PHYVER: Physics-Grounded Material Claim Verification with Multi-Fidelity Physical Evidence

Jianpeng Chen[*,1], **Wangzhi Zhan**[*,1], **Haohui Wang**[1], **Brian Mayer**[1], **Dongqi Fu**[2], **Dawei Zhou**[1]

[1]Virginia Tech, Blacksburg, VA, USA

[2]Meta AI, Menlo Park, CA, USA

Correspondence: {jianpengc,wzhan24}@vt.edu

Demo Link: http://zhoulab-1.cs.vt.edu:5557/demo

Code Link: https://github.com/cjpcool/PhyVer

Abstract

Material claims in papers, patents, and technical reports often involve physical feasibility (*e.g.*, stability under conditions, property consistency), not just textual feasibility. Yet most claim verifiers operate over language, therefore producing ungrounded judgments. On the other hand, direct first-principles verification (*e.g.*, density functional theory, DFT) is inflexible and hard to invoke from underspecified free-form claims. Therefore, we introduce PHYVER, a physics-grounded material claim verification system that bridges this gap by translating claims into multi-fidelity physical evidence and interpretable verdicts. To support human-in-the-loop inspection, we present an interactive web interface that visualizes the instantiated structure, optimization trajectories, DFT summaries, and the final decision. On expert-labeled claims, PHYVER improves agreement with experts over text-only GPT-5.1, reducing MAE from 1.54 to 1.20 and Signed MAE from 0.95 to 0.82, and increasing Accuracy@$\pm$1 from 50% to 70%.

1 Introduction

Material-related claims appear throughout scientific papers, patents, technical reports, and even everyday product descriptions (Smith et al., 2022). Unlike purely textual factuality, many material claims are related to whether a proposed material configuration is physically feasible and whether its implied properties are consistent with underlying thermodynamic and structural constraints (Zhan et al., 2025; Jain et al., 2013; Chen et al., 2025b). As large language models (LLMs) become a popular interface for scientific reading and discovery, an increasingly practical question emerges: *given a free-form material claim, can we verify it with evidence that is grounded in physics rather than linguistic plausibility?*

Existing claim verification pipelines in natural language processing (NLP) predominantly operate over text: they retrieve and assess evidence sentences, or rely on an LLM's internal priors (Thorne et al., 2018a; Wadden et al., 2020a; Manakul et al., 2023). However, for materials, the core evidence for many material claims is not another sentence in a corpus, but quantitative physical signals, such as whether a structure can relax to a low-energy configuration, whether computed properties align with the claim, and whether the claim remains consistent under reasonable assumptions (Jain et al., 2013). A claim may be commonly stated yet physically implausible; conversely, a novel but valid claim may be absent from the literature and therefore hard to "verify" by retrieval. This mismatch makes text-only verification brittle and encourages failure modes where an LLM produces fluent but ungrounded rationales.

A seemingly natural alternative is to verify material claims by appealing directly to first-principles simulations such as density functional theory (DFT). Yet DFT is not a verifier by itself: it requires a well-defined atomic structure, careful choices of computational settings, and most importantly, a mapping from natural-language claims to executable tests and measurable proxies (Lejaeghere et al., 2016). In practice, the bottleneck is often *knowledge and specification*: translating an underspecified claim into a concrete material candidate (formula and atomic positions), deciding what to compute, and interpreting results against the claim's semantics. As a result, physics tools are powerful but difficult to invoke reliably from natural language, while NLP tools are accessible but lack intrinsic physics awareness. This leaves a clear gap for methods that can translate free-form claims into well-specified simulations and return verdicts grounded in quantitative physical evidence.

We address this gap with **phy**sics-grounded

[*]These authors contributed equally to this work.

[+]A demo video can be accessed at This Link.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 546–555

