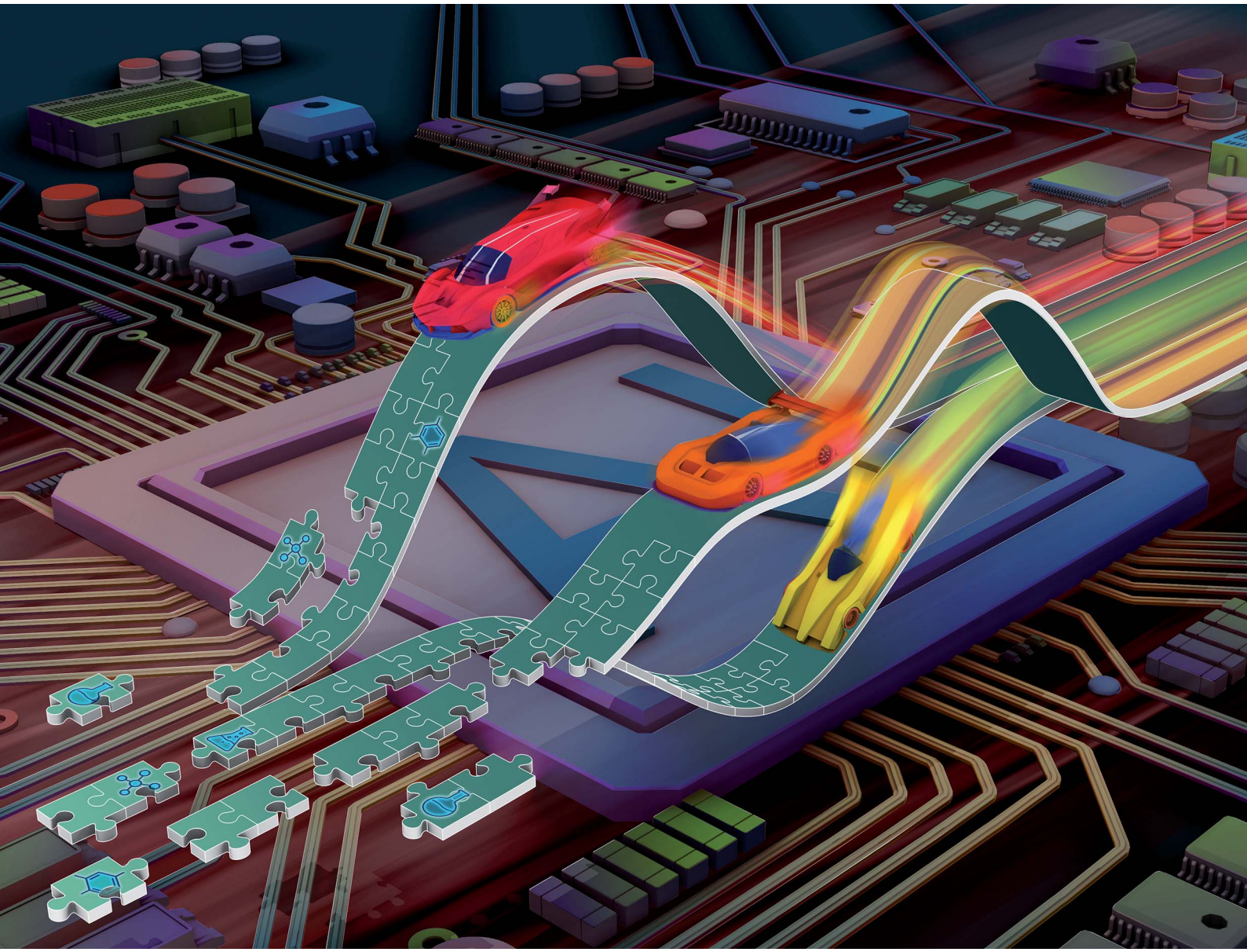
Volume 4
Number 10
October 2025
Pages 2643-3054

Digital Discovery

rsc.li/digitaldiscovery



ISSN 2635-098X

ROYAL SOCIETY OF CHEMISTRY

PAPER

Omar M. Yaghi *et al.*

Comparison of LLMs in extracting synthesis conditions and generating Q&A datasets for metal–organic frameworks

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 2676

Comparison of LLMs in extracting synthesis conditions and generating Q&A datasets for metal–organic frameworks†

