noether's theorem linking them to conservation laws like $d\mathbf{P}/dt = 0$ for translational symmetry) to discrete ones dictating selection rules or fundamental properties (e.g., CP violation in particle physics [54]).

3.4 Code Generation and Execution for Physics

Skill Physics-aware code generation that correctly translates physical models and algorithms, bridging theory, computation, and experiment. **Analysis** LLMs can generate code (NumPy/SciPy) for Monte Carlo simulations in solid state physics and molecular dynamics, perform numerical analysis of equations [55], and assist with data analysis [56], helping to accelerate prototyping. They might assist in maintaining/extending legacy code (e.g., Fortran in large collaborations [57]) or translating

to modern languages. This capability could help bridge theory, computation, and experiment: a theorist might use an LLM to quickly prototype a simulation for a new model; an experimentalist might use it to apply computational analysis to their data without extensive programming expertise. However, physics-aware code generation [58] demands correct translation of physical models. For instance, implementing the Hubbard model $H=-t\sum_{\langle i,j\rangle,\sigma}(c_{i\sigma}^{\dagger}c_{j\sigma}+\text{h.c.})+U\sum_{i}n_{i\uparrow}n_{i\downarrow}$ [59] requires understanding its Hilbert space, symmetries (particle number, S_z conservation), and numerical algorithms (exact diagonalization, Quantum Monte Carlo [60]). A naive LLM agent might miss crucial physical constraints like fermionic anticommutation rules $\{c_{i\sigma},c_{j\sigma'}^{\dagger}\}=\delta_{ij}\delta_{\sigma\sigma'}$ or boundary conditions (e.g., periodic $c_{N+1}=c_1$). Similarly, translating Lattice Gauge Theory (LGT) formalisms [61], like the SU(N_c) Hamiltonian $H=\frac{g^2}{2a}\sum_{l,\alpha}E_l^{\alpha}E_l^{\alpha}-\frac{1}{ag^2}\sum_p \operatorname{ReTr}(U_p)$ (where E_l^{α} are electric field operators, U_p plaquette operators, a lattice spacing, g coupling), into code requires handling complex group theory and ensuring constraints like Gauss's law (e.g., its lattice version $G_n^{\alpha}|{\rm phys}\rangle = (\sum_{l \text{ at } n} E_l^{\alpha})|{\rm phys}\rangle = 0)$ are correctly implemented.

4 LLM Techniques as Augmentation for Physics Research

Various techniques in LLM reasoning can be adapted to tackle several common tasks within physics research, as we detail in this section.

4.1 Literature Review by Retrieval-Augmented Generation and Long-Context Reasoning

By leveraging Retrieval-Augmented Generation (RAG) [62], frontier agentic research systems like DeepResearch [63] can access massive up-to-date literature. The rise of long-context LLMs (e.g., 200K [6] to over 1M tokens [4]) enables workflows that require comprehensive summarization across various data sources such as multiple *Physical Review* papers, PhD theses, or graduate-level textbook chapters (e.g., following the derivation of the Bethe Ansatz solution [64, 65] for the 1D Heisenberg model $H = J \sum_{i} \mathbf{S}_{i} \cdot \mathbf{S}_{i+1}$ across several chapters of a textbook). However, practical limitations persist as performance often degrades as context length increases (the "lost in the middle" phenomenon [66]), and models can be easily distracted by irrelevant information embedded within the context [67]. Effectively combining and synthesizing information from a wide range of diverse documents longer than the context window remains an open research frontier [68].

4.2 Exploratory Reasoning by In-Context Few-Shot Learning

LLMs can adapt their behavior based on in-context demonstrations [1, 69]. For example, LLM agents could infer how to tackle a particular type of equation from a few examples (e.g., the time-independent Schrödinger equation $(-\hbar^2/2m\cdot\psi''+V\psi)=E\psi$ for different potentials V(x) like the harmonic oscillator $V(x)=m\omega^2x^2/2$) and then apply a similar methodology to a new potential, such as the microwave shielding for cold molecules [70, 71], where experimental setups require analyzing a new long-range potential. Using few-shot examples of similar long-range potential analyses, an LLM could help researchers apply established analysis procedures to these novel experimental configurations.

This is particularly relevant in pursuit of new physics that involves new conditions or classes of models where the general solution methodology is known and the solutions are verifiable. LLM agents could study multiple variants of the same problem or multiple solution paths for the same conjecture simultaneously to help human researchers.

4.3 Tool Usage and Reliable Scientific Reasoning by Self-Reflection

Tool Usage LLMs are not inherently calculators or symbolic reasoners, but they can effectively use external tools like symbolic math engines (Mathematica, SymPy), numerical libraries (via code execution), or databases via dedicated portals such as Model-Context Protocols (MCP) [72, 73, 74, 75]. Models need to learn when and how to call these tools effectively, formulate valid queries for them (e.g., correctly translating a subproblem like "calculate $\int_0^\infty x^2 e^{-ax} dx$ for a > 0" (a standard integral found in texts like [76]) into Integrate[x^2 Exp[-a x], {x, 0, Infinity}, Assumptions -> a > 0] for Mathematica), and interpret their output correctly within the physics context. Tool use allows for a more dynamic, nonsequential workflow: LLM agents can query a tool, analyze the output, and then decide on subsequent actions, effectively optimizing their solution path. This contrasts with purely auto-regressive generation, enabling more robust self-correction and complex problem decomposition. **Reliable Scientific Reasoning by Self-Reflection** LLMs suffer from hallucinations or confabulations, which may produce factually

incorrect information that sounds plausible at first [37, 77]. This can lead to flawed conclusions 298 or even potentially dangerous outcomes in an experimental setting. Ensuring the factual accuracy 299 and logical consistency of LLM outputs, especially for complex reasoning chains, remains a major 300 challenge [78]. Techniques like self-critique [46] and RAG [62] with physics-specific knowledge 301 bases show promise for improving factual accuracy but need further development for scientific 302 domains. Self-reflection [44, 45, 46] by external modules or human oversight [78, 79, 80] has shown 303 304 promising performance gains on scientific tasks [78, 81]. This is particularly valuable for catching logical inconsistencies, sign errors in derivations, or violations of conservation laws. 305

Multi-agent simulations [? 82] open the door for streamlining verification, where specialized agents verify different physical constraints separately (e.g., one agent checks dimensional consistency, another checks symmetry properties). A combined system might help accelerate the hypothesis-verification cycle of scientific discovery.

