Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 869

MOFReasoner: think like a scientist—a reasoning large language model *via* knowledge distillation

Xuefeng Bai,[ab] Zhiling Zheng,[c] Xin Zhang, *[ab] Hao-Tian Wang,[ab] Rui Yang[ab] and Jian-Rong Li *[ab]

Large Language Models (LLMs) have the potential to transform chemical research. Nevertheless, their general-purpose design constrains scientific understanding and reasoning within specialized fields like chemistry. In this study, we introduce MOFReasoner, a domain model designed to enhance scientific reasoning, using Metal–Organic Framework (MOF) adsorption as a case study. By employing knowledge distillation from teacher models and Chain-of-Thought (CoT) reasoning extracted from a corpus of over 8242 research articles and 500 reviews, we developed a domain-specific chemical reasoning dataset. Using domain-specific chemical reasoning datasets, general chemistry datasets, and general reasoning datasets, the LLMs were fine-tuned. The model's performance was evaluated across four tasks: experimental studies, chemical mechanisms, application scenarios, and industrialization challenges. MOFReasoner outperformed existing general-purpose models, such as GPT-4.5 and DeepSeek-R1. Furthermore, the model achieves prediction accuracy comparable to DFT, enabling material recommendations. This work underscores the potential of integrating domain-specific knowledge, CoT reasoning, and knowledge distillation in creating LLMs that support scientific inquiry and decision-making within the discipline of chemistry.

Received 25th September 2025
Accepted 8th January 2026

DOI: 10.1039/d5dd00429b

rsc.li/digitaldiscovery

Introduction

AI has revolutionized chemistry research by driving advancements in molecular design, reaction prediction, and materials discovery.[1–3] In recent years, large language models (LLMs) have shown remarkable potential in knowledge extraction, complex reasoning, and automated data analysis, positioning them as powerful tools for accelerating scientific innovation.[4–9] Due to their strong natural language understanding capabilities, LLMs are used for processing vast amounts of literature, generating insightful hypotheses, and assisting in experiment planning, among other tasks.[10–13] To enhance their applications in chemistry, various optimization strategies have been explored by researchers. Prompt engineering has proven effective in extracting scientific data,[14,15] while multi-agent collaboration frameworks distribute tasks and improve decision-making for complex chemical problems.[16–18] In addition, LLMs can further enhance their mastery of chemical knowledge through techniques like fine-tuning and RAG, enabling them to perform more advanced tasks such as reaction prediction and Q&A.[19–22]

Despite the enhancement of expertise of LLMs through fine-tuning and RAG, their ability to tackle complex problems, especially those requiring multi-step chemical reasoning, remains insufficient. This limitation constrains their applications in areas such as materials design and chemical reasoning Q&A.[23,24] Therefore, LLMs need to possess scientific reasoning abilities akin to those of scientists. Once they acquire such thinking skills, they can derive accurate conclusions through more rigorous logical deduction. These accurate conclusions will further enhance the performance of LLMs in areas such as materials design, performance prediction, and multi-objective optimization, enabling them to achieve the experimental paradigm of autonomous AI research in the future.

Chain-of-Thought (CoT) reasoning enhances LLMs by enabling structured, step-by-step logical inference, allowing them to tackle answer chemical questions with improved accuracy and interpretability.[25,26] This is particularly beneficial for tasks requiring multi-step reasoning, such as reaction prediction and experimental design, where breaking down intricate processes leads to more reliable and scientifically sound conclusions. However, the reasoning patterns of LLMs trained on general knowledge still differ significantly from those used in scientific research, as scientific reasoning often involves a process of making numerous hypotheses followed by verification. Therefore, it is crucial to further train LLMs in specialized domain knowledge and scientific reasoning. Building such a domain-specific reasoning model requires fine-

[a]*Department of Chemical Engineering, College of Materials Science & Engineering, Beijing University of Technology, Beijing 100124, P. R. China. E-mail: jrli@bjut.edu.cn; zhang.xin@bjut.edu.cn*

[b]*State Key Laboratory of Materials Low-Carbon Recycling, Beijing University of Technology, Beijing 100124, China*

[c]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02142, USA*

tuning based on domain-specific CoT data. Knowledge distillation provides an efficient way to transfer domain expertise from larger parameter models or structured knowledge sources to small parameter models.[27–29] By integrating CoT reasoning with knowledge distillation, LLMs can acquire not only more domain-specific knowledge but also domain-specific thinking. This approach enables them to achieve structured, step-by-step reasoning, leading to more reliable inference and more efficient knowledge utilization. Recently, ChemMatch and the Scholar-ChemQA dataset[30] have highlighted the potential of lightweight, domain-specific models for chemical QA; in contrast, our work emphasizes equipping LLMs with multi-step scientific reasoning through literature-derived CoT data and knowledge distillation. It should be noted that while CoT reasoning improves interpretability and task performance, recent studies suggest that such outputs may reflect structured pattern generation rather than genuine scientific understanding. As a highly designable class of three-dimensional materials, MOFs, which are widely used in adsorption separation,[31] catalysis,[32] and other fields,[33,34] greatly benefit from AI assistance due to their vast potential for synthesizing a diverse range of materials. Herein, taking the field of MOF adsorption as an example, we extracted domain-specific reasoning pathways from scientific papers and refined them with the aid of large-parameter language models to construct a domain-specific CoT database and trained MOFReasoner (as shown in Fig. 1). The model can be found at https://huggingface.co/baixuefeng/ChemReasoner-7B. Specifically, through a hard-label knowledge distillation approach inspired by recent large-model compression studies,[35] we conducted high-throughput analysis with large-parameter, long-text teacher LLMs, guiding them to extract and structure chain-of-thought reasoning from the literature, and organized the results into a domain-specific reasoning dataset. Additionally, we used these teacher models to transform existing chemical datasets into chemical reasoning datasets. By integrating these datasets with general CoT datasets, we constructed a reasoning model for chemistry named MOFReasoner. MOFReasoner, with its enhanced structured reasoning capabilities and effective integration of chemical knowledge, significantly outperforms ChatGPT and DeepSeek on a dataset consisting of four types of tasks in Q&A testing. Moreover, MOFReasoner can be further coupled with existing knowledge bases and knowledge graphs,[36–38] and through its robust reasoning capability, it recommends materials that are consistent with DFT calculation results.