Yuang Shi, ‡[ab] Nakul Rampal, ‡[abc] Chengbin Zhao, [abc] Dongrong Joe Fu, [c] Christian Borgs, [cd] Jennifer T. Chayes [cdefg] and Omar M. Yaghi *[abch]

Received 2nd March 2025
Accepted 13th May 2025

DOI: 10.1039/d5dd00081e

rsc.li/digitaldiscovery

Artificial intelligence, represented by large language models (LLMs), has demonstrated tremendous capabilities in natural language recognition and extraction. To further evaluate the performance of various LLMs in extracting information from academic papers, this study explores the application of LLMs in reticular chemistry, focusing on their effectiveness in generating Q&A datasets and extracting synthesis conditions from scientific literature. The models evaluated include OpenAI's GPT-4 Turbo, Anthropic's Claude 3 Opus, and Google's Gemini 1.5 Pro. Key results indicate that Claude excelled in providing complete synthesis data, while Gemini outperformed others in accuracy, characterization-free compliance (obedience), and proactive structuring of responses. Although GPT-4 was less effective in quantitative metrics, it demonstrated strong logical reasoning and contextual inference capabilities. Overall, Gemini and Claude achieved the highest scores in accuracy, groundedness, and adherence to prompt requirements, making them suitable benchmarks for future studies. The findings reveal the potential of LLMs to aid in scientific research, particularly in the efficient construction of structured datasets, which can help train models, predict, and assist in the synthesis of new metal–organic frameworks (MOFs).

Introduction

The integration of large language models (LLMs) in reticular chemistry—the branch of chemistry concerned with the synthesis of extended crystalline materials connected by strong bonds[1]—is transforming the way laboratory research is being conducted. Recent studies have demonstrated that LLMs can be applied across a variety of different tasks in reticular chemistry, such as, (i) assisting in the design and optimization of synthesis conditions for metal–organic frameworks (MOFs),[2] (ii) enabling the high-throughput extraction of scientific insights, including graphical data such as X-ray diffraction patterns and adsorption isotherms,[3] (iii) editing and generating new linker chemistries to design MOFs for applications such as water harvesting,[4] (iv) generating novel MOF structures given specific prompts or desired performance criteria,[5] and (v) creating computational scripts and polishing papers.[6]

Importantly, all the use cases described above require high quality datasets as they provide the foundation for the training, testing, and benchmarking of different LLMs. Most datasets involve using the published scientific literature as the primary input. This input data is diverse and very often, unstructured, making the task of using this information less easy. Given these challenges, it is critical to ensure that the datasets generated are both comprehensive, that is representative of the broad scientific literature available, and accurate, where the information contained within the dataset is correct. Since LLMs are now also often being used to generate these datasets,[7,8] knowledge of which LLM is useful for a particular task, we believe, is helpful for the community.

In this study, we compare the performance of different LLMs—OpenAI's GPT-4 Turbo (hereafter abbreviated as GPT-4),[9] Anthropic's Claude 3 Opus (Claude)[10] and Google's Gemini 1.5 Pro(Gemini)[11]—in two tasks, (i) generating question–answer (Q&A) pairs, and (ii) the extraction of synthesis conditions of

[a]*Department of Chemistry, University of California, Berkeley, California 94720, USA. E-mail: yaghi@berkeley.edu*

[b]*Kavli Energy NanoScience Institute, University of California, Berkeley, California 94720, USA*

[c]*Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, University of California, Berkeley, California 94720, USA*

[d]*Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720, USA*

[e]*Department of Mathematics, University of California, Berkeley, California 94720, USA*

[f]*Department of Statistics, University of California, Berkeley, California 94720, USA*

[g]*School of Information, University of California, Berkeley, California 94720, USA*

[h]*KACST-UC Berkeley Center of Excellence for Nanomaterials for Clean Energy Applications, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5dd00081e

‡ Equal contribution.

MOFs, from the scientific literature by subject matter experts. Building on the RetChemQA dataset,[12] we provide a thorough quantitative and qualitative comparison of each model's performance identifying the key strengths, limitations, and specific challenges encountered when generating the datasets. We find that for extracting synthesis conditions from the literature, Claude is the most comprehensive and accurate LLMs, while for generating Q&A pairs datasets Gemini performance is the best. Furthermore, we report the particulars of where these LLMs fail, but also where they demonstrate "amazing" and unexpected results. These findings provide insights towards advancing the use of LLM powered AI agents in reticular chemistry.

