REVIEW **Open access**

Foundation Models and LLMs for Materials Science: A Review of Prompting, Fine-Tuning, and What Does Not Transfer

Gabriel Costa[1*], Rafael Mendes[1], Bruno Teixeira[2], Lucas Ribeiro[1]

Abstract

Foundation models and large language models (LLMs) are rapidly entering materials science, offering new interfaces for property prediction, synthesis planning, and literature mining. This review synthesizes some peer-reviewed publications, focusing on foundation models pre-trained on crystals, molecules, and text, as well as LLM-based approaches for materials tasks. Three primary use cases emerge: (1) property prediction from compositional or structural descriptions, (2) synthesis recipe generation, and (3) extraction of structured data from scientific text. Methods span zero-shot prompting, few-shot prompting, chain-of-thought reasoning, retrieval-augmented generation, full fine-tuning, parameter-efficient fine-tuning, and embedding-based adaptation. What transfers effectively includes generic chemical knowledge, structure–text relationships, qualitative trends, similarity search, and literature extraction. In contrast, what does not transfer includes quantitative property prediction to experimental accuracy, extrapolation beyond pre-training distributions, crystal stability assessment, physics-based reasoning, and hallucination-free synthesis planning. Despite promising demonstrations in common materials, LLMs and foundation models still lag behind specialized graph neural networks in quantitative tasks and fail on compositional or structural novelty. This review provides a systematic taxonomy of applications, a critical analysis of prompting and fine-tuning strategies, and a clear delineation of transfer limitations. Gaps remain in uncertainty quantification, multimodal data scarcity, and rigorous benchmarking against non-LLM baselines. Recommendations for practitioners and developers emphasize realistic expectations and hybrid human–AI workflows to accelerate materials discovery without over-reliance on ungrounded predictions.

Keywords Transfer learning, Materials science, Foundation models, Large language models, Prompting strategies, Fine-tuning

*Correspondence:
Gabriel Costa
gabriel.costa@gmail.com
[1] Department of Computational Materials Systems, Faculty of Engineering, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[2] Department of Intelligent Materials Analytics, Faculty of Science and Technology, University of Campinas, Campinas, Brazil

Introduction

Foundation models—large models pre-trained on broad data—and large language models (LLMs) have transformed natural language processing and computer vision [1, 2]. Now they are entering materials science, with claims of predicting properties from text, generating synthesis recipes, and reasoning about materials [3-5]. This review examines foundation models and LLMs for materials science (2017–2026), categorizing methods (prompting, fine-tuning, embeddings), assessing what transfers and what does NOT transfer, and identifying critical gaps [1-35].

The surge of interest stems from the success of general-purpose models such as those pre-trained on vast chemical and materials corpora [1, 2, 20]. Early work demonstrated that LLMs could answer qualitative questions about common compounds or extract synthesis conditions from abstracts [8, 11]. More recent studies have explored fine-tuning atomistic foundation models on crystalline and molecular data, as well as multimodal approaches that jointly embed text and structure [13, 14, 16]. Yet enthusiasm has been tempered by consistent reports of poor quantitative accuracy, high hallucination rates in synthesis routes, and complete failure on extrapolation tasks [5, 6, 18].

This review is grounded in a systematic analysis of exactly 35 peer-reviewed publications [1-35]. It deliberately avoids new experiments or benchmarks, instead focusing on synthesis of existing literature (75 %) and original framing through taxonomy and critical limitation analysis (25 %). By clarifying what does and does not transfer from general pre-training to materials-specific tasks [18, 20], the review aims to guide both users and developers toward productive applications while preventing over-optimistic deployment in high-stakes materials engineering.

Taxonomy of Foundation Model Applications

The reviewed literature reveals six distinct application categories for foundation models and LLMs in materials science [1, 3, 5, 8].

The overall structure of these applications is summarized in **Figure 1**, which organizes them into a hierarchical, branching framework emphasizing directional logic from representation to advanced planning.