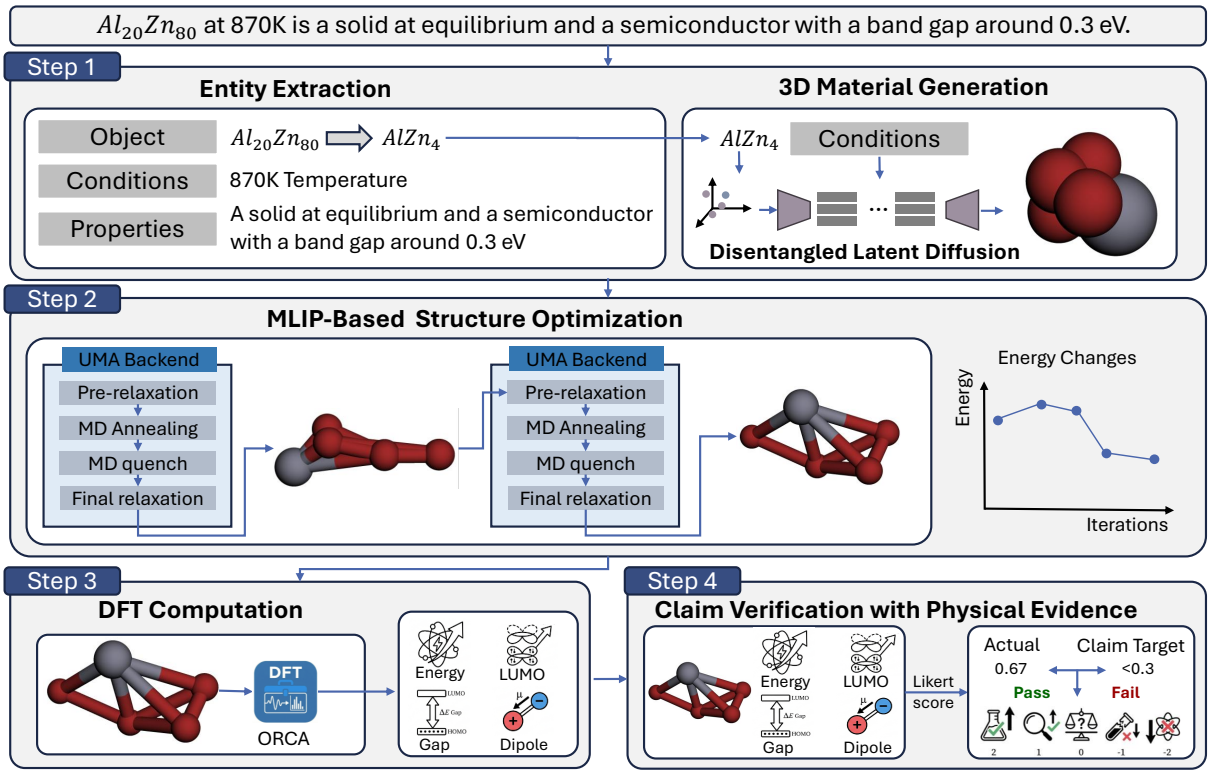

Figure 1: PHYVER system for claim verification contains four steps: Step 1. material instantiation, Step 2. MLIP-based optimization, Step 3. DFT computation, and Step 4. evidence-grounded claim verification.

material claim **ver**ification (PHYVER system), an end-to-end system that makes claim verification both *knowledge-aware* and *physics-grounded*. Given a textual claim, PHYVER at first instantiates material candidates from text, then gathers physical evidence through multi-fidelity simulation (low-fidelity machine-learned interatomic potentials, MLIP, and high-fidelity DFT computation), and finally produces an interpretable verdict explicitly supported by the multi-fidelity evidence. For material instantiation, we introduce a text-to-structure module, which translates free-form descriptions into material structures, including chemical formula and atomic positions. Physical evidence is then acquired efficiently via a multi-fidelity workflow: a fast optimization stage based on MLIP performs iterative structural relaxation toward energy minimization, producing an optimization trajectory and energy-change profile; for higher-confidence confirmation, we selectively invoke DFT to compute key properties needed to evaluate the claim. By grounding verification in physical evidence, the system reduces reliance on linguistic priors and enables judgments that an LLM alone cannot reliably infer.

To demonstrate PHYVER system, we present an interactive demo designed for human-in-the-loop scientific use. Users input a textual claim and receive a structured, inspectable verification report comprising a 3D visualization of the instantiated material structure, the energy evolution during structural optimization, computed DFT properties, and a final verdict with five graded outcomes (*true*, *possibly true*, *uncertain*, *possibly false*, *false*) according to Likert scale (Likert, 1932). Crucially, the interface exposes intermediate artifacts and evidence summaries, allowing users to understand *why* a claim is supported or contradicted, and to iteratively refine ambiguous claims rather than treating verification as a black-box decision.

Our contributions are threefold:

- **System:** We introduce a physics-grounded claim verification pipeline that connects natural-language material claims to executable, evidence-producing workflows, and aggregates multi-fidelity physical evidence (MLIP optimization and DFT) into evidence-grounded verdicts.
- **Demo:** We build an interactive interface that visualizes instantiated structures, optimization trajectories, computed properties,

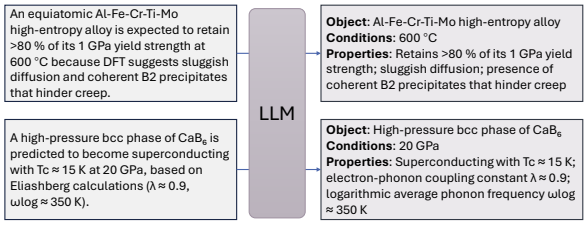

Figure 2: Two example results of entity extraction.

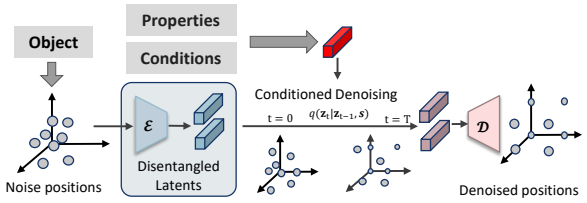

Figure 3: Disentangled latent diffusion model.

and supporting evidence to enable transparent claim verification.