5 Open Directions and Opportunities for LLM agents

5.1 Advancing Multimodal Reasoning

310

311

325

326

327

328

330

331

332 333

337

338

341

342 343

345

Physics is inherently multimodal, relying on text, equations, diagrams, and various forms of data.

LLM agents must evolve to efficiently integrate these diverse information types by parsing, interpreting, and generating specialized visual representations such as Feynman diagrams (see Figure 2), tensor network notations [83] (as shown in the example below), dual unitary circuit diagrams [84] (used in studies of quantum chaos), and phase diagrams.

Current vision-language models show potential in interpreting general plots but struggle with highly specialized physics notations [85]. The ability to seamlessly reason across modalities—for example, connecting a mathematical formalism with its graphical representation and experimental data—would be valuable. This extends beyond calculation to translating diagrams into executable programs (e.g., a quantum circuit diagram into code for a quantum simulator) and assisting in graphical proofs or derivations [86]. Recent advances like OpenAI-o3 demonstrate improved image analysis by calling tools to crop/zoom-in images, but understanding the deeper semantics of physics visualizations requires further progress and careful benchmarking.

Example: Tensor Network Diagram Understanding and manipulating diagrams in specialized fields, such as the tensor network notation often used in quantum information and condensed matter theory, illustrates the type of complex visual-symbolic language that future LLMs should handle. Consider the following tensors (all indices are 3D, indexed from 0):

$$\underbrace{A}_{j} = i^{2} - 5j, \quad \underbrace{B}_{k} = 4^{ij} - k, \quad \underbrace{j}_{i} = j, \quad \underbrace{j}_{k} = ijk, \quad \underbrace{A}_{k} = 0.$$

The task is to calculate the value of the last tensor network (perturbed from [87]). An ideal LLM assistant would parse the diagram, identify tensors and their connectivity, translate this into an algebraic expression $\sum_{a,b,c,d,e,f} A_{af} B_{abc} D_{cd} C_{fed}$, substitute definitions, and perform the computation via generated code. LLMs could assist by creating such diagrams from text or formula descriptions. Mastery of such visual-symbolic languages could extend to interpreting Feynman diagrams by extending them into scattering amplitude calculations, or parsing quantum circuit diagrams to determine their unitary evolution [88] or even carrying out graphical proofs of mathematical physics with such a graphical language [86].

5.2 Developing Agentic Capabilities for Scientific Discovery

Future LLM systems may evolve into more autonomous agents capable of performing full-cycle scientific tasks with greater independence under human oversight.

5.2.1 Agentic AI for Hypothesis Generation and Verification

Future AI agents might propose new models by analyzing anomalies [89] and inconsistencies among different theories, exploring multiple branches of a solution tree and alternative physical models or mathematical ansätze, and then systematically validating each option. By scanning through parameter spaces [82, 90, 91, 92] guided by physical principles, AI-assisted scientific discovery [93, 89] may eventually contribute to scientific hypothesis generation akin to AlphaGo Move 37 [94, 95], though this remains an ambitious prospect.

Variational methods in computational physics work by devising an appropriate parameterized class of 348 variational wavefunction $|\Psi(\{\alpha_i\})\rangle$. This state should capture the essential physics of the system 349 (e.g., correlations, symmetries) while being computationally verifiable by minimizing the energy with 350 respect to $\{\alpha_i\}$. An LLM might assist by suggesting functional forms for $|\Psi\rangle$ based on the known 351 properties of the Hamiltonian (e.g., suggesting a Gutzwiller-projected wavefunction for strongly 352 correlated systems, or a Matrix Product State for 1D systems [83]), incorporating specific symmetries 353 (e.g., in lattice gauge theory [96, 97]), or using non-Gaussian state ansätze [98] that require long 354 analytical calculations. The verifiability of variational methods is direct: a better guess leads to a 355 lower calculated energy, providing a clear objective function for iterative optimization. 356

5.2.2 Automated Simulation, Experimental Design, and Verifying Theoretical Results at Scale

LLM agents might assist in optimizing experimental designs or simulation parameters, particularly where theoretical models can guide the process, to maximize information gain or test specific hypotheses [99]. This could involve suggesting appropriate measurement techniques (e.g., choosing between different spectroscopic methods to probe a material's electronic structure), identifying key parameters to calibrate in an experiment with a quantum gas microscope [100] (where images from a CCD camera are used to reconstruct particle configurations), or even interfacing with automated cloud labs [93].

A significant challenge would be for an LLM to assist in verifying highly complex proofs in mathematical physics, such as Hastings's proof of the super-additivity of Holevo information [101]. It would be valuable to automatically cross-check the use of inequalities against established databases at scale. Such verification is particularly important given that very technical results typically take years to fact-check, and errors in published proofs are not uncommon [102].

5.3 Fine-tuned LLM Physicists and Towards AI Physicists

Specialized LLMs fine-tuned for physics could offer advantages over general-purpose models [103]. Such models would prioritize domain knowledge (e.g., quantum mechanics principles), eliminate irrelevant information (e.g., historical facts unrelated to physics), and focus on physical reasoning patterns (e.g., dimensional analysis, order-of-magnitude estimation, symmetry arguments). Fine-tuning could involve supervised [104] and reinforcement learning [105]. To train physics-dedicated LLMs, we need to collect more nuanced and effective reward signals, similar to how coding agents are improving. Beyond simple pass/fail on benchmark problems, rewards should ideally capture the quality of the reasoning process [78], the physical insightfulness of solutions, and alignment with established scientific methodology by incorporating feedback from domain experts.

A long-term vision would be for LLM agents to become effective AI collaborators, or building blocks of automated "AI physicists", capable of full-cycle capabilities including proposing novel research ideas, theorizing experimental phenomena, verifying hypothesis and assisting in all research stages [89, 93]. Apart from significant technical advances, this would require a synergistic partnership where AI models augment human intellect, supported by suitable UI/UX and appropriate guardrails. Imagine investigating a novel material: an AI assistant might synthesize literature, formulate computational models using solid-state physics principles, generate simulation code, explore analytical approximations, and report results. For theoretical physics, a distant prospect is to deploy AI agents for tackling open problems in the field [106].