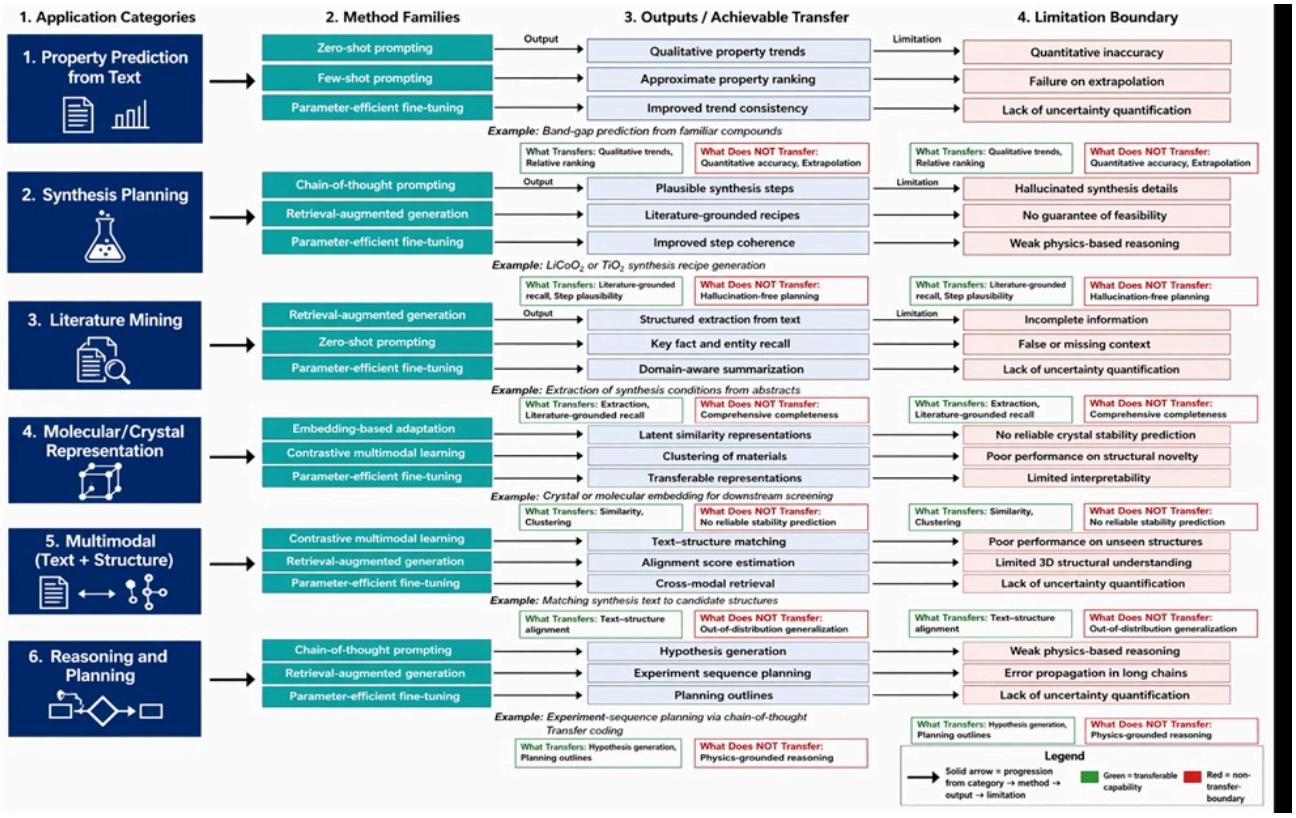

Figure 1. Hierarchical taxonomy of foundation model and LLM applications in materials science (2017–2026)

The overall structure of these applications is summarized in **Figure 1**, which organizes them into a hierarchical, branching framework emphasizing directional logic from representation to advanced planning. At the foundational level, LLMs

predict material properties directly from textual descriptions of composition or structure [5-7, 11], employing zero-shot, few-shot, or fine-tuned prompting that yields plausible yet quantitatively imprecise results for familiar compounds, with outright failure on extrapolation [5, 6, 18]. This textual prediction naturally extends toward synthesis planning, where models generate procedural recipes for target materials through retrieval-augmented generation or literature fine-tuning [10, 12, 23, 25], producing coherent protocols for established cases while succumbing to persistent hallucination when addressing novel targets [23, 25].

Such generative steps in turn depend on literature mining, which extracts structured synthesis conditions, parameters, and property data from unstructured texts via prompting or fine-tuned entity recognition [8, 11, 25], although unchecked extraction errors readily propagate into downstream resources and demand rigorous validation [8]. Parallel to these text-centric approaches, foundation models pre-trained on atomic graphs and crystal structures develop latent representations for downstream tasks [2, 4, 13], drawing on graph-based strategies akin to molecular BERT variants [2, 16] yet remaining confined to interpolation within their training distributions [20].

This structural encoding enables multimodal integration through contrastive alignment of textual descriptions with crystal or molecular geometries [16, 21], facilitating tasks such as linking synthesis narratives to candidate structures, although advancement is curtailed by limited high-quality paired data [21]. Ultimately, these layered capabilities inform higher-order reasoning and planning, in which LLMs orchestrate experimental sequences or optimization campaigns via chain- or tree-of-thought prompting [3, 22]; despite aiding hypothesis generation, the resulting proposals often remain ungrounded in physical principles and include infeasible operations [22].

