Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: F. Yang, W. Chen and J. D. Evans, *Digital Discovery*, 2025, DOI: 10.1039/D5DD00499C.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D5DD00499C

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Large language models in materials science and the need for open-source approaches

Fengxu Yang,[a] Weitong Chen[b] and Jack D. Evans*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Large language models (LLMs) are rapidly transforming materials science. This review examines recent LLM applications across the materials discovery pipeline, focusing on three key areas: mining scientific literature, predictive modelling, and multi-agent experimental systems. We highlight how LLMs extract valuable information, such as synthesis conditions from text, learn structure-property relationships, and can coordinate agentic systems integrating computational tools and laboratory automation. While progress has been largely dependent on closed-source commercial models, our benchmark results demonstrate that open-source alternatives can match performance while offering greater transparency, reproducibility, cost-effectiveness, and data privacy. As open-source models continue to improve, we advocate their broader adoption to build accessible, flexible, and community-driven AI platforms for scientific discovery.

1 Introduction

Large Language Models (LLMs) have been evolving at an unprecedented pace over the last few years and have shown remarkable capabilities across a wide range of tasks and domains.[1,2] Trained on immense and diverse data from across different knowledge areas, they have developed an extraordinary ability to understand, process, and generate complex text.[3] Beyond general-purpose applications such as education[4] and health[5], LLMs are increasingly emerging as powerful tools for scientific research.

The discovery of materials, such as metal-organic frameworks (MOFs)[7–9], has become pivotal to breakthroughs in areas like energy storage, catalysis, and chemical separation.[10,11] However, the traditional labour-intensive and trial-and-error research paradigm in materials research moves at a slow pace. Decades of research have produced a vast, yet fragmented, archive of knowledge scattered across millions of scientific publications, among other data repositories.[12] Therefore, systematically extracting unstructured data with minimal manual effort is crucial for data collection and analysis. Traditional techniques such as regular expressions (RegEx) and part-of-speech tagging rely heavily on rule-based patterns. Although transformer-based extraction methods represent a significant shift toward statistical learning, they are still often constrained by the scope of carefully curated training datasets. As a result, these approaches may struggle to capture the full diversity and variability of natural language expressions.[13] In contrast, by leveraging their extensive pre-trained knowledge, LLMs offer an advanced opportunity to understand the intricate chemistry language and textual context, enabling them to process unstructured data with higher flexibility without the need for domain-specific training or fine-tuning.[14,15]

Building upon clean, actionable data as a foundation unlocks new possibilities in materials science research. One of the ambitious goals in materials science research is to establish the structure-property relationships that govern material performance. By training on vast and comprehensive datasets of chemical information, LLMs can potentially learn these intricate connections and provide valuable insights into the fundamental principles.[16,17] Additionally, LLMs are transitioning from passive assistants to active participants in the research process. The most advanced applications now integrate LLMs as a central "brain" into research workflows, where these agentic systems can plan multi-step procedures, interface with computational simulation tools, and even operate robotic platforms.[18–20]

LLMs are poised to reshape the materials science landscape. This potential, however, has largely been explored using closed-source, commercial models such as the GPT series from OpenAI. These models often benefit from diverse and extensive training data that capture a broader spectrum of knowledge, along with proprietary reinforcement learning from human feedback (RLHF) pipelines, which are key to stronger reasoning capabilities and better semantic alignment essential for handling complex tasks.[21] While these models are industry-leading, their closed-source nature presents many drawbacks: high costs for large-scale or high-throughput tasks, data privacy concerns, reduced reproducibility, and limited flexibility for model customisation.[22] In parallel, the

[a] School of Physics, Chemistry and Earth Sciences, Adelaide University, Adelaide 5005, Australia
[b] School of Computer and Mathematical Sciences, Adelaide University, Adelaide 5005, Australia
E-mail: j.evans@adelaide.edu.au