Results and discussion

In the initial phase of our study, we amassed a collection of over 8200 research papers and more than 500 review articles, which were organized into a text format using Python code. To characterize the dataset, we further summarized the distribution of journals, publishers, and dominant scientific keywords, as shown in Fig. S1–S7 and Tables S1–S3. These analyses confirm that the corpus covers diverse publication sources and broadly representative MOF-related topics, without exhibiting over-concentration in any single journal, publisher, or material family. This balanced distribution indicates that the data collection process did not introduce systematic bias and provides a reliable foundation for downstream model training

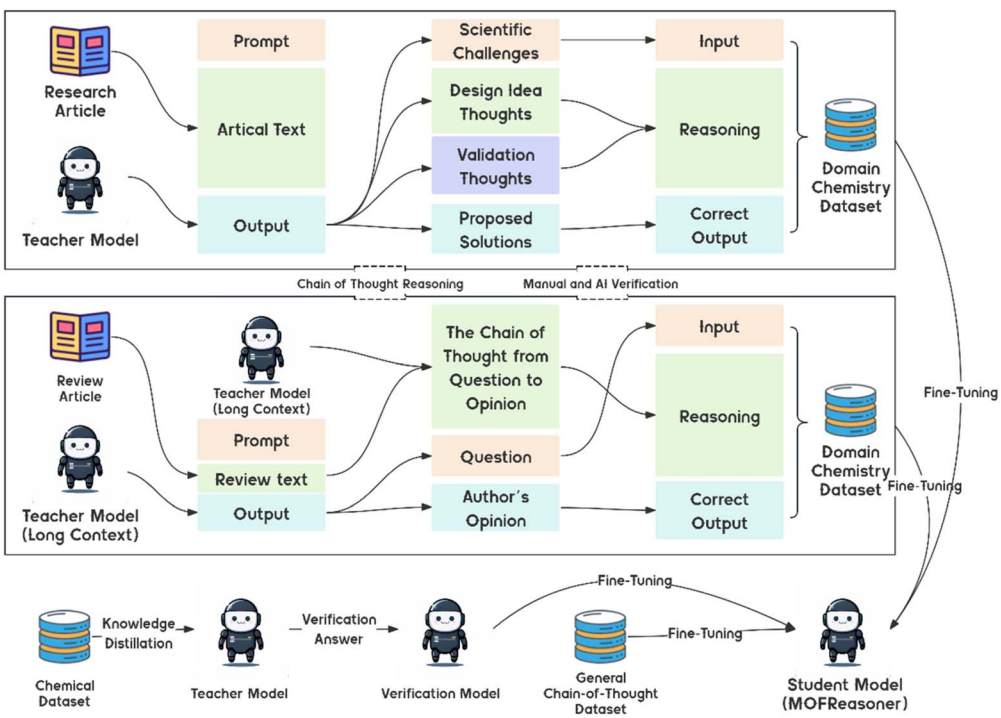

Fig. 1 Workflow for constructing the domain-specific reasoning dataset from domain-specific papers, which is then combined with chemical reasoning datasets and general CoT datasets to train the domain-specific reasoning LLM.