- **Evaluation:** We provide empirical results on expert-rated material claims, demonstrating improved alignment with expert judgments compared to text-only LLM assessment.

2 System Design

LLMs can assess a claim using their internal knowledge, but they struggle when a claim expresses a novel hypothesis or requires physical feasibility beyond what is stated in text. For materials, whether a claim is true often depends on evidence from the 3D physical world, *e.g.*, whether a proposed structure can relax to a stable configuration under realistic conditions. Motivated by this gap, our system performs verification through four steps as shown in Figure 1: Step 1. **3D material instantiation**, which maps a claim to a plausible candidate structure; Step 2. **MLIP-based material optimization** which runs molecular dynamics (MD) to obtain a stable 3D structure and optimization evidence; Step 3. **DFT computation**, which selectively computes claim-relevant physical indicators via a developed DFT toolbox; and Step 4. **evidence-grounded claim verification**, which aggregates the computed evidence to produce a knowledge-aware and physics-grounded verdict and rationale.

2.1 Material Instantiation

Entity Extraction. A free-form material claim is often underspecified for computation, and it may mix composition, operating conditions, and target properties in a single sentence, which makes it difficult to directly instantiate a concrete structure and verify it against physical evidence. To make the claim executable, we first decompose it into a small set of structured entities: (i) the material object (*e.g.*, chemical formula, dopants, or prototype keywords), (ii) conditions (*e.g.*, temperature, pressure, or environment), and (iii) claimed properties (such as stability, phase, or electronic/mechanical descriptors). This structured representation serves as the interface between language and physics, and it determines what candidate structures to generate and what properties to compute and compare. In practice, we find that LLMs are reliable at extracting such entities, especially in the molecular domain (Jung et al., 2024). Therefore, we implement entity extraction by prompting an LLM to return a structured output for the following modules. Figure 2 shows two entity extraction examples.

3D Material Generation. Building on previous work (Chen et al., 2026b), we introduce the disentangled latent diffusion framework to generate a plausible 3D material candidate conditioned on the extracted entities. As illustrated in Figure 3, given the extracted material object (composition) and optional property/condition constraints, the generator first assigns atom types according to the composition and initializes their 3D coordinates by sampling noisy (Gaussian) positions. An encoder then maps this noisy structure into a disentangled latent space consisting of node-wise latents that capture local geometric patterns and a global property latent that summarizes property-relevant semantics. In this latent space, the model performs an iterative denoising diffusion process conditioned on the property and condition entities, gradually steering the sample toward structures that satisfy the claim specification. Finally, a decoder transforms the denoised latents back into an explicit 3D geometry (atomic positions and lattice), which is used as the instantiated candidate for physical optimization.

2.2 MLIP-Based Material Optimization

The structure produced by the previous step is a hypothesis. It is often noisy, metastable, or even physically inconsistent (*e.g.*, unrealistically short bonds) because it is created to match language constraints rather than to minimize energy. Running DFT directly on such a raw candidate is brittle and

misleading. We therefore introduce an optimization stage: an MLIP can approximate the potential-energy surface at orders-of-magnitude lower cost than first-principles methods, allowing us to repair generated structures into a stable local minimum. This process produces MLIP-level fidelity evidence (energy/forces) that is directly useful for claim verification (Faiyad and Martini, 2025).

We implement this stage with UMA model (Wood et al., 2025) as the default MLIP backend, wrapped through ASE (Hjorth Larsen et al., 2017) to standardize optimization, MD, and trajectory logging for inspection in the demo. As shown in Step 2 of Figure 1, because a single relaxation can be trapped in a poor local minimum (especially for generated candidates), we adopt an iterative workflow, *i.e.*, *pre-relax* $\rightarrow$ *anneal* $\rightarrow$ *quench* $\rightarrow$ *re-relax*. Concretely, we repeat an *anneal* $\rightarrow$ *quench* $\rightarrow$ *relax* loop, *i.e.*, heating to escape unfavorable basins, cooling to 0 K to stabilize, and performing 0 K optimization to refine the geometry, while retaining the best-energy structure across loops.

This stage outputs more than a "final structure", and it produces physical evidence that later supports the verdict: (i) the optimized structure (for visualization and as DFT input), (ii) the optimization trajectory (to show whether relaxation was smooth or unstable), and (iii) an energy evolution profile (to quantify whether the candidate consistently moves toward a stable low-energy basin). In the web interface, these are directly rendered as the optimized 3D structure plus an energy/metric plot over time, enabling users to interpret the results. As a demo intended for interactive use, we expose different presets (*e.g.*, quick/sprint/standard/thorough) that adjust the number of loops, peak anneal temperature, cooling schedule, and force thresholds. This lets users quickly screen claims and tune the simulation settings.