View Article Online
DOI: 10.1039/D5DD00499C

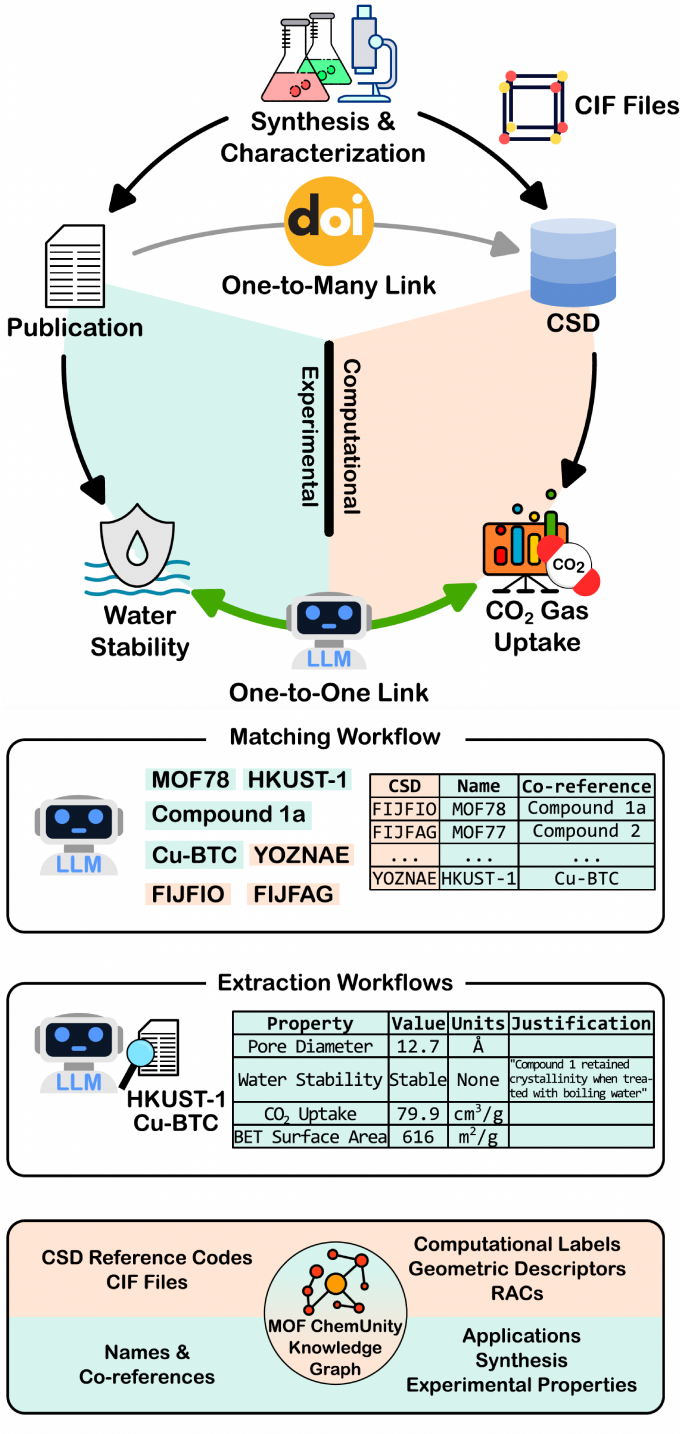

Fig. 1 The MOF-ChemUnity workflow. LLMs are used to link publications and CSD entries by extracting experimental properties and matching compound names across literature and structure files. This structured data populates the knowledge graph which combines synthesis, applications and more. Reproduced from ref. 6, licensed under CC BY-NC 4.0.

open-source LLM ecosystem has expanded significantly. While "open-source" in the context of LLMs typically refers to accessible weights rather than fully transparent training data or pre-training code, the ability to adapt these models through fine-tuning or reinforcement learning is generally sufficient for most research applications. The release of the Llama 3 family[23] by Meta in early 2024 marked the first time open-source models achieved true commercial-grade competitiveness with their closed-source counterparts. This milestone established a strong foundation for both research and industry applications. Subsequently, the Qwen[24] and GLM[25] series have made substantial progress toward matching and even surpassing proprietary models.

In this review, we outline the use of LLMs in materials science applications, with particular focus on MOFs. We also examine the capabilities of rapidly emerging open-source models across these diverse tasks.

2 Intelligent Data Extraction and Curation

The traditional approach to discovering new materials has largely been driven by trial and error, which requires extensive experimentation and validation.[26] With the efforts of thousands of dedicated researchers, a vast amount of valuable information has been accumulated over time. However, these data remain sporadic and scattered across different sources, therefore transforming this fragmented knowledge into a unified, standardised database can significantly facilitate the entire research process by enabling faster information retrieval and data-driven analysis.

In the work by Ghosh et al., an LLM-driven workflow was developed to autonomously draw out key thermoelectric properties (e.g., Seebeck coefficient, thermal conductivity) and associated structural properties (crystal class, space group, and doping strategy) from approximately 10,000 materials science articles.[27] They also benchmarked different Gemini and GPT models and found GPT-4.1 mini offered the best cost-performance balance. This effort resulted in the creation of the largest LLM-curated thermoelectric dataset which contains 27,822 temperature-resolved property records for a diverse class of materials. A key strength of this work lies in its explicit focus of tables and their associated captions as distinct, high-value data sources, rather than relying solely on unstructured text. Similarly, Li et al. developed an extraction workflow called "ReactionSeek" which is capable of directly interpreting reaction scheme images using a multimodal LLM (GLM-4V) and achieved an accuracy of 91.5% when tested on a set of 42 diverse images.[28] These multimodal expansions effectively broaden the scope of accessible data and enable a more comprehensive understanding of scientific information. It is worth noting that the models employed here are open-source and demonstrate impressive performance, highlighting the growing potential of community-driven models in specialised scientific applications.

Towards the development of MOFs, Pruyn et al. developed "MOF-ChemUnity" (Figure 1) that not only extracts key information such as material properties and synthesis procedures, but also links the various names used for these materials to their corresponding co-reference names and crystal structures.[6] This linkage bridges between textual synthesis, property knowledge and

View Article Online
DOI: 10.1039/D5DD00499C

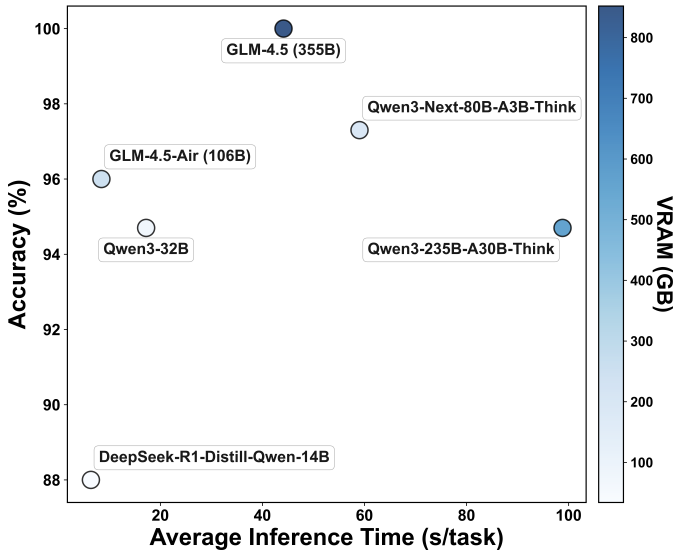

Fig. 2 Performance benchmark of open-source LLMs on the MOF-ChemUnity synthesis conditions extraction task. Model performance is plotted across three key dimensions: accuracy (%), average inference time (s/task), and estimated VRAM (GB) usage under bfloat16 precision (colour scale). The full content of each paper is processed in a single pass to extract the structured information. Note that Qwen3 models exhibit significantly higher average inference times compared to similarly sized models. This is attributed to their reasoning process prior to output generation. All models were evaluated with thinking mode enabled.

the atomic-level structural insights. Finally, the mined datasets form a knowledge graph that serves as a structured, scalable, and queryable foundation for materials discovery. However, while this approach captures the important details such as the overall synthesis duration, it only extracts static attributes and ignores the sequential order and relationships among synthesis actions. The work by Zhao et al. directly targets this gap by presenting a "sequence-aware" extraction, capturing the step-by-step experimental workflow as a directed graph, where each node represents an action (e.g., "mix", "heat", "filter"), and edges define the experimental sequence.[29] This workflow achieved high F1-scores for both entity (0.96) and relation (0.94) extraction. All of these studies highlight a significant shift from simple data extraction toward creating dynamic, AI-ready knowledge bases that enable sophisticated, data-driven discovery.