Methods

Dataset selection and model processing

To evaluate the performance of different LLMs, we utilized the refined CSD MOF subset, building on prior work.[12] To ensure consistency and facilitate meaningful comparisons across LLMs, we randomly selected 50 Digital Object Identifiers (DOIs) from this subset for each specific task: synthesis condition extraction, single-hop Q&A generation, and multi-hop Q&A generation. The number of the selected DOIs for each task can be found in the ESI (Table S1).†

The datasets for synthesis conditions, single-hop Q&A, and multi-hop Q&A were generated using the same workflow as described in our previous work.[12] Briefly, this workflow involved initializing the environment, parsing the relevant files, and tokenizing the combined text. Once tokenized, these texts were processed by three different LLMs: OpenAI's GPT-4 Turbo (gpt-4-0125-preview, data generated on March, 2024), Anthropic's Claude 3 Opus (data generated from June to October, 2024), and Google's Gemini 1.5 Pro (gemini-1.5-pro-001, data generated from May to October, 2024).

Task prompts and processing

The prompts used for each task are detailed in the ESI (Fig. S1).† For the synthesis conditions extraction task, the LLMs were instructed to extract synthesis parameters for MOF products, including temperature, concentration, reagent quantities, and solvents. Importantly, the LLMs were explicitly instructed to exclude any characterization data, such as analytical measurements, ensuring the focus remained solely on synthesis-related details. This exclusion criterion was a key part of our evaluation to maintain the relevance of extracted information.

For the single-hop and multi-hop Q&A tasks, LLMs were instructed to generate and answer questions based on both the main content of the paper and its ESI. Single-hop questions involved straightforward inquiries that could be addressed using a single section of the text, while multi-hop questions required synthesizing information from multiple sections, thereby assessing the models' capabilities in complex reasoning and information integration.

Furthermore, we specified the type and number of questions for the single-hop and multi-hop Q&A tasks: 6 factual, 7 reasoning, and 7 true-or-false questions for each DOI. Evaluating whether the LLMs adhered to these requirements was an important metric, allowing us to assess their ability to follow the prompt specifications accurately.

Evaluation criteria for the synthesis conditions extraction task

After obtaining the outputs from the LLMs, we applied standardized criteria for evaluation. For the synthesis conditions task, we used three criteria: completeness, correctness, and characterization-free compliance.

Completeness is a measure of whether all relevant parameters, such as temperature, concentration, and reagent amount for the synthesis of a specific product, are included.

Correctness refers to the accuracy of the extracted information, meaning that all extracted data had to be precise and without errors. This criterion ensured that every piece of information provided by the LLMs was factually correct and aligned with the original content.

Characterization-free compliance is a measure of how obedient the LLM in following the instructions given in the prompt. In this case, whether the LLMs exclude characterization details—such as temperatures for adsorption isotherms or chemical shifts from nuclear magnetic resonance (NMR)—or not.

As Fig. S2† shows, for each of these criteria, the LLM output was marked as "Y" if it met the requirement and "N" if it did not. In cases where multiple products were reported for a single DOI, we collected the synthesis conditions for all products and human-evaluated the LLM-extracted information for each product individually based on the aforementioned criteria. Occasionally, the LLMs failed to extract synthesis condition information for some products; in such situations, we assigned a value of "\" for each of the three criteria.

After assigning "Y", "N", or "\" for each criterion of each product extracted by each LLM, we calculated the corresponding proportions of Y, N, and \. Additionally, we introduced a parameter called net-Y-ratio, which is defined as the ratio of "Y" to the total extracted information (*i.e.*, Y + N). This parameter measures the accuracy of the extracted information, independent of the overall completeness of the dataset.

This evaluation framework allowed us to thoroughly assess the performance of each LLM, not only in terms of the quality of the information extracted but also in their adherence to specific requirements and ability to exclude irrelevant details.

Evaluation criteria for single-hop and multi-hop Q&A tasks

For the evaluation of single-hop and multi-hop Q&A tasks, we employed two criteria: accuracy and groundedness.

Accuracy is the measure of whether the answer provided is correct, meaning that the response must be factually accurate and directly address the question posed.

Groundedness is the measure of the questions are based on the provided article, rather than relying on common sense or hallucinated information. This criterion assesses whether the generated questions and answers are properly anchored in the context of the provided material.