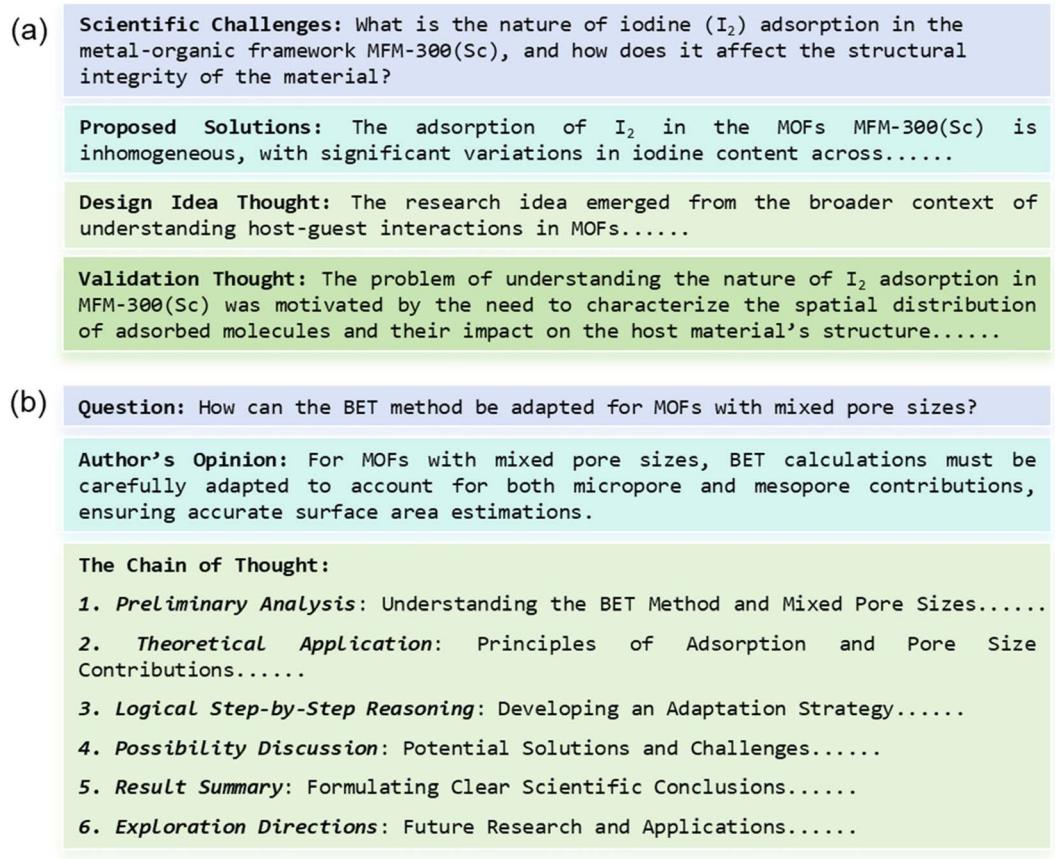

Fig. 2 Knowledge distillation from teacher models on the article: (a) distillation of research article knowledge using DeepSeek-V3; (b) distillation of review article knowledge using Qwen-Turbo.

and evaluation. We note that such literature-based extraction may introduce bias, since prompts can influence the reasoning obtained, and scientific papers typically emphasize successful cases while underreporting failures. For research articles, we employed prompt engineering to leverage the DeepSeek-V3(671B) model (as shown in Table S4 for model versions and access dates), allowing us to extract scientific challenges, proposed solutions, design idea thoughts, and validation thoughts from the texts. As illustrated in Fig. 2a, this LLM effectively translates the ideas and validation logic from the papers into a CoT, resembling the problem-solving approach of a scientist. To enhance transparency, Section S2 (Fig. S8–S11) now provides representative examples of the mined CoT data and their transformation process. Although these CoT sequences inevitably deviate from the full complexity of real scientific thinking, they nonetheless provide a reasonable and practical representation of how reasoning is articulated in the published literature.

In the case of review articles, which consist of experts' summaries of existing research and contain profound scientific insights, we employed the long-text model Qwen-Turbo to distill and summarize the scientific perspectives. We then transformed these insights into question–answer pairs (as shown in Fig. S12–S16). Subsequently, each question-and-answer pair is matched with the original comment content as context, and then presented to the LLM. The LLMs provide a detailed CoT process from multiple dimensions (as shown in Fig. 2b and S17–S21). The current pipeline processes only textual information, and visual data such as adsorption isotherms, PXRD patterns, and microscopy images were not directly included, which may lead to partial underrepresentation of information conveyed exclusively through figures.

In addition to constructing a domain-specific dataset, we also utilized a general chemistry dataset and a general reasoning dataset to enhance the model's chemical knowledge and reasoning capabilities. The camel-ai chemistry dataset,[39] which includes 20 000 chemistry questions across 25 topics, serves as an excellent foundation dataset for general chemistry knowledge. However, since this dataset includes only questions and answers without detailed problem-solving procedures, we applied DeepSeek-R1(671B), an inference LLM, to better equip the LLMs with comprehensive chemistry knowledge and CoT. As shown in Fig. S22, the LLMs delivered exhaustive reasoning processes through logical inference. The CoT dataset utilized STILL,[40] a slow-thinking reasoning dataset, which did not require additional processing, as it is originally presented in the CoT format within the SFT structure (Fig. S23–S25).

Subsequently, we conducted a systematic validation of the dataset to ensure the quality and reliability of the training data. For datasets with clearly defined standard answers, we