6 Conclusion

357

370

380

381

382

383

386 387

388

389

390

392

393

394

395

396

397 398

399

LLMs have the potential to contribute meaningfully to modern physics research. Their potential to help accelerate scientific discovery, automate repetitive tasks, and assist in conceptual breakthroughs is considerable. However, realizing this potential requires substantial effort to address current limitations in rigorous reasoning, physical grounding, reliability, and multimodal understanding. By fostering collaboration between physics and AI communities to develop specialized models, robust verification techniques, and effective human-AI interfaces, we can work toward using LLMs to contribute to expanding our understanding of the physical universe. Furthermore, applying LLMs to physics serves as a demanding testbed for studying LLMs, including interpretability, faithfulness of reasoning, adversarial robustness, and scalable oversight for safety.

References

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind
 Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss,
 Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens
 Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark,
 Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language
 Models are Few-Shot Learners. In Adv. Neural Inf. Process. Syst., volume 33, pages 1877–1901. Curran
 Associates, Inc., 2020. 1, 6
- 408 [2] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understand-409 ing by generative pre-training. 2018. 1
- 410 [3] OpenAI. GPT-4 technical report, 2023. 1
- 411 [4] Gemini Team. Gemini: A Family of Highly Capable Multimodal Models, June 2024. 1, 6
- 412 [5] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 413 December 2024. 1
- [6] Anthropic. Claude 3 model card. Technical report, Anthropic, PBC, March 2024. 1, 6
- [7] Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T. Bergstrom, Colin Allen, Daniel
 Schad, Dirk Wulff, Jevin D. West, Qiong Zhang, Richard M. Shiffrin, Samuel J. Gershman, Vencislav
 Popov, Emily M. Bender, Marco Marelli, Matthew M. Botvinick, Zeynep Akata, and Eric Schulz. How
 should the advancement of large language models affect the practice of science? Proceedings of the
 National Academy of Sciences, 122(5):e2401227121, February 2025.
- 420 [8] DeepSeek-AI. DeepSeek-V3 Technical Report, February 2025. 1
- [9] DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning,
 January 2025. 1
- [10] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, 423 Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe 424 Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, 425 Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, 426 Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, 427 Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, 428 Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying 429 Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, 430 Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, 431 Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying 432 Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, 433 434 Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. 435 Kimi k1.5: Scaling Reinforcement Learning with LLMs, March 2025. 1
- 436 [11] H.-Q. Ding and M. S. Makivić. Kosterlitz-Thouless transition in the two-dimensional quantum XY model.
 437 Phys. Rev. B, 42(10):6827–6830, October 1990. 2
- [12] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie
 Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Rev. Mod. Phys.*,
 91(4):045002, December 2019. 2, 17
- Imagir Azerbayev, Denys Kocetkov, Artem Kocetkov, Shubham Toshniwal, Yuntao Bai, Charles Sutton, Jared Kaplan, Azalia Mirhoseini, Aakanksha Chowdhery, Roger Grosse, Ilya Sutskever, Jean-Baptiste Alayrac, Alhussein Fawzi, and Jascha Sohl-Dickstein. Llemma: An open language model for mathematics.
 In Advances in Neural Information Processing Systems, volume 36, pages 26233–26249, 2023.
- Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, January 2024. 2
- [15] Stanislas Polu, Jesse Michael Han, Kaiyu Zheng, Ambrose Baksys, Igor Babuschkin, Richard Sutton,
 Arthur Szlam, Jean-Baptiste Alayrac, and Ilya Sutskever. AI assistance for proving mathematical theorems
 in Lean. *Nature*, 626:115–122, 2024. 2
- 450 [16] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 451 Augmenting large language models with chemistry tools. *Nat Mach Intell*, 6(5):525–535, May 2024. 2

- Varuni Sarwal, Gaia Andreoletti, Viorel Munteanu, Ariel Suhodolschi, Dumitru Ciorba, Viorel Bostan,
 Mihai Dimian, Eleazar Eskin, Wei Wang, and Serghei Mangul. A benchmark for large language models
 in bioinformatics, April 2025. 2
- Thilo Hagendorff, Ishita Dasgupta, Marcel Binz, Stephanie C. Y. Chan, Andrew Lampinen, Jane X. Wang,
 Zeynep Akata, and Eric Schulz. Machine Psychology, August 2024. 2, 3
- [19] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a
 mirage? In Advances in Neural Information Processing Systems 36 (NeurIPS 2023), 2023. 2, 3, 17
- Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862. Association for Computational Linguistics, June 2024. 2, 3
- [21] D. J. Thouless, M. Kohmoto, M. P. Nightingale, and M. den Nijs. Quantized Hall Conductance in a
 Two-Dimensional Periodic Potential. *Phys. Rev. Lett.*, 49(6):405–408, August 1982.
- [22] K. v. Klitzing, G. Dorda, and M. Pepper. New Method for High-Accuracy Determination of the Fine-Structure Constant Based on Quantized Hall Resistance. *Phys. Rev. Lett.*, 45(6):494–497, August 1980.
 3
- [23] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob
 Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. 3
- 472 [24] Zhihong Sun, Kuan Kuang, Yiming Ji, Wenxiang Wang, Qun Liu, Hua Wu, Haifeng Wang, Shiqi Wu, 473 and Zhi-Hong Dou. SciEval: A multi-level large language model evaluation benchmark for scientific 474 research, 2023. 3
- [25] Daniel J. H. Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic
 Sala, and Sai Chaitanya Tadepalli. Theoretical Physics Benchmark (TPBench) a Dataset and Study of
 AI Reasoning Capabilities in Theoretical Physics, February 2025.
- Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng
 Shou, Lingling Zhang, and Jun Liu. PhysReason: A Comprehensive Benchmark towards Physics-Based
 Reasoning, February 2025. 3
- [27] Shi Qiu, Shaoyang Guo, Zhuo-Yang Song, Yunbo Sun, Zeyu Cai, Jiashen Wei, Tianyu Luo, Yixuan Yin, 481 Haoxu Zhang, Yi Hu, Chenyang Wang, Chencheng Tang, Haoling Chang, Qi Liu, Ziheng Zhou, Tianyu 482 Zhang, Jingtian Zhang, Zhangyi Liu, Minghao Li, Yuku Zhang, Boxuan Jing, Xianqi Yin, Yutong Ren, 483 Zizhuo Fu, Weike Wang, Xudong Tian, Anqi Lv, Laifu Man, Jianxiang Li, Feiyu Tao, Qihua Sun, Zhou 484 Liang, Yushu Mu, Zhongxuan Li, Jing-Jun Zhang, Shutao Zhang, Xiaotian Li, Xingqi Xia, Jiawei Lin, 485 Zheyu Shen, Jiahang Chen, Qiuhao Xiong, Binran Wang, Fengyuan Wang, Ziyang Ni, Bohan Zhang, Fan 486 Cui, Changkun Shao, Qing-Hong Cao, Ming-xing Luo, Muhan Zhang, and Hua Xing Zhu. PHYBench: 487 Holistic Evaluation of Physical Perception and Reasoning in Large Language Models, April 2025. 3 488
- 489 [28] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba,
 490 Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating college-level scientific problem491 solving abilities of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller,
 492 Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st*493 International Conference on Machine Learning (ICML), volume 235 of Proceedings of Machine Learning
 494 Research, pages 50622–50649. PMLR, 2024. 3
- [29] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R
 Narasimhan. SWE-bench: Can language models resolve real-world github issues? In The Twelfth
 International Conference on Learning Representations, 2024. 3
- [30] Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel
 Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal
 Patwardhan. PaperBench: Evaluating AI's Ability to Replicate AI Research, April 2025.
- [31] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman
 Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett,
 Robert Sandler, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart,
 Bogdan Grechuk, Tetiana Grechuk, and Shreepranav Varma Enugandla. FrontierMath: A benchmark for
 evaluating advanced mathematical reasoning in AI. arXiv preprint arXiv:2411.04872, November 2024. 3