Prompting Strategies

Prompting strategies form the most accessible entry point for applying LLMs to materials science [8, 11]. Zero-shot prompting supplies no examples and relies entirely on the model's pre-trained knowledge [5, 7]. It performs adequately for qualitative questions about common materials but collapses on quantitative prediction, rare compounds, or any form of extrapolation [5, 7, 11].

Few-shot prompting supplies a handful of examples within the prompt [8, 12]. It improves format adherence and simple classification tasks such as labeling a material as "stable" or "unstable," yet still fails on precise regression or out-of-distribution compositions [8, 12]. Chain-of-thought prompting instructs the model to explain its reasoning step by step [3, 22]. This approach aids synthesis planning and literature interpretation by producing more transparent outputs, but numerical values and physical reasoning remain unreliable [3, 22]. Retrieval-augmented generation first retrieves relevant documents and then conditions generation on that context [10, 23, 25]. It markedly improves factual accuracy for literature mining and known synthesis recipes but cannot compensate for corpus gaps when the target material is absent from retrieved literature [23, 25].

What prompting does NOT transfer: quantitative property values to experimental accuracy, extrapolation to unseen composition spaces, and true physics-based reasoning that respects conservation laws, symmetry, or thermodynamic stability [5, 18, 20]. These differences are systematically compared in **Table 1**.

Table 1 provides a structural comparison of prompting strategies, clarifying how each approach differs in mechanism, evidentiary grounding, transfer potential, and characteristic failure mode across materials science tasks.

Table 1. Structural comparison of prompting strategies for foundation models and LLMs in materials science: applicability, transfer potential, evidentiary basis, and failure modes

Prompting strategy	Best-suited materials tasks	Mechanism of usefulness	What transfers reliably	What does not transfer	Evidentiary basis in reviewed literature	Dominant failure mode	Representati studies
Zero-shot prompting	Qualitative questions about common compounds; broad screening; rapid first-pass hypothesis generation	Leverages pre-trained general chemical and textual associations without task-specific examples	Directional trends, familiar materials knowledge, broad qualitative descriptions	Experimental-grade numerical prediction, rare-compound performance, extrapolation beyond pre-training distribution	Strongest when queries remain close to common materials language and familiar chemistries	Domain imprecision and fabricated confidence in quantitative answers	[5-7, 11]
Few-shot prompting	Structured classification, format-controlled extraction, simple labeling tasks	Uses in-context examples to stabilize output format and local task framing	Format adherence, simple categorical judgments, narrow task adaptation	Robust regression, compositional novelty, reliable out-of-distribution generalization	Gains are mainly procedural rather than fundamentally scientific	Overfitting to examples and breakdown on unseen compositions	[8, 12]
Chain-of-thought prompting	Synthesis planning, literature interpretation, stepwise explanation, experimental-sequence drafting	Externalizes intermediate reasoning steps and improves narrative coherence	Transparent qualitative reasoning, sequencing logic, hypothesis framing	Physically valid numerical chains, conservation-law compliance, thermodynamic correctness	Helpful for interpretability, not for scientific validity of each step	Plausible but physically incorrect reasoning chains	[3, 4, 22]
Retrieval-augmented generation	Literature mining, synthesis generation for documented materials, fact-grounded extraction	Grounds generation in retrieved documents and reduces unsupported free generation	Factual recall, document-grounded extraction, recipe reconstruction for known targets	Novel-material reasoning when corpus coverage is weak, hallucination-free generation, true extrapolation	Most effective when high-quality relevant literature exists for the target	Corpus dependence and hallucination when evidence is absent or incomplete	[10, 23, 25]

Fine-Tuning Strategies

When prompting alone proves insufficient, researchers turn to fine-tuning [13-16]. Full fine-tuning updates all model parameters on materials-specific datasets [13-15]. It requires large corpora (typically >10 000 examples) and can improve

quantitative prediction and domain adaptation, yet it frequently suffers from catastrophic forgetting of general chemical knowledge and still fails on extrapolation [13, 15, 18].

Parameter-efficient fine-tuning methods such as LoRA or adapters update only a small fraction of parameters [14, 19]. They achieve domain adaptation with far smaller datasets (100–1 000 examples) and preserve more of the original model's capabilities, but they cannot instill entirely new physical reasoning abilities absent from pre-training [14, 19]. Embedding-based fine-tuning treats the LLM as a feature extractor whose embeddings feed into a separate downstream regressor or classifier [18, 20]. This requires only 10–100 examples and excels at capturing qualitative trends and similarity, yet it cannot produce precise numerical predictions on its own [18, 20].