2.3 DFT Computation

While MLIP optimization provides low-fidelity evidence, some claims require higher-confidence confirmation or property evaluation. We therefore develop a DFT (Kohn and Sham, 1965) toolbox for selective, higher-fidelity evidence. Our DFT toolbox follows two principles: (i) it bridges fixed, reproducible DFT settings with flexible, claim-specific property queries, (ii) it runs selectively as a higher-fidelity check beyond MLIP evidence. Concretely, we use ORCA as the backend and run single-point energy and analytical gradients (*e.g.*, SP and EnGrad) on a relaxed structure (Neese, 2012, 2022).

The demo exposes both the raw ORCA input/output and the parsed summary, so users can audit the evidence behind a verdict rather than treating DFT as a black box.

2.4 Evidence-Grounded Claim Verification

To avoid text-only verification dominated by linguistic plausibility, we ground the final judgment in physical evidence produced by the MLIP optimization and DFT computation. Given the extracted entities, we translate the claim into a small set of checkable predicates (*e.g.*, feasibility, stability, and property constraints, *etc*), which resolves the specification gap between free-form language and numerical physics tests. We then collect the simulation outputs into a structured evidence record, including MLIP optimization logs (*e.g.*, energy decrease, final stability indicators, convergence flags) and DFT properties (*e.g.*, dipole, HOMO/LUMO gap), and compute an evidence-grounded confidence score by comparing each obtained property with the target in the claim. The evaluation follows the Likert scale standard (Likert, 1932). Specifically, the system outputs a five-level verdict (*true*, *possibly true*, *uncertain*, *possibly false*, *false*) together with concise reasons (Manakul et al., 2023). In this verification step, the LLM is used as an interpreter that maps these physical signals to a five-level Likert verdict and a natural-language rationale. To improve transparency, the interface exposes the intermediate evidence used for the final verdict, allowing users to inspect whether the rationale is consistent with the computed physical signals.

3 Interactive Interface

PHYVER provides a web-based interactive interface that enables step-wise execution and transparent inspection of the full claim verification pipeline. As shown in Figure 4, the interface is organized into two modules: a **control module** (left panel) and a **visualization module** (right panel).

Control Module. The left panel allows users to configure and execute each stage of the pipeline. Users can input a free-form material claim, select the LLM model, choose execution presets (*e.g.*, quick/sprint/standard/thorough), configure hard-

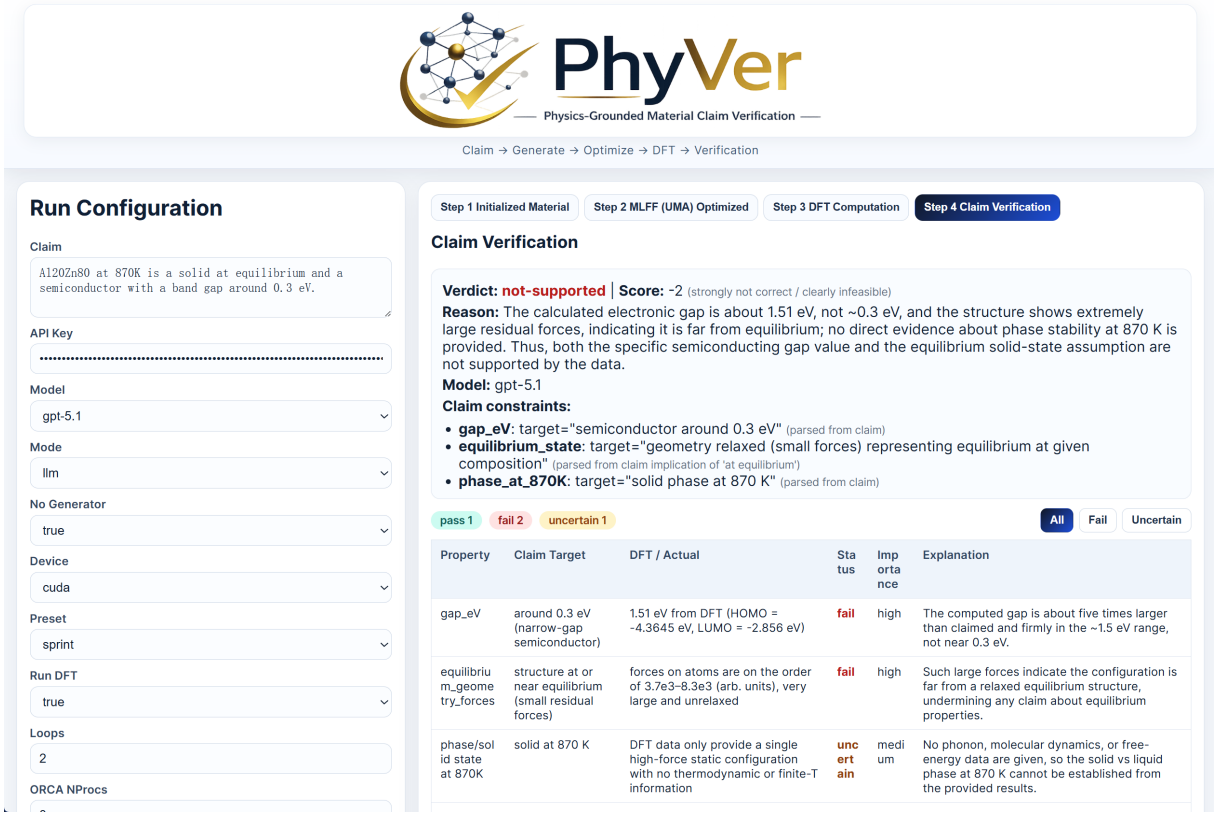

Figure 4: An example of PHYVER interactive interface.