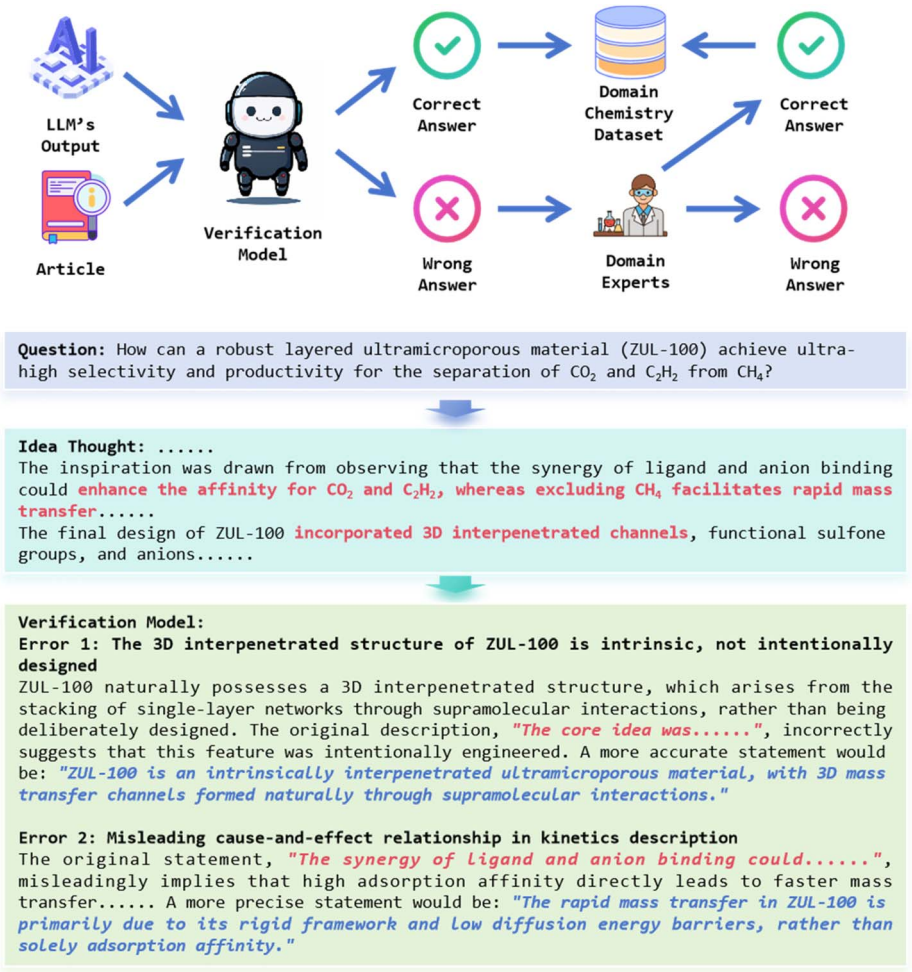

Fig. 3 Construction of domain-specific datasets *via* validation models and manual evaluation of large model responses.

employed DeepSeek-V3 for answer verification to assess the reasoning accuracy and response quality of the LLMs (Fig. S26). Specifically, we compared the standard answers provided in the dataset with the responses generated by the LLMs, evaluating their consistency. If the generated answers matched the standard ones, we considered the reasoning process to be reliable. For literature-extracted datasets without predefined standard answers, we adopted a hybrid approach combining LLM-based filtering with human verification. Initially, LLMs were used to screen the data, after which the original text from the research papers and the reasoning process of LLMs were simultaneously provided to a validation model. This new model then assessed the logical soundness of the reasoning. In cases where the responses were ambiguous or controversial, domain experts were consulted for further judgment. As shown in Fig. 3, the validation model, guided by prompts and review content, effectively identified errors in long-text model responses and provided original text excerpts as supporting evidence.

Finally, as shown in Fig. S27, we performed knowledge distillation by training the student model to emulate the reasoning behaviors of the teacher model (DeepSeek-R1) on general chemistry datasets. This procedure achieved only a ~50% success rate, reflecting the difficulty of answering challenging out-of-domain chemistry questions where relevant knowledge is often absent from the pretraining corpus. In contrast, when distillation was conducted using research papers and reviews, the model could rely on contextual information provided in the documents, leading to an accuracy exceeding 90%. This demonstrates the importance of context-augmented reasoning: rather than recalling memorized facts, the model synthesizes information from scientific literature into structured reasoning traces. The final data distribution is shown in Fig. S28, with a total of 35.8 K data points utilized for LLM training.

In this work, we fine-tuned a small-parameter reasoning LLM, DeepSeek-R1-Distill-Qwen-7B, using the LLaMA-Factory framework. Specifically, we employed supervised fine-tuning to adapt the model to domain-specific tasks and utilized low-rank adaptation to enhance efficiency by reducing trainable parameters while maintaining model performance. This approach enabled efficient adaptation of the LLM with reduced computational cost and memory footprint. As shown in Fig. S29, after a single epoch of training comprising 717 steps, the training loss was reduced to 0.8036.

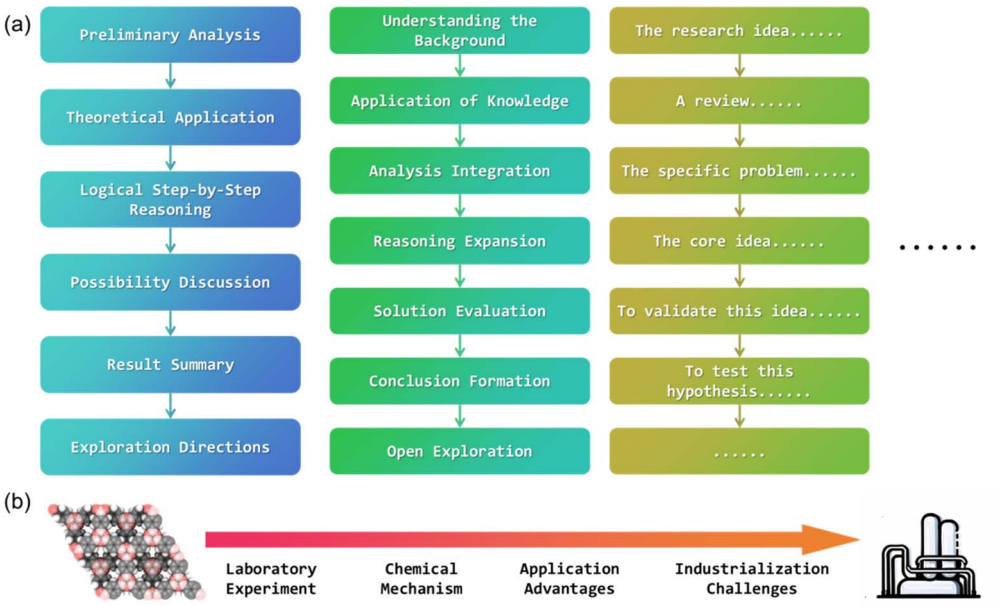

Fig. 4 (a) Examples of MOFReasoner's reasoning process; (b) four types of tasks for large language model evaluation.