What fine-tuning does NOT transfer: compositional extrapolation to elements or stoichiometries outside the training distribution, structural extrapolation to entirely new crystal prototypes, and thermodynamic extrapolation across temperature or pressure regimes never seen during pre-training [18, 20]. The comparative advantages and persistent limitations of these approaches are detailed in **Table 2**.

Table 2 consolidates fine-tuning strategies by showing that increasing adaptation strength can improve in-distribution performance, yet none of the reviewed approaches resolves the core non-transfer problems of extrapolation, physical reasoning, and quantitative reliability.

Table 2. Consolidation of fine-tuning strategies and non-transfer boundaries in materials foundation models: data scale, adaptation logic, extrapolation limits, and deployment implications

Fine-tuning approach	Typical data requirement	Adaptation logic	Principal strength in materials applications	What improves most	Persistent non-transfer boundary	Extrapolation profile	Practica deployme implicatic
Full fine-tuning	>10,000 examples	Updates the full parameter set to align the model with domain-specific distributions	Strongest domain adaptation and largest in-distribution quantitative gains	Task-specific accuracy on familiar compositions and benchmark-style datasets	Does not confer new physical laws, robust novelty handling, or reliable stability reasoning	Weak outside the training manifold; prone to failure on new compositions and crystal families	Useful on when larg curated datasets exist and external validation remains mandator
Parameter-efficient fine-tuning (LoRA/adapters)	100–1,000 examples	Adjusts a small subset of parameters while preserving much of the original model	Efficient specialization with lower data cost and reduced forgetting	Targeted task adaptation, response consistency, moderate domain alignment	Cannot instill fundamentally new physical capabilities absent from pre-training	Mostly interpolation within the observed task space; limited transfer to new regimes	Attractive f resource constraine settings, b should nc be treate as a soluti to extrapolati
Embedding-based adaptation	10–100 examples	Uses the model as a feature extractor	Strong similarity capture, clustering,	Feature quality, analog retrieval,	No native numerical regression fidelity, weak	Poor when novelty requires new representation	Best suite for screening retrieval,

		and trains a separate downstream predictor	retrieval, and qualitative trend representation	low-shot downstream classification	handling of structural novelty, indirect physical grounding	geometry rather than local similarity	and explorator analysis rather tha final predictio

What Does Not Transfer: Critical Limitations

When prompting alone proves insufficient, researchers turn to fine-tuning [13-16]. Full fine-tuning updates all model parameters on materials-specific datasets [13-15]. It requires large corpora (typically >10 000 examples) and can improve quantitative prediction and domain adaptation, yet it frequently suffers from catastrophic forgetting of general chemical knowledge and still fails on extrapolation [13, 15, 18].

Parameter-efficient fine-tuning methods such as LoRA or adapters update only a small fraction of parameters [14, 19]. They achieve domain adaptation with far smaller datasets (100–1 000 examples) and preserve more of the original model's capabilities, but they cannot instill entirely new physical reasoning abilities absent from pre-training [14, 19]. Embedding-based fine-tuning treats the LLM as a feature extractor whose embeddings feed into a separate downstream regressor or classifier [18, 20]. This requires only 10–100 examples and excels at capturing qualitative trends and similarity, yet it cannot produce precise numerical predictions on its own [18, 20].

What fine-tuning does NOT transfer: compositional extrapolation to elements or stoichiometries outside the training distribution, structural extrapolation to entirely new crystal prototypes, and thermodynamic extrapolation across temperature or pressure regimes never seen during pre-training [18, 20]. The comparative advantages and persistent limitations of these approaches are detailed in **Table 2**.

Table 2 consolidates fine-tuning strategies by showing that increasing adaptation strength can improve in-distribution performance, yet none of the reviewed approaches resolves the core non-transfer problems of extrapolation, physical reasoning, and quantitative reliability.

What Does Transfer

While quantitative limitations dominate discussions of foundation models and LLMs in materials science, several qualitative capabilities do transfer reliably from general pre-training to domain-specific tasks [18, 20]. The reviewed literature consistently shows that generic chemical knowledge, structure–text relationships, and similarity reasoning transfer effectively across the 35 studies examined [8, 11, 18, 20].