- 506 [32] Long Phan, Alice Gatti, Ziwen Han, and et al. Humanity's Last Exam, April 2025. 3
- 507 [33] bench.science. bench.science: Benchmarking the future of science, 2025. Initiative for standardized AI benchmarks across scientific domains. 3
- 509 [34] Ernest Davis and Scott Aaronson. Testing GPT-4 with Wolfram Alpha and Code Interpreter plug-ins on 510 math and science problems, February 2025. 3
- 511 [35] Ernest Davis. Testing GPT-4-o1-preview on math and science problems: A follow-up study, October 2024. 3
- [36] Haining Pan, Nayantara Mudur, William Taranto, Maria Tikhanovskaya, Subhashini Venugopalan,
 Yasaman Bahri, Michael P. Brenner, and Eun-Ah Kim. Quantum many-body physics calculations with
 large language models. Commun Phys, 8(1):1–8, January 2025. 3
- [37] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao,
 Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's
 Song in the AI Ocean: A Survey on Hallucination in Large Language Models, September 2023. 3, 7, 16
- [38] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh,
 Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy
 Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In Advances
 in Neural Information Processing Systems 36 (NeurIPS 2022), 2022. 3, 17
- 523 [39] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor 524 Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali 525 Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda 526 Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, 527 Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, 528 529 Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, 530 Aykut Erdem, Ayla Karakas, and et al. Beyond the Imitation Game: Quantifying and extrapolating the 531 capabilities of language models, 2022. 3 532
- [40] Kaixuan Huang, Jiacheng Guo, Zihao Li, Xiang Ji, Jiawei Ge, Wenzhe Li, Yingqing Guo, Tianle Cai, Hui
 Yuan, Runzhe Wang, Yue Wu, Ming Yin, Shange Tang, Yangsibo Huang, Chi Jin, Xinyun Chen, Chiyuan
 Zhang, and Mengdi Wang. MATH-Perturb: Benchmarking LLMs' Math Reasoning Abilities against
 Hard Perturbations, February 2025. 3
- [41] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones,
 Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez,
 Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam
 McCandlish, Chris Olah, and Jared Kaplan. A General Language Assistant as a Laboratory for Alignment,
 December 2021. 4
- 542 [42] Charles Kittel. Introduction to Solid State Physics. John Wiley & Sons, 8th edition, 2004. 4
- [43] David J. Griffiths and Darrell F. Schroeter. *Introduction to Quantum Mechanics*. Cambridge University
 Press, 3rd edition, 2018. 4
- [44] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon,
 Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder,
 Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative Refinement with Self-Feedback, May 2023. 4, 7
- [45] Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu
 Yao. Reflexion: Language Agents with Verbal Reinforcement Learning, October 2023. 4, 7
- [46] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike.
 Self-critiquing models for assisting human evaluators, June 2022. 4, 7
- 553 [47] Larry Weinstein. Fermi Questions. The Physics Teacher, 48(7):490, October 2010. 4
- 554 [48] Frederick Reif. Fundamentals of Statistical and Thermal Physics. McGraw-Hill, New York, 1965. 5
- [49] J. J. Sakurai and Jim Napolitano. *Modern Quantum Mechanics*. Cambridge University Press, 2 edition,
 2017. 5