When presented with scientific questions, the trained MOFReasoner is capable of reasoning logically like a scientist and providing well-founded answers. As shown in Fig. 4a, several typical reasoning pathways utilized by MOFReasoner are demonstrated, including understanding the background, application of knowledge, analysis integration, reasoning expansion, solution evaluation, conclusion formation, and open exploration. In addition to knowledge-based Q&A tasks, MOFReasoner can also generate ideas when prompted. In such cases, it follows a scientific reasoning chain that involves steps like extracting key points, reviewing historical studies, identifying core problems, proposing central ideas, and providing verification strategies and hypothesis-testing procedures. While these reasoning processes are somewhat similar to the CoT patterns found in the training dataset, MOFReasoner adapts its reasoning pathways depending on the nature of the question, indicating that MOFReasoner has effectively learned scientific reasoning through supervised fine-tuning. It is important to note that these reasoning pathways are not manually predefined templates, but rather patterns learned from diverse chain-of-thought examples distilled from research articles, review papers and general CoT datasets. Different question types naturally elicit different combinations of these learned patterns, so the pathways shown in Fig. 4a represent a post hoc summary of recurrent reasoning behaviors rather than fixed decision routes. As illustrated in the expected reasoning path shown in Fig. S30 and further demonstrated in Fig. S31 and S32, compared with DeepSeek R1, MOFReasoner exhibits a more disciplined scientific reasoning style, characterized by systematic contextualization, theory-grounded analysis, and coherent integration of evidence.

To further validate that MOFReasoner has not only learned to reason but also acquired domain knowledge for answering specialized questions, we designed a benchmark consisting of four task categories: experimental studies of MOFs, chemical mechanisms of adsorption, applications of MOF-based adsorbents, and industrialization-related issues (as shown in Fig. 4b and Tables S5–S7). Each question in these tasks was broken down into multiple scoring points. The complete text of all evaluation questions and the detailed scoring points associated with each question are provided in the SI Section S3 to ensure full transparency and reproducibility. Domain experts evaluated the responses based on four criteria: a correct answer (+1), a correct but imprecise answer (+0.5), a wrong or controversial answer (−0.5), and a serious error answer (−1). Key missing information in the model's response was marked as "missing." Since the correct content was already rewarded, no additional penalty was applied for missing points. All models were assessed using exactly the same expert-curated questions and scoring scheme.

As shown in Section S3 and Fig. S33–S104, when comparing different LLMs, we found that the fine-tuned MOFReasoner consistently provided precise answers, addressing the core of each question, without producing severe errors or misleading information. For instance, when asked "How are the dynamic and static adsorption performances of MOFs usually evaluated?", the model correctly distinguished that dynamic adsorption tests employ breakthrough experiments, while static adsorption involves measuring adsorption isotherms. However, possibly due to the imbalance between reasoning chains and final answers in the training dataset (with reasoning tokens significantly outnumbering answer tokens), and the fact that research papers often focus narrowly on single points, MOFReasoner's responses tend to be concise. After thorough reasoning, it retains only the most credible conclusions. For example, for the question "How to determine the adsorption sites in MOF adsorbents?", MOFReasoner conservatively reported DFT calculations and GCMC simulations as methods,

Table 1 The evaluation results of the MOFReasoner, Qwen series, DeepSeek series, and GPT series models

Model	Correct	Inaccurate	Wrong or controversial	Serious error	Missing	Total score
MOFReasoner	25	2	1	0	10	25.5
DeepSeek-R1-Distill-Qwen-7B	15	13	13	23	20	−8
DeepSeek-R1-Distill-Llama-8B	13	8	8	18	22	−5
Qwen-Max	26	16	10	12	9	17
Qwen-Plus	20	11	7	8	15	14
QwQ-32B	24	11	9	13	11	12
DeepSeek-R1-671B	25	14	9	9	10	18.5
o1-preview	26	17	9	16	9	14
GPT-4.5-preview	26	9	11	9	8	16

while omitting single-crystal X-ray diffraction that was considered during reasoning. As summarized in Table 1, MOFReasoner achieved the highest score of 25.5, significantly outperforming its base model DeepSeek-R1-Distill-Qwen-7B, the reasoning model DeepSeek-R1-671B, and even the widely recognized GPT-4.5 and o1 models. Notably, in our benchmark, we observed that GPT-4.5 and o1 occasionally generated literature-style references that were inconsistent with the underlying scientific content or could not be verified (Fig. S74, S83 and S101). Additional control experiments indicate that this performance improvement is not solely due to increased exposure to domain-specific terminology. When trained using only final answers, the resulting model showed limited ability to integrate multiple physicochemical factors and often failed to articulate coherent structure–property relationships relevant to MOF adsorption. Furthermore, comparisons between different model initializations suggest that starting from a reasoning-aligned model facilitates the learning of chemically meaningful reasoning patterns, which are more critical for adsorption-related analysis than increasing the model size alone (Table S8 and Fig. S105–S120).