ware and DFT parameters (*e.g.*, device, ORCA cores, memory), and optionally enable or disable DFT. The system supports both step-wise execution (*Generate* → *Optimize* → *DFT* → *Verify*) and one-click end-to-end runs. Each step can be triggered independently, and the current job state, logs, and runtime metadata are displayed in real time. This design enables flexible experimentation, ablation testing, and interactive debugging.

Visualization Module. The right panel presents structured outputs corresponding to each pipeline stage. For **Step 1**, the initialized 3D structure is rendered using an interactive molecular viewer, together with parsed structure metadata. For **Step 2**, the MLIP-optimized structure is shown alongside quantitative optimization evidence, including energy evolution, volume changes, molecular root-mean-square deviation, and trajectory logs. For **Step 3**, DFT results are summarized as structured evidence cards (*e.g.*, total energy, forces, dipole, HOMO/LUMO gap), with raw output accessible for transparency. Finally, **Step 4** displays the verification summary, including verdict, confidence score, constraint-level checks, and explanation, together with the raw JSON response returned by the verification module.

4 Experiments

Case Study. To demonstrate the practicability and effectiveness of PHYVER, we conduct two case studies that are textually plausible but physically infeasible via model o4-mini (Hurst et al., 2024) as LLM backbone. As shown in Figure 5, the case study presents a failing example where PHYVER evaluates a claim by instantiating a structure, running UMA relaxation and DFT, and concluding the claim is not supported due to unstable geometry and failed evidence.

From this case study, we have the following observations: **First**, entity extraction is essential to make an underspecified claim executable, decomposing it into *object/conditions/properties* directly determines what structure to build and what checks to run. **Second**, the MLIP optimization stage provides strong feasibility evidence: large residual forces after relaxation offer a physics-grounded signal that the instantiated interstitial configuration is far from equilibrium, supporting a negative verdict even before relying on textual priors. **Third**, property-level claims that are inherently *comparative* (interstitial *vs.* substitutional)

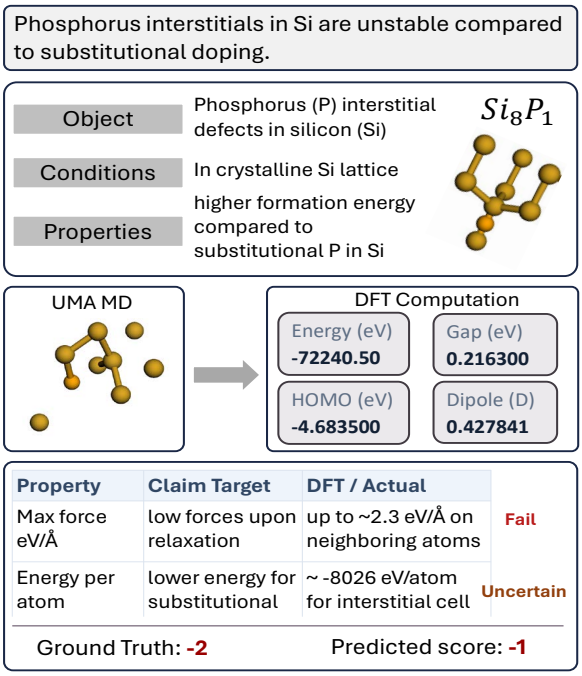

Figure 5: Case study for claim verification. The predicted score and gold score are correctly matched.

Metric	GPT-5.1	PHYVER
MAE ↓	1.5417	**1.2000**
Accuracy@±1 ↑	0.5000	**0.7000**
Signed MAE ↓	0.9545	**0.8182**

Table 1: Comparison between GPT-5.1 (Singh et al., 2025) and PHYVER on expert-labeled material claims.

require matched baselines and consistent reference energies; without an explicitly computed substitutional counterpart, the system can correctly flag the claim as fail, highlighting where additional computations are needed for a verification. More case studies can be found in Appendix A.2.

Quantitative Results. We compare PHYVER-GPT-5.1 with vanilla GPT-5.1 on expert-created material claims, each labeled with a 5-point Likert gold score in $\{-2, -1, 0, 1, 2\}$. We report MAE, Accuracy@±1 (within one Likert level), and Signed MAE (directional bias), with implementation details in Appendix A.1. As shown in Table 1, PHYVER improves all metrics: MAE decreases from 1.54 to 1.20 (22.2% relative reduction) and Accuracy@±1 increases from 50% to 70% (+20 points), indicating more reliable decisions when grounding verification in physical evidence. Signed MAE also drops from 0.95 to 0.82, suggesting reduced systematic bias in the predicted direction.