- [50] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain,
 Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom
 Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott
 Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown,
 Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and
 harmless assistant with reinforcement learning from human feedback, 2022.
- [51] Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model, 2023. 5
- [52] Ramamurti Shankar. Principles of Quantum Mechanics. Plenum Press, 2 edition, 1994. 5
- [53] David J. Gross. The role of symmetry in fundamental physics. *Proceedings of the National Academy of Sciences*, 93(25):14256–14259, December 1996.
- [54] T. D. Lee and C. N. Yang. Question of Parity Conservation in Weak Interactions. *Phys. Rev.*,
 104(1):254–258, October 1956.
- 570 [55] Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. LLMPhy: Complex Physical 571 Reasoning Using Large Language Models and World Models, December 2024. 5
- [56] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and
 Graham Neubig. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*,
 pages 10764–10799. PMLR, 2023. 5
- 576 [57] Anthony Zhou, Linnia Hawkins, and Pierre Gentine. Proof-of-concept: Using ChatGPT to translate and modernize an earth system model from fortran to python/JAX, February 2024. 5
- 578 [58] Yufei Tian, Libo Wang, and Lei Wang. SciCode: A comprehensive benchmark for evaluating code generation capability of large language models in science, 2024. 6
- [59] Assa Auerbach. Interacting Electrons and Quantum Magnetism. Springer-Verlag, 1994. 6
- [60] Federico Becca and Sandro Sorella. Quantum Monte Carlo Approaches for Correlated Systems. Cambridge University Press, Cambridge, 2017.
- 583 [61] Kenneth G. Wilson. Confinement of quarks. Phys. Rev. D, 10(8):2445–2459, October 1974. 6
- [62] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,
 Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela.
 Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc' Aurelio
 Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020,
 NeurIPS 2020, December 6-12, 2020, Virtual, 2020. 6, 7
- [63] OpenAI. Introducing deep research, February 2025. AI-powered multi-step research tool for complex
 tasks. 6
- 592 [64] H. Bethe. Zur Theorie der Metalle. Z. Physik, 71(3):205–226, March 1931. 6
- [65] Fabian H. L. Essler, Holger Frahm, Frank Göhmann, Andreas Klümper, and Vladimir E. Korepin. *The One-Dimensional Hubbard Model*. Cambridge University Press, 2005.
- [66] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy
 Liang. Lost in the middle: How language models use long contexts, July 2023.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli,
 and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31224. PMLR, July 2023.
- [68] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020. 6
- [69] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. Language models of code
 are few-shot commonsense learners, 2022. 6
- [70] Xing-Yan Chen, Andreas Schindewolf, Sebastian Eppelt, Roman Bause, Marcel Duda, Shrestha Biswas,
 Tijs Karman, Timon Hilker, Immanuel Bloch, and Xin-Yu Luo. Field-linked resonances of polar molecules.
 Nature, 614(7946):59–63, February 2023. 6

- Fulin Deng, Xing-Yan Chen, Xin-Yu Luo, Wenxian Zhang, Su Yi, and Tao Shi. Effective Potential and Superfluidity of Microwave-Shielded Polar Molecules. *Phys. Rev. Lett.*, 130(18):183001, May 2023. 6
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. arXiv, December 2023. 6
- [73] Shishir G. Patil, Tianjun Chen, Pingchuan Liang, Xuechen Luan, Dung Tran, Hyung Won Choi, Charles
 Sutton, Sören Mindermann, Alicia Parrish, Judy Hoffman, Le Hou, Brandon Sapp, Joseph E. Gonzalez,
 Ion Stoica, and Dawn Song. Gorilla: Large language model connected with massive apis. In Advances in
 Neural Information Processing Systems, volume 36, pages 1885–1909, 2023.
- 615 [74] Anthropic. Introducing the model context protocol, November 2024. Open standard for connecting AI systems with data sources. 6
- [617] Pathintegral Institute. MCP.science: Open Source MCP Servers for Scientific Research. GitHub repository,
 618 2025. Collection of Model Context Protocol servers for scientific applications. 6
- [76] George B. Arfken, Hans J. Weber, and Frank E. Harris. *Mathematical Methods for Physicists*. Academic Press, 7 edition, 2013. 6
- 621 [77] L. Bottou and B. Schölkopf. The fiction machine. SIAM News, 58(3), April 2025. 7
- [78] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike,
 John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step, May 2023. 7, 8
- [79] Nat McAleese, Rai Michael Pokorny, Juan Felipe Ceron Uribe, Evgenia Nitishinskaya, Maja Trebacz,
 and Jan Leike. LLM Critics Help Catch LLM Bugs, June 2024.
- [80] Jan Hendrik Kirchner, Yining Chen, Harri Edwards, Jan Leike, Nat McAleese, and Yuri Burda. Prover Verifier Games improve legibility of LLM outputs, August 2024.
- [81] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha
 Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language
 Models. In The Eleventh International Conference on Learning Representations, September 2022. 7
- [82] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang,
 Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang.
 AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, October 2023.
- [83] Ignacio Cirac, David Perez-Garcia, Norbert Schuch, and Frank Verstraete. Matrix Product States and
 Projected Entangled Pair States: Concepts, Symmetries, and Theorems. Rev. Mod. Phys., 93(4):045003,
 December 2021. 7, 8
- 637 [84] Bruno Bertini, Pavel Kos, and Tomaz Prosen. Exact correlation functions for dual-unitary lattice models 638 in 1 + 1 dimensions. *Phys. Rev. Lett.*, 123:210601, Nov 2019. 7
- [85] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia,
 Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual,
 multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, 2023. 7
- [86] Arthur Jaffe, Zhengwei Liu, Jacob D. Biamonte, John Ewing, and Alina Vdovina. Mathematical picture
 language program. *Proc. Natl. Acad. Sci.*, 115(1):81–86, 2018.
- [87] Jacob C Bridgeman and Christopher T Chubb. Hand-waving and interpretive dance: An introductory
 course on tensor networks. J. Phys. A, 50(22):223001, May 2017.
- [88] Pavel Kos, Bruno Bertini, and Tomaž Prosen. Correlations in Perturbed Dual-Unitary Circuits: Efficient
 Path-Integral Formula. *Phys. Rev. X*, 11(1):011022, February 2021. 7
- [89] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak,
 Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P.
 Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng
 Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, August 2023. 7, 8, 19
- [90] Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language
 Agent Tree Search Unifies Reasoning Acting and Planning in Language Models, December 2023.