To demonstrate the performance of open-source models on these data extraction tasks, we reproduced the benchmark for six synthesis conditions provided by the MOF-ChemUnity code repository.[30] The models tested included the Qwen3 and GLM-4.5 series, featuring both dense and Mixture-of-Experts architectures with sizes ranging from 14B to 355B parameters. As shown in Figure 2, most models achieved accuracies exceeding 90%, with the largest model reaching 100%. This highlights the strong potential of open-source models which demonstrates their capability to effectively handle data-mining tasks. Notably, small models such as Qwen3-32B yielded an accuracy of 94.7%, suggesting that compact models can also handle the task effectively. This is significant as these smaller models require far fewer computational resources; Qwen3-32B, for instance, can be readily deployed on a standard Mac Studio with an M2 Ultra or M3 Max chip. The original study divided full-text literature into smaller chunks during pre-processing to identify relevant experimental paragraphs, which were then fed to GPT-4o for extraction. While this approach helps narrow the search space, it inevitably results in some loss of contextual detail, potentially omitting valuable features. In contrast, we processed entire papers using open-source models, which enabled the capture of additional information dispersed throughout the document that may have been missed by chunk-based approaches.

3 Predictive Power of LLMs

Beyond simply reciting information, Kang and coworkers developed a system called "L2M3" that not only extracts MOF synthesis conditions for database construction, but also provides a "recommender" tool by predicting synthesis conditions based on provided precursors by users.[31] As most LLMs are designed for general-purpose applications, they may not inherently excel at certain domain-specific tasks especially in materials science where there is complex terminology or unique patterns. Nevertheless, their extensive pre-trained knowledge provides a powerful foundation for transfer learning, enabling them to adapt efficiently even when only small, limited datasets are available. In this case, they fine-tuned GPT-3.5-turbo and GPT-4o using a dataset linking MOF precursors to their corresponding synthesis conditions. Both models achieved a moderate recommendation score of 82% compared to true experimental conditions. For further details on how the recommendation score is calculated, please refer to the Supporting Information.

Further, the work by Liu et al. demonstrated the predictive performance for MOF properties by providing compositions as well as high-level structural features like node connectivity and topology through rich, natural language descriptions.[32] The fine-tuned model achieved 94.8% accuracy in predicting hydrogen storage performance which is a substantial 46.7% improvement over models using only the precursor names.

To better encode comprehensive atomic-level details for LLMs, Song and colleagues proposed a new material representation format called "Material String" (see Figure 3), which is designed to be significantly shorter and more information-dense than standard crystal structure files like CIF or POSCAR.[33] This atomic-level representation encodes essential structure details such as space group, lattice parameters, and Wyckoff positions, allowing the complete mathematical reconstruction of a material's primitive cell in 3D. The fine-tuned model showed remarkable accuracy on the synthesisability test (98.6%). More importantly, it exhibited excellent generalisation, maintaining an average accuracy of 97.8% even when tested on complex experimental structures with up to 275 atoms and is far beyond the 40-atom limit of its training data. The model also achieved impressive performance on prediction of synthesis routes (91.0%). All of these results collectively underscore the ability of LLMs to learn and capture complex structural patterns or property features. They also highlight the importance of incorporating high-quality structural representations as input features to enable more reliable and physically meaningful predictions.

View Article Online
DOI: 10.1039/D5DD00499C

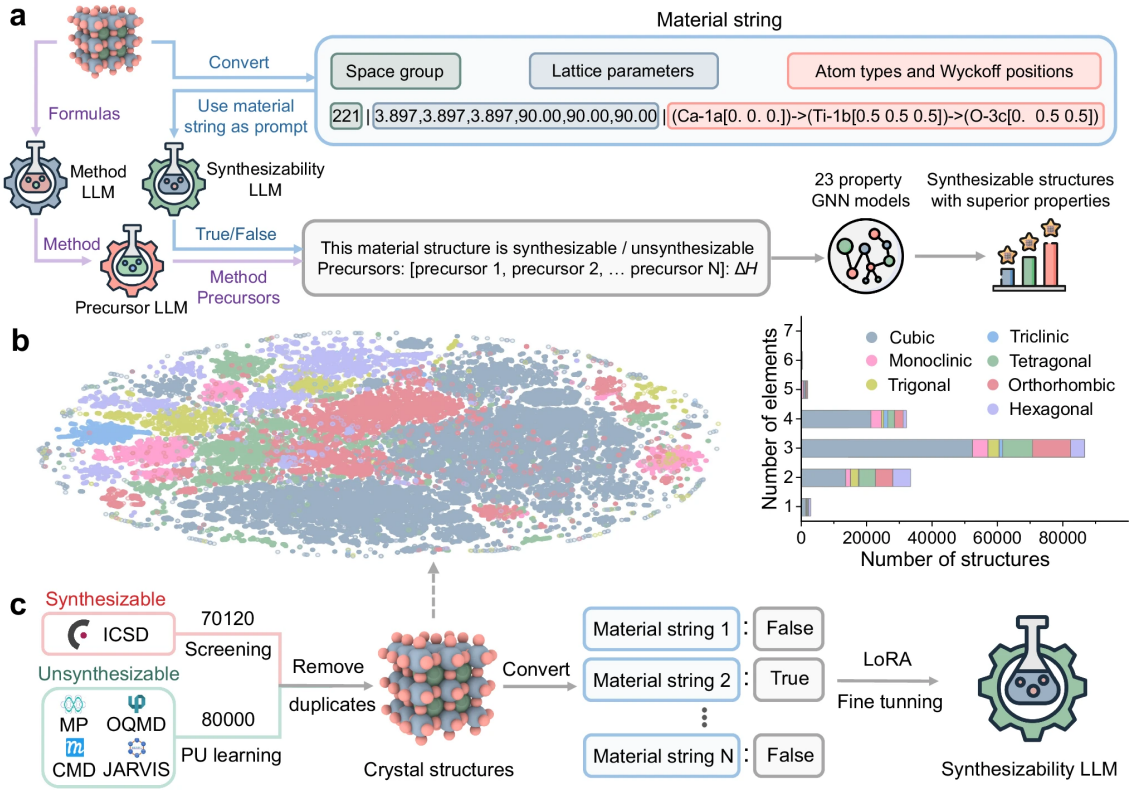

Fig. 3 Framework for predicting material synthesisability and synthesis routes (a). t-distributed Stochastic Neighbor Embedding (t-SNE) visualisation of the material structures used in the dataset that combined both experimental structures and non-synthesisable structures (b). Material string encoding structural data is used to train a "Synthesizability LLM" (c). Reproduced from ref. 33, licensed under CC BY-NC-ND 4.0.