The capability of reasoning large models should not be limited to Q&A tasks but should extend to providing meaningful scientific assistance. To further illustrate this potential, we tested MOFReasoner with a rarely mentioned guest molecule in the dataset (benzothiophene) and tasked the model with identifying metal clusters that may exhibit strong binding affinity. As shown in Fig. 5a, MOFReasoner reasoned through factors such as coordination strength and charge density and paid particular attention to the sulfur atom in benzothiophene. During the reasoning process, MOFReasoner comprehensively considered factors such as the Lewis acidity of the metal centers, the size and charge density of the metal ions, electronic structure, coordination environment, geometric configuration, and adsorption enthalpy. We observed that MOFReasoner struggled significantly to distinguish between Zn and Co

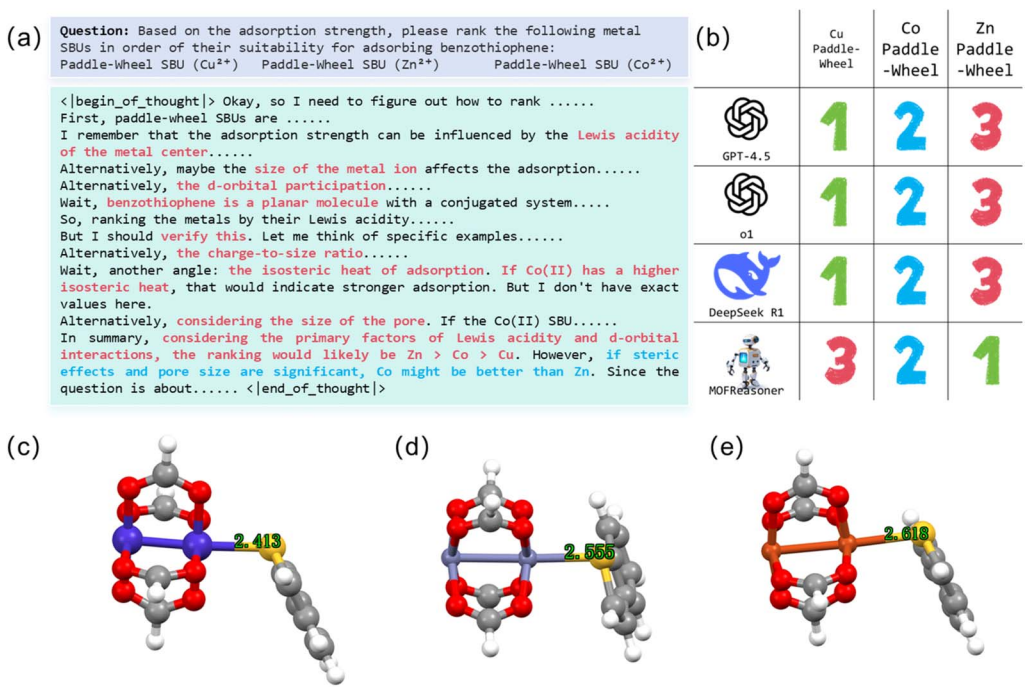

Fig. 5 (a) The reasoning process of MOFReasoner for selecting different SBUs in the adsorption of benzothiophene; (b) the selection results of different models; interaction configurations of benzothiophene with Co paddle-wheel (c), Zn paddle-wheel (d), and Cu paddle-wheel (e).

(Section S4, Table S9) before ultimately ranking the metal ions as Zn^{2+} > Co^{2+} > Cu^{2+}. In contrast, both GPT-4.5 and o1 produced the ranking $Cu^{2+} > Co^{2+} > Zn^{2+}$ (Fig. 5b). This case also reveals limitations in current reasoning behaviors. As shown in the benzothiophene adsorption example and its expected reasoning path (Fig. S121), MOFReasoner shows difficulty in consistently weighting multiple competing physicochemical factors, while the reasoning trace of DeepSeek R1 (Fig. S122) does not explicitly incorporate coordination geometry or framework-level constraints. Through subsequent DFT calculations (Fig. 5c–e), we found that although none of the models initially selected the optimal Co paddle-wheel structure, the Zn and Co paddle-wheel configurations recommended by MOFReasoner exhibited substantially stronger binding affinities than the Cu paddle-wheel structure suggested by GPT-4.5 and o1. Specifically, the Co paddle-wheel structure outperformed Zn by 14.21 kJ mol^{-1} and Cu by 25.96 kJ mol^{-1}, indicating that the Co metal node forms a stronger interaction with benzothiophene and therefore provides a more favorable adsorption configuration. These results indicate that, while MOFReasoner's reasoning still deviates from the actual optimal choice, its inference process can provide useful qualitative guidance and serve as a proof-of-concept example for assisting scientific reasoning tasks.

Conclusion

In this study, we developed MOFReasoner, a domain-specific large language model fine-tuned for scientific reasoning in chemical research, with a particular focus on MOF adsorption. By combining knowledge distillation, CoT reasoning extraction, and systematic dataset validation, MOFReasoner achieves substantial improvements in accuracy, reliability, and scientific depth compared to general-purpose LLMs. Our results show that MOFReasoner not only performs well in knowledge-based Q&A tasks but also suggests the potential of domain-specific language models for scientific reasoning tasks, including hypothesis generation and qualitative material screening that are consistent with DFT trends. This work provides a promising framework for future development of domain-specific scientific LLMs and highlights the importance of integrating structured knowledge, reasoning mechanisms, and expert validation. We note that the current framework is text-based and does not directly process graphical data such as adsorption isotherms, diffraction patterns, or microscopy images; extending MOFReasoner toward multimodal figure understanding represents an important direction for future improvement. MOFReasoner sets a foundation for advancing AI-assisted scientific research, paving the way for more intelligent, reliable, and application-oriented models in the field of chemistry.