- [91] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi,
 Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of
 Thoughts: Solving Elaborate Problems with Large Language Models. AAAI, 38(16):17682–17690, March
 2024. 7
- [92] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen
 Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng
 Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench:
 Evaluating LLMs as agents. In *The Twelfth International Conference on Learning Representations*, 2024.
- [93] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist:
 Towards Fully Automated Open-Ended Scientific Discovery, August 2024. 7, 8
- [94] David Silver, Aja Huang, Christopher J Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- [95] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez,
 Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui,
 Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of Go
 without human knowledge. *Nature*, 550(7676):354–359, October 2017. 7
- [96] Erez Zohar and J. Ignacio Cirac. Combining tensor networks with Monte Carlo methods for lattice gauge
 theories. *Phys. Rev. D*, 97(3):034510, February 2018.
- [97] Julian Bender, Patrick Emonts, and J. Ignacio Cirac. Variational Monte Carlo algorithm for lattice gauge theories with continuous gauge groups: A study of (2 + 1) -dimensional compact QED with dynamical fermions at finite density. *Phys. Rev. Research*, 5(4):043128, November 2023. 8
- [98] Tao Shi, Eugene Demler, and J. Ignacio Cirac. Variational Study of Fermionic and Bosonic Systems with
 Non-Gaussian States: Theory and Applications. *Ann. Phys.*, 390:245–302, July 2017.
- [99] Jan Kaiser, Anne Lauscher, and Annika Eichler. Large language models for human-machine collaborative particle accelerator tuning through natural language. *Science Advances*, 11(1):eadr4173, January 2025. 8
- [100] Christian Gross and Immanuel Bloch. Quantum simulations with ultracold atoms in optical lattices.
 Science, 357(6355):995–1001, September 2017.
- 687 [101] M. B. Hastings. Superadditivity of communication capacity using entangled inputs. *Nat. Phys.*, 688 5(4):255–257, April 2009. 8
- [102] Thomas Vidick. It happens to everyone...but it's not fun. MyCQstate blog, September 2020. Personal
 reflection on errors in published physics research. 8
- [103] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In
 Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th
 International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620,
 Hong Kong, China, November 2019. Association for Computational Linguistics. 8
- [104] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke
 Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe.
 Training language models to follow instructions with human feedback. In Advances in Neural Information
 Processing Systems, volume 35, pages 27730–27744, December 2022. 8
- [105] Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. ReFT: Reasoning with
 Reinforced Fine-Tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the* 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages
 7601–7614, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [106] Institute for Quantum Optics and Quantum Information Vienna. Open quantum problems, 2025. Collection
 of open problems in quantum information theory. 8

- Yoshua Bengio, Geoffrey E. Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin
 Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian K. Hadfield, Jeff Clune, Tegan Maharaj, Frank Hutter,
 Atilim Günes Baydin, Sheila A. McIlraith, Qiqi Gao, Ashwin Acharya, David Krueger, Anca D. Dragan,
 Philip H. S. Torr, Stuart Russell, Daniel Kahneman, Jan Brauner, and Sören Mindermann. Managing AI
 risks in an era of rapid progress. Science, 384(6698), May 2025. 16
- [108] Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh
 Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario
 Günther, Anton Korinek, José Hernández-Orallo, Lewis Hammond, Eric J. Bigelow, Alexander Pan,
 Lauro Langosco, Tomasz Korbak, Heidi Zhang, Ruiqi Zhong, Seán Ó hÉigeartaigh, Gabriel Recchia,
 Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Yoshua Bengio, Danqi Chen, Samuel
 Albanie, Tegan Maharaj, Jakob Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and
 David Krueger. Foundational challenges in assuring alignment and safety of large language models, 2024.
 16
- [109] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding
 in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198. Association for Computational Linguistics, 2020. 16
- Theodore Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas Griffiths. Cognitive Architectures for
 Language Agents. Transactions on Machine Learning Research, October 2023. 16
- [111] Lu Wang, Fangkai Yang, Chaoyun Zhang, Junting Lu, Jiaxu Qian, Shilin He, Pu Zhao, Bo Qiao, Ray
 Huang, Si Qin, Qisheng Su, Jiayi Ye, Yudi Zhang, Jian-Guang Lou, Qingwei Lin, Saravan Rajmohan,
 Dongmei Zhang, and Qi Zhang. Large action models: From inception to implementation, 2025. 17
- [112] Huaping Liu, Di Guo, and Angelo Cangelosi. Embodied intelligence: A synergy of morphology, action,
 perception and learning. ACM Comput. Surv., 57(7):186:1–186:36, 2025.
- 729 [113] Dan Guest, Kyle Cranmer, and Daniel Whiteson. Deep Learning and Its Application to LHC Physics.
 730 Annual Review of Nuclear and Particle Science, 68(Volume 68, 2018):161–181, October 2018. 17
- 731 [114] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017. 17
- [115] Ivan Glasser, Nicola Pancotti, Moritz August, Ivan D Rodriguez, and J. Ignacio Cirac. Neural-Network
 Quantum States, String-Bond States, and Chiral Topological States. *Phys. Rev. X*, 8(1):011006, January
 2018, 17
- [116] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang.
 Physics-informed machine learning. *Nat. Rev. Phys.*, 3(6):422–440, June 2021. 17
- [117] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese,
 Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de las Casas, Craig Donner, Leslie Fritz,
 Cristian Galperti, Andrea Huber, James Keeling, Maria Tsimpoukelli, Jackie Kay, Antoine Merle,
 Jean-Marc Moret, Seb Noury, Federico Pesamosca, David Pfau, Olivier Sauter, Cristian Sommariva,
 Stefano Coda, Basil Duval, Ambrogio Fasoli, Pushmeet Kohli, Koray Kavukcuoglu, Demis Hassabis, and
 Martin Riedmiller. Magnetic control of tokamak plasmas through deep reinforcement learning. Nature,
 602(7897):414–419, February 2022. 17
- 745 [118] R. P. Feynman. The Theory of Positrons. *Phys. Rev.*, 76(6):749–759, September 1949. 17
- [119] Michael E. Peskin and Daniel V. Schroeder. An Introduction to Quantum Field Theory. Addison-Wesley,
 1995. 17

8 A Risks, Limitations and Ethical Considerations

Risk

749

763

775

776

777

778

781

783

784 785 786

787

790

791

792

793

796

797

Over-reliance on LLMs without rigorous verification could embed subtle errors into research [37]. 750 The potential for LLMs to "cheat" reward functions during fine-tuning, producing plausible but 751 physically invalid outputs (e.g., a simulation appearing to conserve energy due to numerical artifacts), 752 requires careful alignment and robust evaluation [107, 108]. Furthermore, depending too heavily on LLMs for tasks like mathematical derivations (e.g., routinely asking an LLM to compute integrals like $\int d^4k/(k^2-m^2+i\epsilon)^2$ instead of learning contour integration techniques), programming, or 756 data interpretation could risk degrading these essential skills among physicists, especially those in training [109]. Deep intuition often arises from performing detailed calculations firsthand. The history of scientific computing shows both warnings and reassurances: tools like Mathematica initially 758 raised de-skilling concerns but ultimately enabled mathematicians to focus on higher-level work by 759 automating repetitive calculations. Similarly, LLMs could potentially elevate physics research by 760 handling routine tasks while humans focus on deeper insights—if used as augmentation rather than 761 replacement for fundamental understanding. 762

Limitations

This position paper presents a high-level overview. The field of LLMs is rapidly evolving, and specific capabilities or limitations discussed may change quickly. Due to the rapid evolution of LLMs, specific examples quickly become outdated. The selected examples are illustrative of general trends observed circa late 2024 and early 2025. The scope is necessarily limited to selected aspects of physics research, and specific examples may not generalize to all subfields.