To evaluate the capability of open-source models on prediction tasks, we fine-tuned three models of varying model sizes and architectures on the training dataset provided by L2M3.[34] As the official test set was unavailable, we further split the training dataset into 85% (4,990 samples) and 15% (1,039 samples) for training and evaluation, respectively. We employed Low-Rank Adaptation (LoRA) with a rank of 32 for efficient fine-tuning, which enabled the largest model tested, GLM-4.5-Air, to fit within four AMD Instinct MI250X Accelerators. When combined with 4-bit quantisation, the fine-tuning could be accommodated on only two MI250X with a minor loss in accuracy.[35] All models achieved a median score identical to that reported for GPT-4o (Figure 4).

Although these models achieved similar performance, closer inspection revealed that the dataset is highly imbalanced. (see Supporting Information) This imbalance can bias models toward the majority class, producing deceptively high accuracy while failing to learn meaningful patterns in the minority class. We also observed that one component of the similarity metric simply repeats the input precursors, yielding nearly 100% accuracy and further inflating the median score. These findings suggest that the models may memorise the most frequently occurring entries. Therefore, it is difficult to determine whether they genuinely learn the correlations between material formula strings and properties, or merely exploiting statistical frequency patterns in the dataset. This analysis also underscores the critical need for transparent reporting and data sharing. We observed that most studies in this domain lack sufficient detail for experimental reproduction, mak-

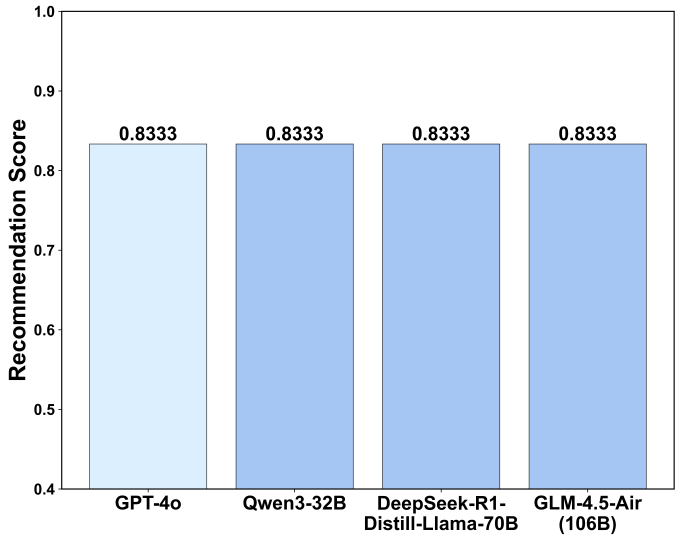

Fig. 4 Comparison of synthesis condition recommendation scores (median per-sample) for different fine-tuned open-source models. The result for GPT-4o (left) is the reported score from the original study.

View Article Online
DOI: 10.1039/D5DD00499C

ing it challenging to validate results or build upon existing work.

4 LLM Agents for Chemical Discovery

The development of LLMs is increasingly transitioning from single-model assistance towards sophisticated, multi-agent systems seemingly capable of performing complex human tasks.[36] This transformation is envisioned to revolutionise materials science, where most material research has long been constrained by laborious cycles of hypothesis formulation, pre-experimental research, and experimentation. By integrating LLMs into the research system, we can minimise unnecessary human intervention and accelerate the pace of discovery.

One of the most impactful roles of LLMs is supporting researchers in the exploration and refinement of their research ideas. As demonstrated by the SciAgents framework, LLMs can navigate a vast knowledge base to uncover previously unseen connections between disparate scientific concepts, which leads to the generation of novel hypotheses.[37] In addition, this framework can iteratively refine and elaborate on initial concepts. Different AI agents within the SciAgents framework can expand upon a hypothesis by adding quantitative details, suggesting specific modelling or simulation priorities, and providing comprehensive critiques that identify strengths, weaknesses, and areas for improvement. This feedback loop effectively mimics and accelerates the traditional scientific process of discussion and peer review, ensuring that the resulting ideas are not only innovative but also scientifically rigorous.

LLMs also hold strong potential to function as central coordinators or even cognitive engines, bridging researchers with complex computational tools in the pre-experimental stage. By leveraging their reasoning capability, LLM-based agents can understand and process queries in natural language, eliminating the need for rigid, formal syntax or other technical knowledge for using a computational tool. The ChatMOF system exemplifies this by orchestrating a sophisticated pipeline built on three core components: an agent, a toolkit, and an evaluator.[38] The toolkit contains a combination of the recognised databases such as QMOF and CoREMOF, and also a machine learning model that predicts material properties (e.g., hydrogen diffusivity), and a genetic algorithm tool for generating new combination of materials. When a user submits a query, the agent (powered by an LLM like GPT-4) functions as the "brain". It analyses the request, formulates a multi-step plan to solve the problem, and selects the appropriate instrument from its toolkit. The evaluator then assesses the output from the tool and synthesises it into a final, coherent answer for the user.

A related system, QUASAR, extends this coordinator paradigm into first-principles simulation workflows.[39] Rather than focusing on database querying and ML-driven property prediction, QUASAR is designed to deal directly with quantum and atomistic simulation. Upon receiving a task description such as calculating a band gap, the Strategist agent decomposes the request into a structured execution plan. The Operator agent then interprets each sub-task, performing detailed technical reasoning to construct validated input files for engines like Quantum ESPRESSO and LAMMPS. Crucially, the system functions not merely as an executor but as an execution-aware controller: it monitors runtime behaviour, diagnoses configuration issues (e.g., inappropriate energy cutoffs or convergence thresholds), applies corrective adjustments, and post-processes raw outputs into scientifically interpretable results.