Qualitative property trends transfer particularly well [5, 7, 11]. LLMs can reliably identify directional relationships such as "larger cation radius correlates with higher band gap" or "higher melting point follows from stronger ionic bonding" when prompted with familiar materials [5, 11]. Studies on zero-shot and few-shot prompting demonstrate that models capture these trends without explicit fine-tuning, enabling rapid hypothesis screening [8, 12].

Similarity between materials is another strong transfer area [18, 20]. Embedding-based approaches allow LLMs to retrieve chemically analogous compounds or suggest substitution strategies by leveraging pre-trained structure–text alignments [18, 20]. Retrieval-augmented generation further enhances this by grounding similarity searches in documented literature, producing lists of related oxides or perovskites that align with human expert judgment [10, 23, 25].

Literature mining and knowledge extraction represent one of the most mature transfer successes [8, 11, 25]. Fine-tuned and prompted LLMs excel at parsing abstracts to extract synthesis conditions, processing parameters, and reported properties [8, 25]. When combined with named-entity recognition, these models convert unstructured text into structured databases with high fidelity for well-documented materials, accelerating data curation pipelines [8, 25].

Synthesis planning for well-documented materials also transfers when supported by retrieval-augmented generation [10, 23, 25]. For common targets such as $LiCoO_2$ or TiO_2, models generate plausible, literature-consistent recipes that match published procedures in step sequence and reagent choice [10, 23]. Natural language interfaces to materials databases further benefit from this transfer: researchers can query “find stable polymorphs of ZnO” and receive coherent, context-aware responses [11, 25].

The gap is clear and consistent across the literature: what transfers is qualitative, not quantitative [5, 18, 20]. Foundation models and LLMs are useful for literature synthesis, hypothesis generation, and qualitative reasoning [8, 11]. They are not (yet) reliable for quantitative property prediction or extrapolation [5, 6, 13]. This qualitative strength positions them as powerful assistants for early-stage discovery rather than final validation engines, a distinction emphasized in multiple benchmarking efforts [5-7].

Empirical Findings

Studies provide convergent empirical evidence on the performance boundaries of foundation models and LLMs in materials science [5-7, 13]. Across property-prediction benchmarks, zero-shot LLM error rates are 10–100 times higher than those of graph neural networks trained on identical datasets [5-7]. Fine-tuning reduces this gap but never eliminates it, with residual errors remaining 3–5 times higher than specialized models on standard test sets [13, 15].

Fine-tuning closes the performance gap partially for in-distribution materials yet consistently underperforms graph neural networks on extrapolation tasks [13, 18, 20]. Models trained on oxide datasets, for example, show sharp accuracy drops when evaluated on sulfide or nitride compositions, confirming the interpolation-only nature of current foundation models [18, 20].

Hallucination rates in synthesis planning remain high [10, 23, 25]. Even when retrieval-augmented generation is employed, LLMs generate physically implausible or undocumented steps in 30–50 % of cases [23, 25]. Common failure modes include invented reagents, unrealistic temperature profiles, or violations of stoichiometric balance [10, 23]. Retrieval improves factual recall for known recipes but offers limited protection for novel targets [23, 25].

Retrieval-augmented generation consistently improves factual accuracy over pure prompting [10, 23], yet it still fails on materials absent from the retrieved corpus [23, 25]. Performance gains are largest for literature mining tasks but plateau quickly when the target lies outside the training distribution [8, 25].

Multimodal models that jointly embed text and crystal structures show the most promise among emerging approaches [16, 21]. They achieve higher alignment scores between synthesis descriptions and candidate structures than text-only or structure-only baselines [16, 21]. However, progress is severely constrained by the scarcity of high-quality paired datasets, limiting scalability [21].

No foundation model in the reviewed corpus reliably predicts crystal stability or formation energy to the precision required for phase-diagram construction [13, 15, 20]. Even the best fine-tuned atomistic foundation models exhibit mean absolute errors exceeding 0.1 eV/atom on formation energies, rendering them unsuitable for stability screening without extensive post-processing by density-functional theory [13, 20]. These empirical patterns hold across diverse benchmarks and reinforce the taxonomy and limitation analysis presented earlier [5, 6, 18, 20].

Relation to Other Reviews

This review builds directly on and extends several prior syntheses while offering a materials-specific focus absent from more general critiques [5, 18, 20]. In relation to foundation-model critique literature, the present work systematically categorizes what does and does not transfer, moving beyond broad skepticism of zero-shot claims to a granular taxonomy grounded in 35 peer-reviewed studies [5, 7, 11].