Author contributions

J.-R. L. conceived and designed the study. X. B. wrote the initial manuscript. H.-T. W. and R. Y. participated in discussing and editing the manuscript. J.-R. L., Z. Z., and X. Z. performed the supervision, review, and editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

The code supporting the findings of this study is publicly available *via* Zenodo at https://doi.org/10.5281/zenodo.18161608, which archives the corresponding GitHub repository (https://github.com/MontageBai/ChemReasoner-Code/). The model resources used in this study, excluding the full model weights, are available *via* Zenodo at https://doi.org/10.5281/zenodo.18163201. The actively maintained version of the model weights is available at https://huggingface.co/baixuefeng/ChemReasoner-7B/.

Supplementary information (SI): detailed methodology for dataset construction and knowledge distillation, chain-of-thought extraction procedures, model training and evaluation details, as well as supplementary tables and figures supporting the main text. See DOI: https://doi.org/10.1039/d5dd00429b.

Acknowledgements

We acknowledge the financial support from the National Natural Science Foundation of China (No. 22225803 and 22278011), Beijing Natural Science Foundation (No. Z230023), and the Beijing Outstanding Young Scientist Program (Project No. JWZQ20240102008)

Notes and references

1 X. Bai, Y. Li, Y. Xie, Q. Chen, X. Zhang and J.-R. Li, High-throughput screening of CO_2 cycloaddition MOF catalyst with an explainable machine learning model, *Green Energy Environ.*, 2025, **10**(1), 132–138.

2 C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, *et al.*, A robotic platform for flow synthesis of organic compounds informed by AI planning, *Science*, 2019, **365**(6453), eaax1566.

3 R. L. Greenaway, K. E. Jelfs, A. C. Spivey and S. N. Yaliraki, From alchemist to AI chemist, *Nat. Rev. Chem.*, 2023, **7**(8), 527–528.

4 X. Bai, Y. Xie, X. Zhang, H. Han and J.-R. Li, Evaluation of Open-Source Large Language Models for Metal–Organic Frameworks Research, *J. Chem. Inf. Model.*, 2024, **64**(13), 4958–4965.

5 M. C. Ramos, C. J. Collison and A. D. White, A review of large language models and autonomous agents in chemistry, *Chem. Sci.*, 2025, **16**(6), 2514–2572.

6 K. M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J. D. Bocarsly, A. M. Bran, S. Bringuier, L. C. Brinson, K. Choudhary, D. Circi, *et al.*, 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon, *Digital Discovery*, 2023, **2**(5), 1233–1250.