Continuing the development of multi-agent systems, their scope has been further refined to focus on material discovery and optimisation. Zheng et al. developed a ChatGPT research group where seven distinct LLM-based assistants collaborate with a single human researcher, who only needs to specify the research objective through prompts.[40] As a result, this system successfully accelerated the finding of optimal synthesis conditions for MOFs and covalent-organic frameworks (COFs) by coupling AI agents with Bayesian optimisation, balancing the exploration and exploitation of a vast parameter space and reducing millions of potential conditions to a manageable number. In a related effort focused on de novo discovery, MOFGen was presented as a system of agentic AI dedicated to discovering novel, synthesisable MOFs.[41] This system employs a pipeline of specialised agents: LinkerGen, an LLM that proposes novel compositions; a diffusion model that generates 3D crystal structures; and other agents that perform quantum mechanical filtering and synthesisability analysis. This generative approach led to the successful synthesis of five "AI-dreamt" MOFs. These "vertical" applications demonstrate how such agentic structures can be specialised to solve deeper problems within a single domain. Moreover, all four systems highlight a core design principle that their strength lies not in the LLM alone, but in its ability to plan, delegate, and distribute sub-tasks across the entire research environment.

Moreover, coupling LLM agents with laboratory automation and robotic synthesis platforms can close the loop between computation and experiment. In the work by Boiko et al., an AI system named Coscientist was developed to autonomously design, plan, and perform complex experiments.[42] Driven by GPT-4, the system coordinates a suite of tools for internet and documentation search, code execution, and experimental automation (Figure 5). Crucially, Coscientist extended beyond in silico planning by directly controlling robotic liquid handlers for precise reagent transfers and managing heater-shaker modules to regulate reaction temperatures and mixing speeds. Its capabilities were demonstrated through the autonomous synthesis of organic compounds such as biphenyl and tolane. Similarly, Song et al. introduced ChemAgents, a multi-agent system powered by Llama-3.1-70B.[43] This system also features a central Task Manager that coordinates four highly specialised agents: Literature Reader, Experiment Designer, Computation Performer, and Robot Operator. Each agent is explicitly linked to a foundational resource, such as a literature database or an automated lab, which makes it a highly flexible framework that is able to execute tasks ranging from literature review to robotic operation. Demonstrated tasks include FTIR characterisation of azobenzene molecules and the synthesis and PXRD characterisation of six metal oxides (including ZrO_2 and ZnO). Notably, ChemAgents physically executes these experiments through Python-based Robot APIs, enabling coordinated control of a fully mobile robot and a benchtop arm across 20 automated stations for operations such as solid weighing, liquid

View Article Online
DOI: 10.1039/D5DD00499C

transfer, and photocatalytic performance evaluation. Both Coscientist and ChemAgents illustrate the view of leveraging LLM agents for general-purpose, "horizontal" platforms that can interface directly with experimental hardware through programmatic control.

The emergence of agentic systems marks a significant step toward the future of autonomous research. However, there are some crucial aspects to consider to ensure the robustness and reliability. At the computational level, the primary hurdle is the non-deterministic nature of LLMs. Unlike traditional hard-coded simulation workflows, an agentic system may generate slightly different reasoning paths or code structures for the same scientific query across multiple runs. In this sense, achieving a research objective is similar to solving a puzzle where the solution is not immediately apparent and multiple plausible paths must be considered. Therefore, these systems must possess the ability to self-correct by identifying when they have reached a dead end and recovering gracefully to maintain progress toward a valid solution. For example, in QUASAR's architecture design, Evaluator agent was used to evaluate the scientific rigorousness of task completion and fix potential problems spotted and form a feedback loop. Similarly, ChemAgents employs a hierarchical reflection and correction mechanism. Instead of relying on a single output, the system uses "critic" and "proofreader" agents to iteratively review, critique, and improve experimental procedures and robot codes against predefined expert rules.

At the physical level, the stakes of non-determinism are significantly higher. In systems like Coscientist, a minor variation in an LLM's interpreted instruction could result in an irreversible physical error, such as an incorrect reagent transfer or a mechanical collision. Because physical lab environments lack an "undo" function, orchestration must incorporate multimodal feedback loops. This means the Robot Operator agent cannot rely solely on the success of a Python API call; it must utilise computer vision and sensor data to verify that a vial is correctly seated or that a liquid transfer actually occurred. Moreover, the system must be execution-aware, capable of responding to asynchronous interruptions such as depleted reagents or hardware malfunctions without causing the entire multi-agent workflow to fail.

Finally, it is worth noting that the LLM used in ChemAgents, Llama-3.1-70B, is an open-source model. This represents a meaningful design choice, as most existing LLM-based agents in this domain continue to rely on closed-source systems, which involve certain trade-offs. Commercial models can cause significant financial cost and accessing them through APIs can expose systems to instability and security risks. APIs are prone to outages, disruptions, and unexpected updates that may compromise reliability, and their dependence on internet connectivity makes it difficult to ensure data privacy and compliance with confidentiality protocols, which is a major concern for sensitive experiments or proprietary research data. In contrast, the rapid advancement of open-source LLMs is beginning to change this landscape. Models such as GLM-4.5 and Qwen3-235B-Thinking-2507 already demonstrate agentic and reasoning performance comparable to that of leading closed-source counterparts.[25] Therefore, selecting an appropriate system should strike a balance between performance, resource availability, and operational independence, and open-source development may ultimately pave the way for sustainable, locally deployable autonomous research systems.

Limitations

While this review demonstrates the transformative potential of LLMs across the materials discovery pipeline, it is important to recognise that this domain-specific success is significantly bolstered by the existence of mature and community-wide infrastructures that underpin reliable model development. Systematic nomenclature, well-established characterisation protocols, and curated databases of MOFs collectively provide the high-quality, structured ground truth essential for training and benchmarking LLMs, making tasks such as automated extraction and structure–property matching considerably more tractable than in many other classes of materials. In contrast, domains such as complex alloys and polymers frequently lack unified nomenclature and consistent reporting standards. Consequently, directly transferring the LLM methodologies described here to new material domains may be challenging.

Despite the rapid advancement of these agentic frameworks, coupling autonomous AI with physical laboratory hardware presents significant safety and operational risks. At the core of these challenges is the lack of deterministic safety bounds; unlike traditional industrial robots that operate on fixed, pre-verified paths, LLM-based agents generate "reasoning-driven" instructions that can be unpredictable or non-reproducible. This stochastic execution introduces the risk of catastrophic hardware collisions or hazardous chemical spills if the agent misinterprets a sensor reading or fails to account for the physical constraints of a benchtop arm. In addition, current robotic platforms remain fundamentally limited in their ability to handle unstructured anomalies, such as a cracked vial or a slightly misaligned microplate. These seemingly minor deviations can propagate uncertainty throughout the workflow, often forcing execution into a single-direction sequence with limited capacity for adaptive recovery.