5 Related Works

LLMs for Claim Verification and Scientific Tool Use. Prior work on claim verification in NLP largely relies on *textual evidence retrieval* followed by entailment-style judgment, ranging from general-domain settings (*e.g.*, FEVER) to scientific variants (*e.g.*, SciFact) (Thorne et al., 2018b; Wadden et al., 2020b). While effective when truth is encoded in the literature, such pipelines do not test *physical feasibility*, which is essential for many material claims. In parallel, *tool-augmented LLM* frameworks connect language models to external solvers or labs (*e.g.*, chemistry toolkits or robotic experimentation) to extend reasoning beyond text (Boiko et al., 2023; M. Bran et al., 2024). Our work follows this direction but differs in purpose: we use the LLM primarily as a *translator* from free-form claims to executable tests, while the final decision is grounded in computed physical evidence rather than the LLMs linguistic priors (Zhong et al., 2024).

Multi-Fidelity Computational Materials Workflows. On the physics side, established infrastructures enable standardized, large-scale *forward* property evaluation from a given structure, including workflow managers and databases built around DFT screening (Jain et al., 2013; Mathew et al., 2017; Huber et al., 2020). More recently, MLIPs have enabled fast structural relaxation and molecular dynamics at near-DFT fidelity, making *multi-fidelity* pipelines (MLIP screening → selective DFT refinement) increasingly common (Batatia et al., 2022; Zuo et al., 2020). However, these systems typically start from a *well-specified structure* and are not designed to verify claims stated in natural language. PHYVER bridges this gap by mapping claims to structures and simulation protocols, and then aggregating MLIP- and DFT-based evidence into an interpretable verification output.

6 Conclusion

PHYVER emphasizes *knowledge-aware*, *physics-grounded* verification for material claims. Rather than relying on textual plausibility, it maps a natural-language claim into structured, physical-checkable predicates guided by domain knowledge (materials entity, conditions, and target properties), then instantiates an executable candidate and validates it with *multi-fidelity physical evidence*. In particular, MLIP-based relaxation of-

fers fast feasibility and stability signals (*e.g.*, convergence trends and residual forces), while optional DFT provides higher-confidence, property-level verification when needed; all evidence is surfaced through an interactive, step-wise interface for transparent inspection. Empirically, this knowledge-aware and physics-grounded verification improves agreement with expert judgment relative to a text-only LLM verifier.

Limitations and Broader Impact

One limitation is that the effectiveness of PHYVER relies on the robustness of the physics computation tools. On the low-fidelity material optimization side, UMA/MLIP enables efficient screening and relaxation, but may be unreliable for out-of-distribution chemistries or bonding environments, leading to biased forces, energies, or relaxed geometries. On the high-fidelity computation side, DFT provides stronger evidence but is computationally expensive, with single-point and gradient calculations often becoming the main bottleneck for large cells, defects, or complex compositions. Its outcomes also depend on methodological choices such as functionals, basis sets/pseudopotentials, and convergence criteria, so non-convergence or small energy differences may affect the final verdict.

By translating free-form claims into executable physical checks, PHYVER can help filter implausible hypotheses, prioritize promising candidates for deeper simulation or experiments, and reduce fluent but ungrounded scientific reasoning. It also connects experimental experts and ML practitioners by presenting inspectable physical evidence while supporting physics-grounded, reproducible model development. PHYVER is intended to complement, not replace, expert judgment and experimental validation.

Acknowledgements

We thank the anonymous reviewers for their constructive comments. This work is supported by the National Science Foundation under Award No. 2339989, No. 2449769 and No. 2406439, DARPA under contract No. HR00112490370 and No. HR001124S0013, U.S. Department of Homeland Security under Grant Award No. 17STCIN00001-08-00, Amazon-Virginia Tech Initiative for Efficient and Robust Machine Learning, Amazon AWS, Google, Cisco, 4-VA, Commonwealth Cyber Initiative, National Surface Transportation Safety Center for Excellence, and Virginia Tech. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

References

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M. Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C. Lawrence Zitnick, and Zachary W. Ulissi. 2024. Open materials 2024 (omat24) inorganic materials dataset and models. *Preprint*, arXiv:2410.12771.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. 2022. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35:11423–11436.

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

Jianpeng Chen, Yawen Ling, Yazhou Ren, Shufei Zhang, and Lifang He. 2026a. Smhgc: Homophily-agnostic multi-view heterophilous graph clustering. *Pattern Recognition*, page 113668.

Jianpeng Chen, Yawen Ling, Jie Xu, Yazhou Ren, Shudong Huang, Xiaorong Pu, Zhifeng Hao, Philip S Yu, and Lifang He. 2025a. Variational graph generator for multiview graph clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11078–11091.

Jianpeng Chen, Wangzhi Zhan, Dongqi Fu, Junkai Zhang, Zian Jia, Ling Li, Wei Wang, and Dawei Zhou. 2026b. Metasymbo: Multi-agent language-guided metamaterial discovery via symbolic latent evolution. *arXiv preprint arXiv:2604.27300*.