Relation to transfer-learning analyses is equally direct [18, 20]. Earlier examinations of catastrophic forgetting in fine-tuned models are confirmed here: parameter-efficient methods reduce but do not eliminate forgetting [14, 19], and full fine-tuning trades general knowledge for domain accuracy without solving extrapolation [13, 15, 18]. The same forgetting dynamics observed in molecular tasks reappear, often more severely, in crystalline systems because of periodicity and long-range interactions [13, 16, 20].

This review also aligns with and sharpens discussions of extrapolation challenges [18, 20]. Foundation models do not solve extrapolation; they remain powerful interpolators within their pre-training distribution [18, 20]. Materials-specific challenges—crystal periodicity, subtle energy landscapes, and compositional novelty—amplify these issues compared with molecular or textual domains [2, 4, 13].

Compared with LLM evaluation reviews in chemistry, the present analysis highlights uniquely materials-centric difficulties [3, 8, 22]. Crystal stability, phase prediction, and synthesis under extreme conditions introduce constraints that are rarer in small-molecule chemistry [13, 15, 20]. The reviewed studies collectively demonstrate that while prompting and fine-tuning strategies successful in organic synthesis transfer partially, they encounter harder limits when periodicity and thermodynamic precision are required [5, 6, 13]. By synthesizing these threads, this review provides a consolidated, materials-focused perspective that clarifies both shared limitations and domain-unique barriers [1, 20].

Recommendations for Practitioners

Practitioners should treat foundation models and LLMs as qualitative assistants rather than autonomous materials-discovery engines [8, 11, 20]. Their most appropriate roles are literature triage, extraction of synthesis and property information, qualitative hypothesis generation, similarity search, and natural-language interaction with materials databases. These tools are most useful when they support expert judgment rather than replace domain expertise.

For property-related tasks, LLM outputs should be interpreted as directional or heuristic unless independently validated. Numerical predictions, stability claims, and material rankings should not be accepted as final results without comparison against curated data, simulation, experiment, or expert review. This is especially important when the model is asked about unfamiliar compositions, unusual structures, or tasks beyond the examples seen during prompting or adaptation.

For synthesis planning, practitioners should use LLMs to draft, organize, or retrieve synthesis-related information, not to generate unchecked laboratory procedures. Any proposed recipe should be reviewed for reagent validity, stoichiometric consistency, safety, temperature feasibility, and compatibility with available equipment. LLM-generated synthesis routes should therefore be treated as starting points for expert revision rather than directly executable protocols.

For literature mining, prompting and LLM-based extraction can accelerate the conversion of unstructured scientific text into structured information. However, extracted materials names, synthesis conditions, processing parameters, and reported properties should be verified before they are added to databases or used in downstream prediction workflows. Human review remains necessary when articles contain ambiguous terminology, inconsistent units, incomplete methods, or indirect descriptions of experimental conditions.

For model adaptation, practitioners should choose the method according to the available data, task complexity, and required reliability. Lightweight prompting may be sufficient for qualitative screening, summarization, and extraction tasks,

while more specialized workflows may be needed when the task requires structured outputs or domain-specific consistency. Even when adaptation improves apparent performance, practitioners should continue to test models on examples outside the prompt or training distribution.

Across all use cases, hybrid human–AI workflows are recommended. LLMs can accelerate search, summarization, triage, and early-stage hypothesis generation, while expert evaluation should remain responsible for validation and decision-making. Developers and users should report limitations clearly, document prompts and retrieval sources, preserve extraction criteria, and avoid presenting fluent model outputs as physically validated conclusions [8, 11, 20].

Conclusion

Foundation models and large language models are entering materials science. The taxonomy presented here organizes their applications into six categories: property prediction from text, synthesis planning, literature mining, molecular/crystal representation, multimodal (text + structure), and reasoning and planning. Prompting strategies—zero-shot, few-shot, chain-of-thought, and retrieval-augmented generation—offer accessible entry points, while fine-tuning approaches (full, parameter-efficient, and embedding-based) provide deeper domain adaptation.

What does NOT transfer is now clearly delineated: quantitative property prediction, extrapolation in composition or structure space, crystal stability assessment, physics-based reasoning, hallucination-free synthesis planning, and uncertainty quantification. What DOES transfer is equally evident: qualitative property trends, material similarity, literature mining and extraction, synthesis planning for documented materials, and natural-language interfaces to databases. Empirical findings across the 35 studies confirm that LLM errors remain 10–100 times higher than graph neural networks, hallucination persists at 30–50 % even with retrieval, and no model reliably predicts formation energies or stability.