Additionally, we would like to highlight some downsides of open-source models. Unlike pay-as-you-go proprietary models, local deployment requires substantial upfront investment in high-performance hardware to meet the VRAM demands of larger models. For instance, the largest model tested in this review, GLM-4.5 (355B total parameters), requires approximately 823 GB of VRAM for inference. Beyond hardware costs, the operational overhead, including the specialised expertise needed for model inference and fine-tuning, can pose a substantial barrier. These total cost of ownership considerations may outweigh the savings from avoiding API fees, particularly for researchers who do not require high-throughput processing or the stringent data privacy guarantees that local hosting provides.

Conclusions and Outlook

In conclusion, the integration of LLMs across data extraction, predictive modelling, and agentic system demonstrates an emerging capability to interpret complex chemistry text, learn structure–property relationships, and orchestrate research workflows. The studies reviewed here together with our benchmarking result

View Article Online
DOI: 10.1039/D5DD00499C

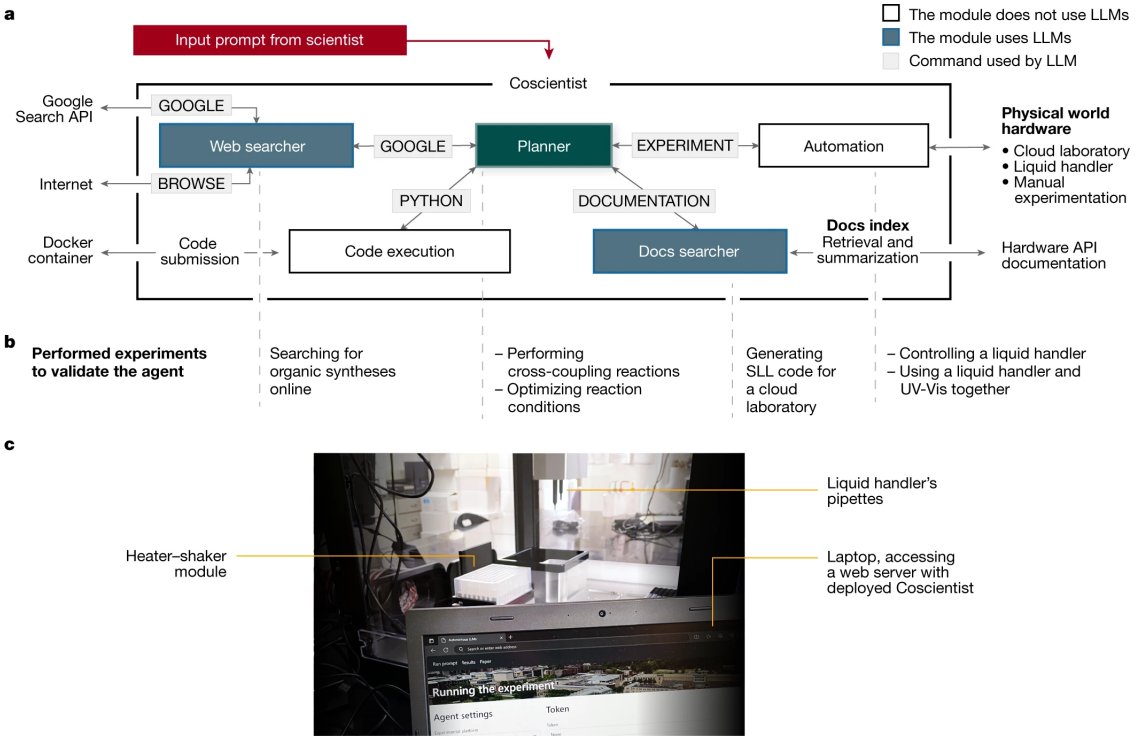

Fig. 5 A comprehensive system featuring a central LLM-based "Planner" that orchestrates and manages the entire research workflow. Reproduced from ref. 42, licensed under CC BY 4.0.

confirm that open-source models can achieve comparable performance to closed-source systems, on the tasks evaluated here, while offering superior transparency, flexibility, and reproducibility. Collectively, these advances mark a transition from isolated, task-specific LLM applications toward unified, AI-driven research ecosystems that accelerate the exploration of complex scientific knowledge, minimise repetitive manual effort, and open new frontiers for discovery.

However, evaluating the reliability of highly autonomous systems remains a challenge. While recent studies have attempted to benchmark the agentic abilities of LLMs such as tool calling,[44] there is still not yet a comprehensive framework for assessing performance beyond one-step reasoning accuracy. Complex agentic systems require the integration of planning, execution, and adaptive decision-making, which current benchmarks do not capture. Moreover, existing metrics predominantly measure procedural correctness rather than the holistic resilience of reasoning when confronted with uncertainty or failure. Developing new benchmarks is therefore crucial to clarify the true competency of models acting as "researchers". Such frameworks would not only guide the need for domain-specific fine-tuning, but also be essential for building trustworthy autonomous systems that can be confidently adopted in real-world scientific research.

Conflicts of interest

There are no conflicts to declare.

Data availability

Code and data related to benchmarking and fine-tuning tools are available on Zenodo: 10.5281/zenodo.17548056.

Acknowledgements

J.D.E. is the recipient of an Australian Research Council Discovery Early Career Award (project number DE220100163) funded by the Australian Government. The Phoenix HPC service at the Adelaide University is thanked for providing high-performance computing resources. This research was supported by the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS), with access to computational resources provided by Pawsey Supercomputing Research Centre through the National Computational Merit Allocation Scheme. We thank Dr Fabien Voisin (Phoenix HPC, Adelaide University) for his assistance managing our resource needs.

Notes and references

1 M. R. AI4Science and M. A. Quantum, *The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4*, 2023, http://arxiv.org/abs/2311.07361, arXiv:2311.07361.

2 Y. Zimmermann, A. Bazgir, A. Al-Feghali, M. Ansari, J. Bocarsly, L. C. Brinson, Y. Chiang, D. Circi, M.-H. Chiu, N. Daelman, M. L. Evans, A. S. Gangan, J. George, H. Harb, G. Khalighinejad, S. T. Khan, S. Klawohn, M. Lederbauer, S. Mahjoubi, B. Mohr, S. M. Moosavi, A. Naik, A. B. Ozhan, D. Plessers, A. Roy, F. Schöppach, P. Schwaller, C. Terboven, K. Ueltzen, Y. Wu, S. Zhu, J. Janssen, C. Li, I. Foster and

View Article Online
DOI: 10.1039/D5DD00499C

B. Blaiszik, *34 Examples of LLM Applications in Materials Science and Chemistry: Towards Automation, Assistants, Agents, and Accelerated Scientific Discovery*, 2025, http://arxiv.org/abs/2505.03049, arXiv:2505.03049.