769 Ethical Considerations

Generating plausible but incorrect claims requires rigorous validation. Training bias could steer research suboptimally. Responsible deployment and human oversight are required. Access and equity issues must be addressed to ensure broad availability of these tools across the global physics community.

B Building Better UI/UX for Human-centered AI

For LLMs to be effective collaborators, intuitive and efficient user interfaces (UIs) and user experiences (UXs) are essential for supervision, tracing, and trust-building. These interfaces should allow physicists to interact with LLMs naturally without extensive prompt engineering and should integrate with existing research workflows and tools (e.g., LaTeX editors, data analysis and simulation environments). Future LLMs should also be robust to specific prompts and offer finer-grained controllability over their reasoning style, level of detail, and assumptions made. For instance, prompting an LLM to "solve the Schrödinger equation for a particle in a box" might yield different solution forms depending on subtle phrasing. Ideally, an LLM should recognize standard conventions (e.g., specific boundary conditions $\psi(0) = \psi(L) = 0$ for a box of length L) or prompt for these if ambiguous. Controllability would allow a physicist to specify, for example, "provide a solution using separation of variables and show all steps for $H\psi=E\psi$ where $H=-\hbar^2/2m\cdot d^2/dx^2+V(x)$ and V(x)=0 for 0< x< L, ∞ otherwise" versus "give the energy eigenvalues $E_n=\frac{n^2\pi^2\hbar^2}{2mL^2}$ and normalized wavefunctions $\psi_n(x) = \sqrt{2/L}\sin(n\pi x/L)$ directly". Such controllability is vital for making LLMs reliable and adaptive research assistants. Effective UI/UX must go beyond simple chat interfaces. Physicists often work with extensive comments, annotations, and margin notes; interfaces supporting these natural workflows would be more effective. For supervising agents, UIs need robust mechanisms for managing experimental/simulation results, tracking context across long interactions (context management [110]), and accommodating human-in-the-loop intervention. Given that LLM outputs can be verbose, tools for generating structured summaries with highlighting are needed. Integration with collaborative platforms (e.g., Overleaf-like features with LLM assistance for consistency checking in LATEX documents, or GitHub-style review tools for coding and derivations) would also be convenient.

C Running Physical Experiments

While our primary focus is on theoretical physics, we acknowledge that in the long term, AI systems might also actively control experimental instrumentation, interpret sensory data in real-time, and adjust experimental parameters accordingly. This would require the seamless integration of perception

(e.g., using computer vision to optimize laser beam path setup for quantum optics experiments where alignment precision is critical for data quality), reasoning (understanding the experimental progress and deciding which measurements to perform next based on acquired data), and action (subsequently adjusting the lensing setup with high-precision robotic arms) in the physical world. This long-term vision enables a dynamic integration of reasoning models with robotics and control theory, bridging high-level human-defined agenda with corresponding physical actions as envisioned by rising interest in Large Action Models (LAM) [111] and Embodied Intelligence [112].

D Related Works

800

801

802

803

804

805 806

807

808

809

810

811

812

813

816

818

819

820

821

822

823

Non-Language Models Already Help Physics Machine learning (ML) is not new to physics [12]. Current applications include analyzing large experimental datasets (e.g., particle identification at the Large Hadron Collider [113]), solving computational physics problems (e.g., finding ground states of quantum Hamiltonians like $H\psi = E\psi$ [114, 115]), accelerating partial differential equation solvers [116], and optimizing experimental controls (e.g., plasma shaping in fusion reactors [117]). These applications typically involve supervised learning (classification, regression), unsupervised learning (clustering, dimensionality reduction, generative modeling), or reinforcement learning for specific, well-defined tasks. While powerful for specific tasks, these methods often differ from the requirements of open-ended theoretical exploration, complex multistep problem solving, or nuanced experimental design where LLMs might offer complementary advantages through their natural language interface and broad knowledge encoding [19, 38].

E More Analysis and Examples

Example: Notational Nuances Another example is the Bogoliubov-de Gennes (BdG) Hamiltonian, which often includes a 1/2 prefactor by convention; LLMs might add or omit this factor inconsistently if not carefully prompted, thereby impacting all subsequent calculations even though the authors intend a different factor. **Example: Lattice Gauge Theory** For example, implementing the pure gauge SU(2) Hamiltonian:

$$H = \frac{g^2}{2} \sum_{l} \hat{E}_l^a \hat{E}_l^a + \frac{1}{2g^2} \sum_{p} \left(2 - \text{Tr}(\hat{U}_p + \hat{U}_p^{\dagger}) \right)$$
 (1)