Jianpeng Chen, Wangzhi Zhan, Haohui Wang, Zian Jia, Jingru Gan, Junkai Zhang, Jingyuan Qi, Tingwei Chen, Lifu Huang, Muhao Chen, and 1 others. 2025b. Metamatbench: Integrating heterogeneous data, computational tools, and visual interface for metamaterial discovery. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5334–5344.

Abrar Faiyad and Ashlie Martini. 2025. Machine learning interatomic potentials enable molecular dynamics simulations of doped mos2. *Journal of Chemical Theory and Computation*.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves,

Bjørk Hammer, Cory Hargus, and 1 others. 2017. The atomic simulation environmenta python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002.

Sebastiaan P Huber, Spyros Zoupanos, Martin Uhrin, Leopold Talirz, Leonid Kahle, Rico Häuselmann, Dominik Gresch, Tiziano Müller, Aliaksandr V Yakutovich, Casper W Andersen, and 1 others. 2020. Aiida 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Scientific data*, 7(1):300.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and 1 others. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1).

Sung Jae Jung, Hajung Kim, and Kyoung Sang Jang. 2024. Llm-based biological named entity recognition from scientific literature. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 433–435. IEEE.

Walter Kohn and Lu Jeu Sham. 1965. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133.

Kurt Lejaeghere, Gustav Bihlmayer, Torbjörn Björkman, Peter Blaha, Stefan Blügel, Volker Blum, Damien Caliste, Ivano E Castelli, Stewart J Clark, Andrea Dal Corso, and 1 others. 2016. Reproducibility in density functional theory calculations of solids. *Science*, 351(6280):aad3000.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Kiran Mathew, Joseph H Montoya, Alireza Faghaninia, Shyam Dwarakanath, Muratahan Aykol, Hanmei Tang, Iek-heng Chu, Tess Smidt, Brandon Bocklund, Matthew Horton, and 1 others. 2017. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Computational Materials Science*, 139:140–152.

F. Neese. 2012. The orca program system. *WIRES Comput. Molec. Sci.*, 2(1):73–78.

F. Neese. 2022. Software update: the orca program system, version 5.0. *WIRES Comput. Molec. Sci.*, 12(1):e1606.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Andrew Smith, Vinayak Bhat, Qianxiang Ai, and Chad Risko. 2022. Challenges in information-mining the materials literature: a case study and perspective. *Chemistry of Materials*, 34(11):4821–4827.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018b. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020a. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020b. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Brandon M Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R Kitchin, Daniel S Levine, and 1 others. 2025. Uma: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*.

Wangzhi Zhan, Jianpeng Chen, Dongqi Fu, and Dawei Zhou. 2025. Unimate: a unified model for mechanical metamaterial generation, property prediction, and condition confirmation. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org.

Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. 2024. Benchmarking large language models for molecule prediction tasks. *ArXiv*, abs/2403.05075.

Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jorg Behler, Gábor Csányi, Alexander V Shapeev, Aidan P Thompson, Mitchell A Wood, and 1 others. 2020. Performance and cost assessment of machine learning interatomic potentials. *The Journal of Physical Chemistry A*, 124(4):731–745.

A Appendix

A.1 Experiment Settings and Dataset Details

Implementation Details. PHYVER consists of four stages: (1) claim-to-structure generation, (2) MLIP-based optimization, (3) optional DFT computation, and (4) evidence-grounded claim verification.

For **generation**, the proposed disentangled diffusion model is revised from MetaSymbO (Chen et al., 2026b) for adapting to chemical materials, checkpoint trained on the OMAT24 (Barroso-Luque et al., 2024) dataset. The LLM scaffold (default: GPT-5.1) is used only for entity extraction and structured prompting, while structure generation is from the disentangled diffusion model.

For **MLIP-based relaxation**, we use the FairChem Universal ML interatomic potential (UMA-S-1p1) checkpoint (`uma-s-1p1.pt`), accessed via ASE wrappers as described in the repository (Wood et al., 2025). We adopt the `quick` preset unless otherwise specified, which performs multi-loop anneal → quench → 0K relax cycles under GPU acceleration.

For **DFT computation**, we use ORCA (version 5/6 compatible) with the default configuration: `M062X 6-31G* SP EnGrad D3BJ def2/J RIJCOSX TightSCF`. Single-point energy and analytical gradients are computed on the MLIP-relaxed structure, and parsed outputs include total energy, dipole moment, HOMO/LUMO energies, and gap when available.

For **claim verification**, we collect all physical evidence, including MD optimization trajectories, DFT computed properties, and 3D geometries. These multi-fidelity evidences are provided to LLM for quantitative verification.

Training Data. The Disentangled Latent Diffusion generator is trained on the OMAT24 crystal dataset. The UMA model is pretrained on large-scale open catalyst and materials datasets as provided by FairChem. No additional fine-tuning is performed for the quantitative evaluation. LLM backbones are used in inference-only mode.