3 D. Grigorov, *Introduction to Python and Large Language Models: A Guide to Language Models*, Apress, Berkeley, CA, 2024, pp. 59–100.

4 E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn and G. Kasneci, *Learning and Individual Differences*, 2023, **103**, 102274.

5 A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan and D. S. W. Ting, *Nature Medicine*, 2023, **29**, 1930–1940.

6 T. M. Pruyn, A. Aswad, S. T. Khan, R. Black and S. M. Moosavi, *MOF-ChemUnity: Unifying metal-organic framework data using large language models*, 2025, https://chemrxiv.org/engage/chemrxiv/article-details/6838df8bc1cb1ecda036f363.

7 S. Kitagawa, *Chemical Society Reviews*, 2014, **43**, 5415–5418.

8 H. Furukawa, K. E. Cordova, M. O'Keeffe and O. M. Yaghi, *Science*, 2013, **341**, 1230444.

9 *Nobel Prize in Chemistry 2025*, https://www.nobelprize.org/prizes/chemistry/2025/press-release/.

10 J. Xing, Y. Liu, G. Mathew, Q. He, J. Aghassi-Hagmann, S. Schweidler and B. Breitung, *Advanced Science*, 2025, **12**, 2411175.

11 D. Li, A. Yadav, H. Zhou, K. Roy, P. Thanasekaran and C. Lee, *Global Challenges*, 2024, **8**, 2300244.

12 K. Stracke and J. D. Evans, *Communications Chemistry*, 2024, **7**, 1–4.

13 S. Bae, M. Jeon and H. Ri Moon, *Chemical Communications*, 2025, **61**, 11083–11094.

14 M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez and K. M. Jablonka, *Chemical Society Reviews*, 2025, **54**, 1125–1150.

15 L. Zhang, J. Du, Z. Xie, L. Chen, W. Li, W. Geng, Y. Zhou, X. Ou, C. Gong, Y. Gao, S. He, C. Yan, C. Zhao, Y. Jiao, S.-Y. Yang, B. Huang, J. W. Y. Lam, J. Qian, J. Jiang, B. Z. Tang and H. Deng, *Nature Chemistry*, 2025, **17**, 1645–1654.

16 *"M"²LLM: Multi-view Molecular Representation Learning with Large Language Models*, https://arxiv.org/html/2508.08657v1.

17 Z. Wang, K. Zhang, Z. Zhao, Y. Wen, A. Pandey, H. Liu and K. Ding, *A Survey of Large Language Models for Text-Guided Molecular Discovery: from Molecule Generation to Optimization*, 2025, http://arxiv.org/abs/2505.16094, arXiv:2505.16094.

18 L. Mitchener, A. Yiu, B. Chang, M. Bourdenx, T. Nadolski, A. Sulovari, E. C. Landsness, D. L. Barabasi, S. Narayanan, N. Evans, S. Reddy, M. Foiani, A. Kamal, L. P. Shriver, F. Cao, A. T. Wassie, J. M. Laurent, E. Melville-Green, M. Caldas, A. Bou, K. F. Roberts, S. Zagorac, T. C. Orr, M. E. Orr, K. J. Zwezdaryk, A. E. Ghareeb, L. McCoy, B. Gomes, E. A. Ashley, K. E. Duff, T. Buonassisi, T. Rainforth, R. J. Bateman, M. Skarlinski, S. G. Rodriques, M. M. Hinks and A. D. White, *Kosmos: An AI Scientist for Autonomous Discovery*, 2025, http://arxiv.org/abs/2511.02824, arXiv:2511.02824.

19 K. Darvish, M. Skreta, Y. Zhao, N. Yoshikawa, S. Som, M. Bogdanovic, Y. Cao, H. Hao, H. Xu, A. Aspuru-Guzik, A. Garg and F. Shkurti, *Matter*, 2025, **8**, 101897.

20 O. A. Mendible-Barreto, M. Díaz-Maldonado, F. J. C. Esteva, J. Emmanuel Torres, U. M. Córdova-Figueroa and Y. J. Colón, *Molecular Systems Design & Engineering*, 2025, **10**, 585–598.

21 Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann and J. Kaplan, *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*, 2022, https://arxiv.org/abs/2204.05862.

22 S. Kumar, M. A. Lones, M. Maarek and H. Zantout, *Navigating Pitfalls: Evaluating LLMs in Machine Learning Programming Education*, 2025, http://arxiv.org/abs/2505.18220, arXiv:2505.18220.

23 A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. v. d. Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. v. d. Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. d. Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Gird-

View Article Online
DOI: 10.1039/D5DD00499C

har, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao and Z. Ma, *The Llama 3 Herd of Models*, 2024, http://arxiv.org/abs/2407.21783, arXiv:2407.21783.

24 A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou and Z. Qiu, *Qwen3 Technical Report*, 2025, http://arxiv.org/abs/2505.09388, arXiv:2505.09388.