where \hat{E}^a_l are electric field operators on links l, \hat{U}_p are plaquette operators, and g is the coupling con-825 stant, requires translating abstract gauge theory concepts into concrete numerical algorithms that pre-826 serve gauge invariance and other symmetries. A critical physical constraint in such simulations is en-827 suring that states satisfy Gauss's law, which in the quantum context becomes $\sum_{l \in \text{star}(n)} \hat{E}^a_l |\psi\rangle_{\text{phys}} = 0$ 828 for each lattice site n and each gauge group generator a. This constraint must be explicitly enforced 829 in the code, typically by projecting onto the physical subspace or by adding an energy penalty term. 830 Example: Explaining the derivations A research paper may state a key result derived from an 831 effective action, $S_{\rm eff}[\phi_c]$, obtained by "integrating out" high-momentum modes ϕ_h from a full action 832 $S[\phi_c,\phi_h]$. An LLM assisting a researcher could be tasked to elaborate on the formal path integral definition $e^{-S_{\text{eff}}[\phi_c]/\hbar} = \int \mathcal{D}\phi_h e^{-S[\phi_c,\phi_h]/\hbar}$. This elaboration might involve expanding the derivations 833 834 with common evaluation techniques like saddle-point approximations or perturbative expansions of 835 with common evaluation techniques like studie-point approximations of perturbative expansions of $S[\phi_c,\phi_h]$ around a background field, all while strictly adhering to the paper's specific notation for the classical fields ϕ_c and quantum fluctuations ϕ_h . **Example: Diagonalizing a 2x2 Hermitian matrix**When asked to diagonalize a general 2x2 Hermitian matrix $H = \begin{pmatrix} a & b - ic \\ b + ic & d \end{pmatrix}$ (where a, b, c, d836 838 are real but have complex expressions), an LLM might default to a brute-force symbolic expansion of 839 the characteristic determinant $det(H - \lambda I) = 0$ to find eigenvalues, followed by solving systems 840 of linear equations for eigenvectors. A 'good AI physicist', however, might recognize the structure and suggest decomposing the matrix in the Pauli basis: $H = a_0 I + \mathbf{a} \cdot \boldsymbol{\sigma}$, where I is the identity matrix, $\sigma = (\sigma_x, \sigma_y, \sigma_z)$ are the Pauli matrices, $a_0 = (a+d)/2$, and $\mathbf{a} = (b, c, (a-d)/2)$. From 843 this decomposition, eigenvalues $(a_0 \pm |\mathbf{a}|)$ and eigenvectors (related to the direction of \mathbf{a}) can be read 844 off with greater physical insight (e.g., connecting to spin precession in a magnetic field) and often 845 less computation. Training LLMs to prefer such insightful decompositions over brute-force methods is key to developing AI assistants that contribute to more elegant theory and can reduce complex algebraic manipulations where errors might occur. Example: Feynman diagram Interpreting a Feynman diagram [118] for Compton scattering ($\gamma e^- \to \gamma e^-$) (see, e.g., [119]) requires identifying

incoming/outgoing photon (wavy lines) and electron (solid lines) lines, internal propagators (e.g., electron propagator $S_F(p)=i(\gamma\cdot p+m_e)/(p^2-m_e^2+i\epsilon)$, where m_e is electron mass, e is elementary charge, γ^μ are Dirac gamma matrices), and vertices (e.g., QED vertex factor $-ie\gamma^\mu$). An LLM should connect these diagrammatic elements to the mathematical terms in the scattering amplitude calculation according to Feynman rules. **Example: Physical Inaccuracy in AI-Generated Images** Current

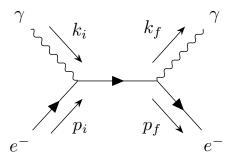


Figure 2: A Feynman diagram for Compton scattering ($\gamma e^- \to \gamma e^-$) in s-channel. LLMs should connect graphical elements (straight lines for fermions, wavy lines for bosons, and vertices for interactions) to mathematical terms in scattering amplitude calculations (e.g., propagators, vertex factors, external leg factors).

 AI image generators lack physical understanding, producing visually appealing but scientifically incorrect visualizations. Figure 3 shows GPT-40's response to "generate an image of a 3D modeling of a two dimensional projected entangled pair state tensor network (4 by 4 square lattice)". The image violates PEPS structure: bulk tensors require exactly five indices (four virtual bonds to neighbors, one physical index), yet many nodes show incorrect connectivity. AI systems fail to encode the physical constraints—here, tensor network geometry and index structure. **Example: AI Material Physicist**

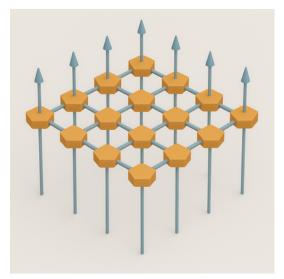


Figure 3: GPT-40-generated PEPS network with incorrect or at least unconventional tensor connectivity. Proper PEPS tensors need 5 indices (4 virtual, 1 physical); many nodes lack required connections.

Imagine a physicist investigating a novel topological material. An 'AI Physicist' assistant might: (1) Synthesize recent literature on related materials and their Berry curvature $\Omega_{n,xy}(\mathbf{k})$ calculations. (2) Assist in formulating a tight-binding Hamiltonian $H(\mathbf{k})$ for the new material based on its crystal structure (e.g., honeycomb lattice for graphene-like systems). (3) Generate Python code using libraries like Kwant or TightBindingTools.jl to numerically calculate the band structure $E_n(\mathbf{k})$ and Chern numbers $C_n = \frac{1}{2\pi} \int_{BZ} d^2k\Omega_{n,xy}(\mathbf{k})$. (4) If numerical results show unexpected edge states, it might help consider analytical approximations (e.g., a low-energy effective Dirac Hamiltonian

 $H_{eff} = v_F(k_x\sigma_y - k_y\sigma_x) + m\sigma_z$, where v_F is the Fermi velocity and m is a mass/gap parameter) to understand their origin. (5) Finally, it might create a slide deck summarizing these findings, 868 869 including generating plots. This collaborative workflow, with the AI handling complex but definable 870 subtasks under human strategic guidance, shows the potential [89]. Example: Verifying analytical 871 calculations by Mathematica Consider the Jordan-Wigner transformation, useful for 1D quantum 872 spin systems. The transverse field Ising model Hamiltonian is $H = -J\sum_{\langle i,j\rangle}\sigma_i^z\sigma_j^z - h\sum_i\sigma_i^x$. 873 The transformation maps spin operators σ_i^{α} to fermionic operators c_j, c_j^{\dagger} , e.g., $\sigma_j^z = 2c_j^{\dagger}c_j - 1$ and 874 $\sigma^x_j = (\prod_{k < j} (1 - 2c_k^\dagger c_k))(c_j + c_j^\dagger)$. An LLM might attempt this transformation and could be asked to 875 verify parts via a Mathematica MCP, such as the anticommutation relation of the fermionic operators after the transformation $\{c_i, c_i^{\dagger}\}$. Using symbolic tools such as Mathematica, the probability of correctness can be increased, if LLMs become better at generating the correct query for a tool and interpreting its output for such (and more nontrivial) operator algebra.