If a question was relevant to the content and the corresponding answer was correct, we classified this instance as True Positive (TP). If the question was hallucinated but the answer was still correct, or if the LLM correctly identified the question as hallucinated—such as by responding with "this question is not related to the content" or "I don't know the answer"—we marked this as True Negative (TN). In cases where the question was well-formed but the answer was incorrect, we categorized the outcome as False Positive (FP). If the question was irrelevant to the main text and the LLM failed to identify it as hallucinated, we classified it as False Negative (FN).

After assigning TP, TN, FP, and FN labels to each question–answer pair, we applied a modified version of our previous evaluation framework.[12] To provide a more intuitive assessment of LLM performance, we used the following four metrics:

Accuracy: this metric evaluates the ability of the LLMs to provide correct answers, regardless of the quality of the question. Accuracy is calculated as the ratio of all correctly handled instances (TP + TN) to the total number of Q&A pairs (TP + TN + FP + FN).

Groundedness: this metric is conceptually the inverse of the "Hallucination Rate" used in our previous framework.[12] It measures the quality of the questions, specifically whether they are based on the context provided. Groundedness is calculated as the ratio of in-context questions (TP + FP) to the total number of Q&A pairs (TP + TN + FP + FN).

Precision: unlike accuracy and groundedness, precision considers both the quality of the question and the correctness of the answer. This means that hallucinated questions or incorrect answers are penalized. Precision is calculated as the ratio of accurately answered, in-context questions (TP) to the total number of Q&A pairs (TP + TN + FP + FN).

Hallucination capture rate: This is a measure to evaluate the ability of LLMs to self-correct when faced with irrelevant or erroneous questions. Hallucination capture rate is calculated as the ratio of hallucinated questions correctly identified (TN) to the total number of hallucinated questions (TN + FN).

We observed that despite explicitly specifying in the prompt that each entry of datasets should contain 6 factual, 7 reasoning, and 7 true-or-false questions, the LLM-generated Q&A datasets did not always adhere to this requirement. To address this, we introduced an alignment parameter for each LLM and each DOI. The alignment parameter was calculated as the proportion of DOIs (out of 50) for which the generated responses met the prompt requirements. This allowed us to evaluate how consistently each LLM adhered to the required question distribution across the entire dataset.