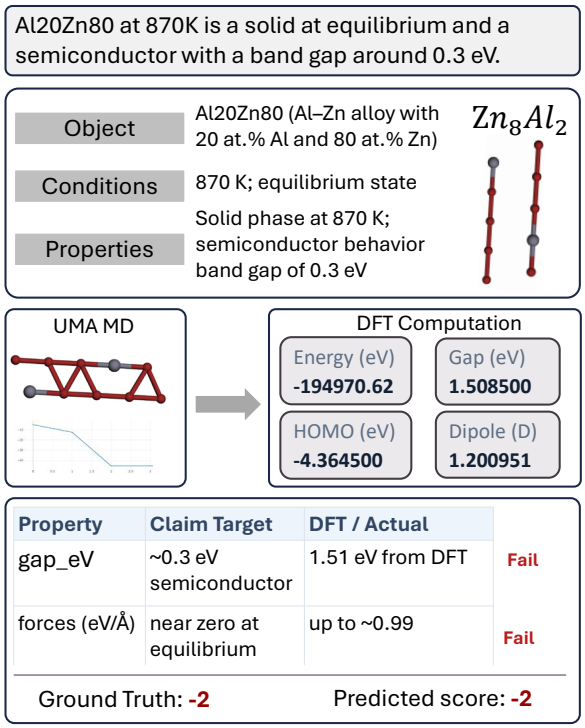

Figure 6: Case study. The predicted score is well matched with the expert-evaluated gold score.

Evaluation Dataset. We evaluate 25 expert-created material claims spanning alloys, semiconductors, and battery systems. Each claim is annotated with a 5-point Likert score in $\{-2, -1, 0, 1, 2\}$ by domain experts. These gold labels are used exclusively for evaluation and were not used during training. We open-sourced the data with gold standard labels at our github codebase.

Quantitative Evaluation Protocol. For each claim, we compare vanilla GPT-5.1 (text-only verification) and PHYVER-GPT-5.1 (LLM + structured physics evidence). We report Mean Absolute Error (MAE), Accuracy@± 1 (prediction within one Likert level), and Signed MAE to assess directional bias. All runs use the same LLM backbone and identical prompts for fairness. PHYVER additionally invokes structure generation, MLIP optimization, and optional DFT computation.

A.2 Additional Case Study

Figure 6 presents a failing example where PHYVER evaluates the claim *$Al_{20}Zn_{80}$ at 870K is a solid at equilibrium and a semiconductor with a band gap around 0.3 eV* and rejects it based on

inconsistent structural stability and electronic evidence.

From this case study, we have the following observations: **First**, entity extraction decomposes the compound claim into multiple verifiable constraints: composition ($Al_{20}Zn_{80}$), thermodynamic condition (870K, equilibrium solid), and electronic property (band gap $\approx$ 0.3 eV), enabling structured and targeted verification rather than a holistic text-level judgment. **Second**, MLIP-based relaxation provides direct physical plausibility signals: the optimized structure exhibits large residual forces (up to $\sim$0.99 eV/Å), contradicting the assumption of equilibrium stability under the claimed condition. **Third**, DFT computation supplies decisive electronic evidence: the calculated band gap ($\sim$1.51 eV) significantly deviates from the claimed 0.3 eV, leading to a clear failure on the semiconducting constraint. By aggregating these independent physical signals, PHYVER outputs a predicted score of -2, fully consistent with the expert-evaluated gold label (-2), demonstrating reliable physics-grounded rejection of an implausible material claim.

A.3 Prompts Used in PHYVER

We summarize the LLM prompts used in PHYVER and clarify how each prompt aligns with the system stages in the following. Please find the detailed prompt in our codebase.

Step 1: Material instantiation.

- `ENTITY_EXTRACTION` (*Claim parsing*): maps a free-form material claim into a structured *claim specification* consisting of *material object* (composition/prototype keywords), *conditions*, and *claimed properties* (what should be verified). This output serves as the executable interface for downstream material generation and physics-tool invocation.

- `INSTRUCTIONS_TRANSLATOR` (*Structure scaffolding*): converts the extracted claim specification into a minimal *generation scaffold* (*e.g.*, FCC/rocksalt/perovskite) with a fixed schema (element types/atomic numbers and fractional coordinates). The scaffold provides a canonical starting point for the generator, improving stability and controllability of the instantiated candidates.

Step 4: Evidence-Grounded Claim Verification.

- `SYSTEM_PROMPT` (*Evidence-grounded verification*): produces the final verdict by reasoning *only* over the provided structured evidence, including the instantiated/relaxed structure summary (Step 1–2 artifacts) and the parsed DFT outputs when available (Step 3). The prompt enforces explicit quantitative checks against claim predicates and outputs a calibrated Likert score with a structured JSON rationale.

Baseline Prompt: LLM-only Scoring Without Physics Evidence.

- `SYSTEM_PROMPT` (*Minimal Likert scoring*): returns a single Likert truthfulness score and a short JSON reason based only on the raw claim (and lightweight metadata), without access to PHYVER's instantiated structures, optimization trajectories, or DFT evidence.