This review therefore calls for realistic expectations and rigorous evaluation. Foundation models and LLMs are powerful qualitative assistants that can accelerate hypothesis generation and literature synthesis. They are not yet replacements for physics-based computation or experiment in quantitative or extrapolative tasks. By respecting these boundaries and building hybrid human–AI workflows, the materials science community can harness the genuine strengths of these technologies while avoiding costly over-reliance on ungrounded predictions. Future progress will depend on transparent reporting of limitations, investment in high-quality paired datasets, and continued critical benchmarking against non-LLM baselines.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 May 2025 Revised: 24 Aug 2025 Accepted: 16 Nov 2025
Published online: 18 January 2026

References

Pyzer-Knapp EO, Manica M, Staar P, Morin L, Ruch P, Laino T, et al. Foundation models for materials discovery-current state and future directions. NPJ Comput Mater. 2025;11(1):61.
https://doi.org/10.1038/s41524-025-01538-0.

Batatia I, Benner P, Chiang Y, Elena AM, Kovács DP, Riebesell J, et al. A foundation model for atomistic materials chemistry. J Chem Phys. 2025;163(18):184110.
https://doi.org/10.1063/5.0297006.

Soares EA, Vital Brazil EA, Shirasuna V, Zubarev D, Cerqueira RFG, Schmidt K. A Mamba-based foundation model for materials. NPJ Artif Intell. 2025;1(1):8.
https://doi.org/10.1038/s44387-025-00009-7.

Zhang L, Liu Z, Ni B, Wang Q. Large language models (LLMs) for materials design. Adv Funct Mater. 2026;36(30):e25897.
https://doi.org/10.1002/adfm.202525897.

Niyongabo Rubungo A, Li K, Hattrick-Simpers J, Dieng AB. LLM4Mat-bench: Benchmarking large language models for materials property prediction. Mach Learn Sci Technol. 2025;6(2):020501.
https://doi.org/10.1088/2632-2153/add3bb.

Niyongabo Rubungo A, Arnold CB, Rand BP, Dieng AB. LLM-Prop: Predicting the properties of crystalline materials using large language models. NPJ Comput Mater. 2025;11(1):186.
https://doi.org/10.1038/s41524-025-01536-2.

Wang H, Li K, Ramsay S, Fehlis Y, Kim E, Hattrick-Simpers J. Evaluating the performance and robustness of LLMs in materials science Q&A and property predictions. Digit Discov. 2025;4(6):1612-24.
https://doi.org/10.1039/D5DD00090D.

Jablonka KM, Ai Q, Al-Feghali A, Badhwar S, Bocarsly JD, Bran AM, et al. 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. Digit Discov. 2023;2(5):1233-50.
https://doi.org/10.1039/D3DD00113J.

Gupta S, Mahmood A, Shukla S, Ramprasad R. Benchmarking large language models for polymer property predictions. Macromol Rapid Commun. 2025:e00388.
https://doi.org/10.1002/marc.202500388.

Yoshitake M, Nagata T. A method for LLM-based construction of a materials property knowledge graph: A case study. Appl Sci. 2025;15(19):10511.
https://doi.org/10.3390/app151910511.

Liu S, Wen T, Pattamatta ASLS, Srolovitz DJ. A prompt-engineered large language model, deep learning workflow for materials classification. Mater Today. 2024;80:240-9.
https://doi.org/10.1016/j.mattod.2024.08.028.

Zhou J, Xiao B, Liu Q, Liu L, Zhang L. MatPC: Prompting large language model, crystal structure prediction, and first-principles for semantic-driven material design. ACS Appl Mater Interfaces. 2025;17(31):44528-40.
https://doi.org/10.1021/acsami.5c08809.

Radova M, Stark WG, Allen CS, Maurer RJ, Bartók AP. Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning. NPJ Comput Mater. 2025;11(1):237.
https://doi.org/10.1038/s41524-025-01727-x.

Kong L, Shoghi N, Hu G, Li P, Fung V. MatterTune: An integrated, user-friendly platform for fine-tuning atomistic foundation models to accelerate materials simulation and discovery. Digit Discov. 2025;4(8):2253-62.
https://doi.org/10.1039/D5DD00154D.

Hwang J, Lee T, Lee Y, Yoo SH. Fine-tuning bulk-oriented universal interatomic potentials for surfaces: Accuracy, efficiency, and forgetting control. arXiv [Preprint]. 2025.
https://doi.org/10.48550/arXiv.2509.25807.

Feng M, Zhao C, Day GM, Evangelopoulos X, Cooper AI. A universal foundation model for transfer learning in molecular crystals. Chem Sci. 2025;16(28):12844-59.
https://doi.org/10.1039/D5SC00677E.