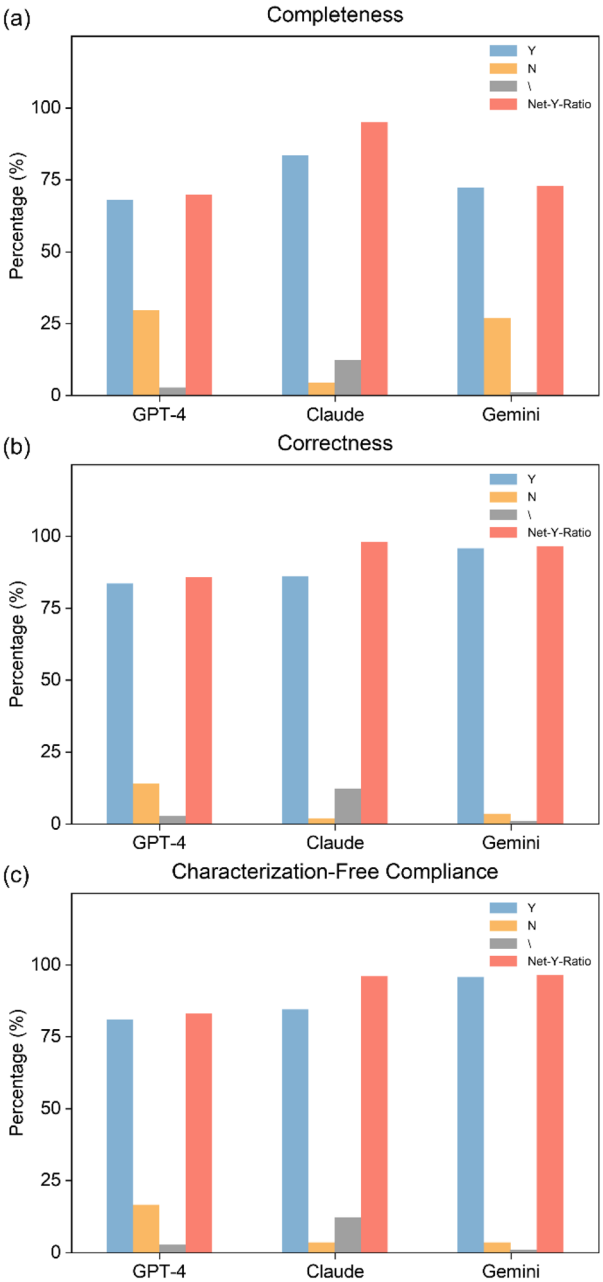


Fig. 1 Performance of LLMs in the synthesis condition extraction task. (a) GPT-4; (b) Claude; (c) Gemini. 'Y' represents the proportion of products where information is successfully extracted and meets the corresponding requirements. 'N' is the ratio where information is extracted but does not meet the requirements. '\' represents the proportion of products whose synthesis conditions are not extracted. Net-Y-ratio is calculated as the ratio of Y to (Y + N).

Results

For the selected 50 DOIs, we extracted a total of 115 products and evaluated the performance of LLMs in extracting information based on completeness, correctness, and characterization-free compliance.

Fig. 1 shows the performance of GPT-4, Claude, and Gemini in the synthesis conditions extraction task. For the first criterion, completeness (Fig. 1a), Claude successfully extracted complete synthesis information for 83.5% of the products, outperforming Gemini's 72.2% and GPT-4's 67.8%. This suggests that Claude is more effective at covering all necessary details for synthesis across a wide range of products.

For the correctness criterion (Fig. 1b), Gemini demonstrated exceptional performance with an accuracy of 95.7%, significantly better than Claude's 86.1% and GPT-4's 83.5%. This highlights Gemini's ability to extract precise information

without errors, which is critical for maintaining the reliability of the database.

Regarding characterization-free compliance (Fig. 1c), Gemini again excelled with a compliance rate of 95.7%, surpassing Claude's 84.4% and GPT-4's 80.9%. This indicates that Gemini was most successful in adhering to the requirement of excluding characterization data, which is essential for creating a synthesis conditions-focused dataset.

Notably, for the net-Y-ratio, which excludes the 12.2% of product data missed by Claude, Claude's performance in both correctness and characterization-free compliance was comparable to that of Gemini (>96%). This indicates that, when focusing solely on the data that was successfully extracted, both Claude and Gemini generated highly accurate and compliant datasets. The primary difference lay in the scope of coverage, with Claude's dataset ignoring some of the products. Considering that Claude achieved high net ratios (>95%) across all three criteria, we conclude that Claude exhibited the best overall performance in generating synthesis condition databases, balancing both accuracy and coverage.

Subjective evaluation of "amazing" and unexpected responses

In addition to quantitatively comparing the performance of the three LLMs, we conducted a subjective qualitative assessment of the generated datasets. We identified several "amazing" responses that demonstrated the LLMs' logical reasoning capabilities when handling complex tasks.

Fig. 2a showcases an "amazing" response from GPT-4. In the original text, the authors provided complete synthesis conditions for compound **1** but only mentioned that "**1**@Y was prepared by the same synthetic procedure as **1** with $ErCl_3 \cdot 6H_2O$ and $YCl_3 \cdot 6H_2O$ in a 1 : 7 molar ratio" when describing the synthesis of compound **1**@Y. While Claude and Gemini successfully extracted the synthesis conditions for compound **1**, they did not capture the synthesis details for compound **1**@Y. In contrast, GPT-4 understood the context and duplicated the synthesis conditions for compound **1** while updating the chemicals used entry to include "$ErCl_3 \cdot 6H_2O$ and $YCl_3 \cdot 6H_2O$ in a 1 : 7 molar ratio". This demonstrates GPT-4's ability to infer and apply contextual information effectively, a key aspect of advanced data extraction.

Fig. 2b presents a remarkable response from Claude. When describing the washing process of a product, the original text stated: "before drying, the wet solid powder was immediately transferred to a 500 mL glass jar, and 400 mL of DMF was added. The jar was heated at 120 °C for 12 h, then cooled to room temperature. The DMF was decanted and replaced with 400 mL of fresh DMF. The jar was reheated at 120 °C, followed by decanting and replacing with fresh DMF. This was repeated four additional times". The total number of DMF washings was therefore 1 + 1 + 4 = 6, which Claude correctly calculated. This response highlights Claude's capacity to perform arithmetic reasoning and accurately interpret iterative procedures, which is crucial for understanding experimental protocols.