7 Z. Zheng, N. Rampal, T. J. Inizan, C. Borgs, J. T. Chayes and O. M. Yaghi, Large language models for reticular chemistry, *Nat. Rev. Mater.*, 2025, **10**, 369–381.
8 J. Büchel, A. Vasilopoulos, W. A. Simon, I. Boybat, H. Tsai, G. W. Burr, H. Castro, B. Filipiak, M. Le Gallo, A. Rahimi, *et al.*, Efficient scaling of large language models with mixture of experts and 3D analog in-memory computing, *Nat. Comput. Sci.*, 2025, **5**(1), 13–26.
9 R. Haase, Towards transparency and knowledge exchange in AI-assisted data analysis code generation, *Nat. Comput. Sci.*, 2025, **5**, 271–272.
10 M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez and K. M. Jablonka, From text to insight: large language models for chemical data extraction, *Chem. Soc. Rev.*, 2025, **54**(3), 1125–1150.
11 L. M. Antunes, K. T. Butler and R. Grau-Crespo, Crystal structure generation with autoregressive large language modeling, *Nat. Commun.*, 2024, **15**(1), 10570.
12 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, Augmenting large language models with chemistry tools, *Nat. Mach. Intell.*, 2024, **6**(5), 525–535.
13 N. Liu, S. Jafarzadeh, B. Y. Lattimer, S. Ni, J. Lua and Y. Yu, Harnessing large language models for data-scarce learning of polymer properties, *Nat. Comput. Sci.*, 2025, **5**(3), 245–254.
14 Z. Zheng, Z. He, O. Khattab, N. Rampal, M. A. Zaharia, C. Borgs, J. T. Chayes and O. M. Yaghi, Image and data mining in reticular chemistry powered by GPT-4V, *Digital Discovery*, 2024, **3**(3), 491–501.
15 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis, *J. Am. Chem. Soc.*, 2023, **145**(32), 18048–18062.
16 Z. Zheng, Z. Rong, N. Rampal, C. Borgs, J. T. Chayes and O. M. Yaghi, A GPT-4 Reticular Chemist for Guiding MOF Discovery, *Angew. Chem., Int. Ed.*, 2023, **62**(46), e202311983.
17 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, ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs, *ACS Cent. Sci.*, 2023, **9**(11), 2161–2170.
18 T. Song, M. Luo, X. Zhang, L. Chen, Y. Huang, J. Cao, Q. Zhu, D. Liu, B. Zhang, G. Zou, *et al.*, A Multiagent-Driven Robotic AI Chemist Enabling Autonomous Chemical Research On Demand, *J. Am. Chem. Soc.*, 2025, **147**(15), 12534–12545.
19 K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, Leveraging large language models for predictive chemistry, *Nat. Mach. Intell.*, 2024, **6**(2), 161–169.
20 Z. Zheng, A. H. Alawadhi, S. Chheda, S. E. Neumann, N. Rampal, S. Liu, H. L. Nguyen, Y.-h. Lin, Z. Rong, J. I. Siepmann, *et al.*, Shaping the Water-Harvesting Behavior of Metal–Organic Frameworks Aided by Fine-Tuned GPT Models, *J. Am. Chem. Soc.*, 2023, **145**(51), 28284–28295.
21 Y. Kang and J. Kim, ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models, *Nat. Commun.*, 2024, **15**(1), 4705.
22 X. Bai, S. He, Y. Li, Y. Xie, X. Zhang, W. Du and J.-R. Li, Construction of a knowledge graph for framework material enabled by large language models and its application, *npj Comput. Mater.*, 2025, **11**(1), 51.
23 C. M. Castro Nascimento and A. S. Pimentel, Do Large Language Models Understand Chemistry? A Conversation with ChatGPT, *J. Chem. Inf. Model.*, 2023, **63**(6), 1649–1655.
24 A. D. White, G. M. Hocky, H. A. Gandhi, M. Ansari, S. Cox, G. P. Wellawatte, S. Sasmal, Z. Yang, K. Liu, Y. Singh, *et al.*, Assessment of chemistry knowledge in large language models that generate code, *Digital Discovery*, 2023, **2**(2), 368–376.
25 X. Chen, H. Yi, M. You, W. Liu, L. Wang, H. Li, X. Zhang, Y. Guo, L. Fan, G. Chen, *et al.*, Enhancing diagnostic capability with multi-agents conversational large language models, *npj Digit. Med.*, 2025, **8**(1), 159.
26 Z. Zhang, Y. Yao, A. Zhang, X. Tang, X. Ma, Z. He, Y. Wang, M. Gerstein, R. Wang, G. Liu, *et al.*, Igniting Language Intelligence: The Hitchhiker's Guide from Chain-of-Thought Reasoning to Language Agents, *ACM Comput. Surv.*, 2025, **57**(8), 1–39.
27 R. Cantini, A. Orsino and D. Talia, Xai-driven knowledge distillation of large language models for efficient deployment on low-resource devices, *J. Big Data*, 2024, **11**(1), 63.
28 C. Yang, Y. Zhu, W. Lu, Y. Wang, Q. Chen, C. Gao, B. Yan and Y. Chen, Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application, *ACM Trans. Intell. Syst. Technol.*, 2024, **16**(6), 1–27.
29 B. Zhang, H. Ma, J. Ding, J. Wang, B. Xu and H. Lin, Distilling implicit multimodal knowledge into large language models for zero-resource dialogue generation, *Inf. Fusion*, 2025, **118**, 102985.
30 X. Chen, T. Wang, T. Guo, *et al.*, Unveiling the power of language models in chemical research question answering, *Commun. Chem.*, 2025, **8**, 4.
31 X. Zhang, Y. Li and J.-R. Li, Metal-organic frameworks for multicomponent gas separation, *Trends Chem.*, 2024, **6**(1), 22–36.
32 N. F. Suremann, B. D. McCarthy, W. Gschwind, A. Kumar, B. A. Johnson, L. Hammarström and S. Ott, Molecular Catalysis of Energy Relevance in Metal–Organic Frameworks: From Higher Coordination Sphere to System Effects, *Chem. Rev.*, 2023, **123**(10), 6545–6611.
33 X. Luo, M. Zhang, Y. Hu, Y. Xu, H. Zhou, Z. Xu, Y. Hao, S. Chen, S. Chen, Y. Luo, *et al.*, Wrinkled metal-organic framework thin films with tunable Turing patterns for pliable integration, *Science*, 2024, **385**(6709), 647–651.
34 I. Abánades Lázaro, X. Chen, M. Ding, A. Eskandari, D. Fairen-Jimenez, M. Giménez-Marqués, R. Gref, W. Lin, T. Luo and R. S. Forgan, Metal–organic frameworks for biological applications, *Nat. Rev. Methods Primers*, 2024, **4**(1), 42.
35 DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, *et al.*, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs

via Reinforcement Learning, *arXiv*, 2025, preprint arXiv.2501.12948, DOI: 10.48550/arXiv.2501.12948.

36 F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca and M. Kraft, OntoKin: An Ontology for Chemical Kinetic Reaction Mechanisms, *J. Chem. Inf. Model.*, 2020, **60**(1), 108–120.

37 J. Bai, S. D. Rihm, A. Kondinski, F. Saluz, X. Deng, G. Brownbridge, S. Mosbach, J. Akroyd and M. T. Kraft, The World Avatar Python Package for Dynamic Knowledge Graphs and Its Application in Reticular Chemistry, *Digital Discovery*, 2025, **4**(8), 2123–2135.

38 L. Pascazio, S. Rihm, A. Naseri, S. Mosbach, J. Akroyd and M. Kraft, Chemical Species Ontology for Data Integration and Knowledge Discovery, *J. Chem. Inf. Model.*, 2023, **63**(21), 6569–6586.

39 G. Li, H. A. A. K. Hammoud, H. Itani, D. Khizbullin, B. Ghanem, CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Model Society, *arXiv*, 2023, preprint arXiv.2303.17760, DOI: 10.48550/arXiv.2303.17760.

40 X. Cheng, J. Li, W. X. Zhao and J.-R. Wen, Think More, Hallucinate Less: Mitigating Hallucinations via Dual Process of Fast and Slow Thinking, *arXiv*, 2025, preprint, arXiv:2501.01306, DOI: 10.48550/arXiv.2501.01306.