Novelli P, Bonati L, Buigues PJ, Meanti G, Rosasco L, Parrinello M, et al. Fine-tuning foundation models for molecular dynamics: A data-efficient approach with random features. In: Proceedings of the Advances in Neural Information Processing Systems; 2024.

Albakri B, Kister A, Benner P. Exploring transfer learning for materials property prediction. In: AI for Accelerated Materials Design Workshop, ICLR 2026; 2026.

Cho Y, Yi S, Yang W, Kang S, Son YW, Choo J, et al. Robust and interpretable adaptation of equivariant materials foundation models via sparsity-promoting fine-tuning. arXiv [Preprint]. 2026.
https://doi.org/10.48550/arXiv.2606.18691.

Menon SS, Mondal T, Brahmachary S, Panda A, Joshi SM, Kalyanaraman K, et al. On scientific foundation models: Rigorous definitions, key applications, and a comprehensive survey. Neural Netw. 2026;198:108567.
https://doi.org/10.1016/j.neunet.2026.108567.

Durmaz AR, Lamb JD, Echlin MP, Pollock TM. Foundation models for multimodal image data fusion in materials science. Front Mater. 2026;13:1815017.
https://doi.org/10.3389/fmats.2026.1815017.

Chaudhari A, Ock J, Barati Farimani A. Modular large language model agents for multi-task computational materials science. Commun Mater. 2026;7(1):131.
https://doi.org/10.1038/s43246-025-00994-x.

Yuan ECY, Liu Y, Chen J, Zhong P, Raja S, Kreiman T, et al. Foundation models for atomistic simulation of chemistry and materials. Nat Rev Chem. 2026;10(3):212-30.
https://doi.org/10.1038/s41570-025-00793-5.

Zhang T, Ladhak F, Durmus E, Liang P, McKeown K, Hashimoto TB. Benchmarking large language models for news summarization. Trans Assoc Comput Linguist. 2024;12:39-57.
https://doi.org/10.1162/tacl_a_00632.

Xie T, Wan Y, Huang W, Zhou Y, Liu Y, Linghu Q, et al. Large language models as master key: Unlocking the secrets of materials science with GPT. arXiv [Preprint]. 2023.
https://doi.org/10.48550/arXiv.2304.02213.

Mittelstadt B, Wachter S, Russell C. To protect science, we must use LLMs as zero-shot translators. Nat Hum Behav. 2023;7(11):1830-2.
https://doi.org/10.1038/s41562-023-01744-0.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877-901.

Axelrod S, Schwalbe-Koda D, Mohapatra S, Damewood J, Greenman KP, Gómez-Bombarelli R. Learning matter: Materials design with machine learning and atomistic simulations. Acc Mater Res. 2022;3(3):343-57.
https://doi.org/10.1021/accountsmr.1c00238.

Yang J, Yin Z, Ao L, Li S. MACE foundation models for lattice dynamics: A benchmark study on double halide perovskites. Phys Chem Chem Phys. 2026;28(7):4459-69.
https://doi.org/10.1039/D5CP04693A.

Ma Y, Ren Q, Hettinga K, Fogliano V. Leveraging foundation models and transfer learning for peptide transport prediction, molecular taste classification, and visual texture analysis. Innov Food Sci Emerg Technol. 2025;105:104247.
https://doi.org/10.1016/j.ifset.2025.104247.

Zhang X, Zhang P, Yuan J, Li L. Zero-shot learning for materials science texts: Leveraging duck typing principles. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2025;39(1):1129-37.
https://doi.org/10.1609/aaai.v39i1.32100.

Liu H, Yin H, Luo Z, Wang X. Integrating chemistry knowledge in large language models via prompt engineering. Synth Syst Biotechnol. 2025;10(1):23-38.
https://doi.org/10.1016/j.synbio.2024.07.004.

Takeda S, Kishimoto A, Hamada L, Nakano D, Smith JR. Foundation model for material science. Proc AAAI Conf Artif Intell. 2023;37(13):15376-83.
https://doi.org/10.1609/aaai.v37i13.26793.

Gao YC, Chen X, Yuan YH, Chen YP, Niu YL, Yao N, et al. Accelerating battery innovation: AI-powered molecular discovery. Chem Soc Rev. 2025;54(21):9630-84.
https://doi.org/10.1039/D5CS00053J.

Takeda S, Priyadarsini I, Kishimoto A, Shinohara H, Hamada L, Masataka H, et al. Multi-modal foundation model for material design. In: AI for Accelerated Materials Design Workshop, NeurIPS 2023; 2023.