25 G.-. . Team, A. Zeng, X. Lv, Q. Zheng, Z. Hou, B. Chen, C. Xie, C. Wang, D. Yin, H. Zeng, J. Zhang, K. Wang, L. Zhong, M. Liu, R. Lu, S. Cao, X. Zhang, X. Huang, Y. Wei, Y. Cheng, Y. An, Y. Niu, Y. Wen, Y. Bai, Z. Du, Z. Wang, Z. Zhu, B. Zhang, B. Wen, B. Wu, B. Xu, C. Huang, C. Zhao, C. Cai, C. Yu, C. Li, C. Ge, C. Huang, C. Zhang, C. Xu, C. Zhu, C. Li, C. Yin, D. Lin, D. Yang, D. Jiang, D. Ai, E. Zhu, F. Wang, G. Pan, G. Wang, H. Sun, H. Li, H. Li, H. Hu, H. Zhang, H. Peng, H. Tai, H. Zhang, H. Wang, H. Yang, H. Liu, H. Zhao, H. Liu, H. Yan, H. Liu, H. Chen, J. Li, J. Zhao, J. Ren, J. Jiao, J. Zhao, J. Yan, J. Wang, J. Gui, J. Zhao, J. Liu, J. Li, J. Li, J. Lu, J. Wang, J. Yuan, J. Li, J. Du, J. Du, J. Liu, J. Zhi, J. Gao, K. Wang, L. Yang, L. Xu, L. Fan, L. Wu, L. Ding, L. Wang, M. Zhang, M. Li, M. Xu, M. Zhao, M. Zhai, P. Du, Q. Dong, S. Lei, S. Tu, S. Yang, S. Lu, S. Li, S. Li, S. Yang, S. Yi, T. Yu, W. Tian, W. Wang, W. Yu, W. L. Tam, W. Liang, W. Liu, X. Wang, X. Jia, X. Gu, X. Ling, X. Wang, X. Fan, X. Pan, X. Zhang, X. Zhang, X. Fu, X. Zhang, Y. Xu, Y. Wu, Y. Lu, Y. Wang, Y. Zhou, Y. Pan, Y. Zhang, Y. Wang, Y. Li, Y. Su, Y. Geng, Y. Zhu, Y. Yang, Y. Li, Y. Wu, Y. Li, Y. Liu, Y. Wang, Y. Li, Y. Zhang, Z. Liu, Z. Yang, Z. Zhou, Z. Qiao, Z. Feng, Z. Liu, Z. Zhang, Z. Wang, Z. Yao, Z. Wang, Z. Liu, Z. Chai, Z. Li, Z. Zhao, W. Chen, J. Zhai, B. Xu, M. Huang, H. Wang, J. Li, Y. Dong and J. Tang, *GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models*, 2025, http://arxiv.org/abs/2508.06471, arXiv:2508.06471.

26 A. W. Thornton, C. M. Simon, J. Kim, O. Kwon, K. S. Deeg, K. Konstas, S. J. Pas, M. R. Hill, D. A. Winkler, M. Haranczyk

View Article Online
DOI: 10.1039/D5DD00499C

and B. Smit, *Chemistry of Materials*, 2017, **29**, 2844–2854.

27 S. Ghosh and A. Tewari, *Automated Extraction of Material Properties using LLM-based AI Agents*, 2025, http://arxiv.org/abs/2510.01235, arXiv:2510.01235.

28 J. Li, M. Li, Q. Yang and S. Luo, *ReactionSeek: LLM-Powered Literature Data Mining and Knowledge Discovery in Organic Synthesis*, 2025, https://chemrxiv.org/engage/chemrxiv/article-details/689328e223be8e43d6f494d3.

29 X. Zhao, J. F. F. Rojas, J. Furst, K. Ardila, K. Langlois, Y. An, X. Hu, F. Uribe-Romo, D. Gomez-Gualdron and J. Greenberg, *Expert-Guided LLM Approach for Sequence-Aware Extraction of MOF Synthesis*, 2025, https://chemrxiv.org/engage/chemrxiv/article-details/689c179c728bf9025e39cb58.

30 *AI4ChemS/MOF_ChemUnity*, 2025, https://github.com/AI4ChemS/MOF_ChemUnity, original-date: 2024-11-14T22:44:44Z.

31 Y. Kang, W. Lee, T. Bae, S. Han, H. Jang and J. Kim, *Journal of the American Chemical Society*, 2025, **147**, 3943–3958.

32 Z. Liu, Y. Su, H. Wang, T. Ban, L. Wang, S. Lu, Z. Xi, W. Li, Y. Guo, C. Wang, H. Gao and G. Wang, *Leveraging Post-Pretrained LLMs for Inverse Engineering High-Capacity Hydrogen-Storage Metal-Organic Frameworks: From Virtual Structures to Synthesized Materials*, 2025, https://chemrxiv.org/engage/chemrxiv/article-details/682f187d1a8f9bdab5390703.

33 Z. Song, S. Lu, M. Ju, Q. Zhou and J. Wang, *Nature Communications*, 2025, **16**, 6530.

34 *Taeun8991/L2M3*, 2025, https://github.com/Taeun8991/L2M3, original-date: 2025-02-18T08:17:15Z.

35 T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer, *QLoRA: Efficient Finetuning of Quantized LLMs*, 2023, http://arxiv.org/abs/2305.14314, arXiv:2305.14314.

36 L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei and J. Wen, *Frontiers of Computer Science*, 2024, **18**, 186345.

37 A. Ghafarollahi and M. J. Buehler, *Advanced Materials*, 2025, **37**, 2413523.

38 Y. Kang and J. Kim, *Nature Communications*, 2024, **15**, 4705.

39 F. Yang and J. D. Evans, *QUASAR: A Universal Autonomous System for Atomistic Simulation and a Benchmark of Its Capabilities*, 2026, https://arxiv.org/abs/2602.00185.

40 Z. Zheng, O. Zhang, H. L. Nguyen, N. Rampal, A. H. Alawadhi, Z. Rong, T. Head-Gordon, C. Borgs, J. T. Chayes and O. M. Yaghi, *ACS Central Science*, 2023, **9**, 2161–2170.

41 T. J. Inizan, S. Yang, A. Kaplan, Y.-h. Lin, J. Yin, S. Mirzaei, M. Abdelgaid, A. H. Alawadhi, K. Cho, Z. Zheng, E. D. Cubuk, C. Borgs, J. T. Chayes, K. A. Persson and O. M. Yaghi, *System of Agentic AI for the Discovery of Metal-Organic Frameworks*, 2025, https://arxiv.org/abs/2504.14110.

42 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, *Nature*, 2023, **624**, 570–578.

43 T. Song, M. Luo, X. Zhang, L. Chen, Y. Huang, J. Cao, Q. Zhu, D. Liu, B. Zhang, G. Zou, G. Zhang, F. Zhang, W. Shang, Y. Fu, J. Jiang and Y. Luo, *Journal of the American Chemical Society*, 2025, **147**, 12534–12545.

44 *ChemAgent: Enhancing LLMs for Chemistry and Materials Science through Tree-Search Based Tool Learning*, https://arxiv.org/html/2506.07551v1.

View Article Online
DOI: 10.1039/D5DD00499C

Data for this article, including code and data related to benchmarking and fine-tuning tools are available at Zenodo at https://doi.org/10.5281/zenodo.17548056.