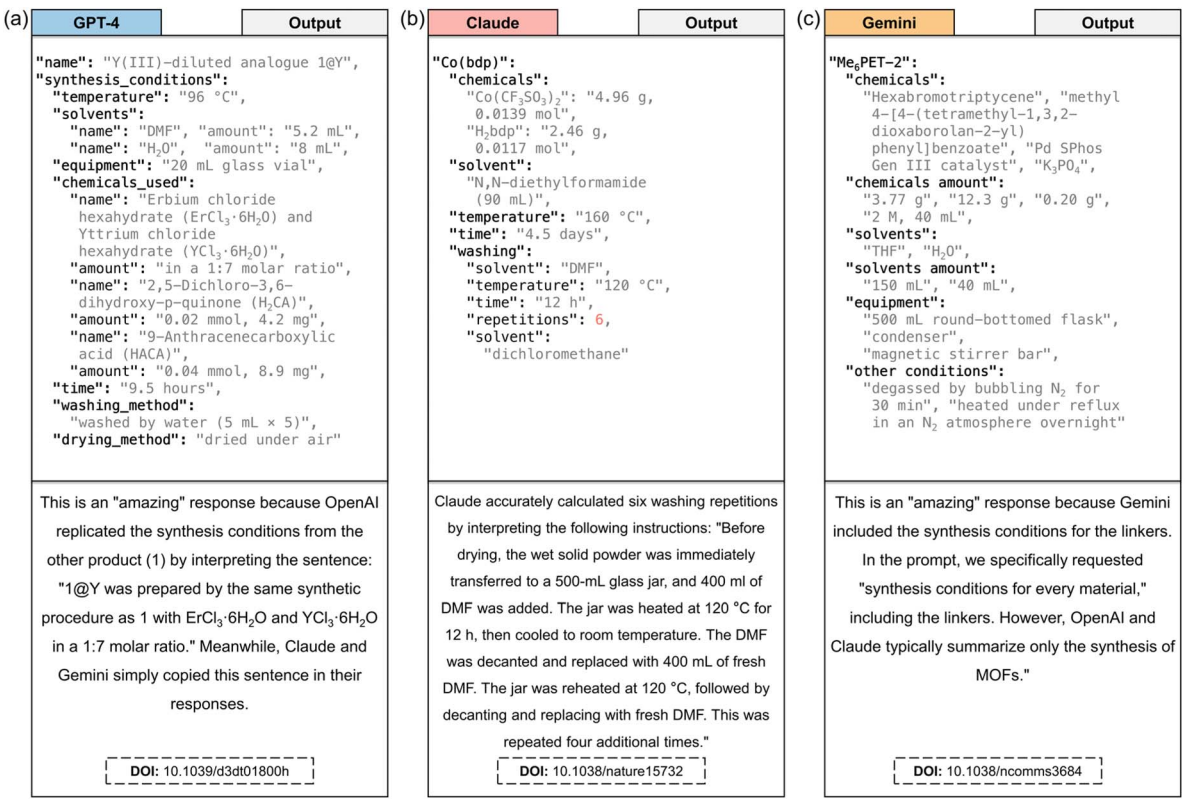


Fig. 2 "Amazing" responses of (a) GPT-4, (b) Claude, and (c) Gemini in the synthesis condition extraction task.

Fig. 2c highlights Gemini's outstanding performance in providing comprehensive synthesis details. In our synthesis conditions prompt, we explicitly requested synthesis conditions for every material, which included both MOFs and organic linkers. While GPT-4 and Claude often focused on extracting synthesis conditions for MOFs only, frequently overlooking the synthesis of organic linkers, Gemini consistently searched both the main text and the ESI to ensure no synthesis detail was omitted. This completeness is reflected in the statistical data, where Gemini received the fewest "\" labels, indicating a more complete and exhaustive extraction process. Gemini's diligence in capturing all relevant synthesis details underscores its potential for tasks that require exhaustive data compilation.

These "amazing" responses illustrate that LLMs possess a considerable level of logical reasoning ability when handling complex tasks, showing their potential to significantly improve dataset construction. Their ability to infer information, perform arithmetic operations, and ensure thorough extraction highlights their versatility in dealing with nuanced scientific content. By refining prompts and applying targeted preprocessing to the literature, we believe LLM-based information extraction can become even more accurate and consistent. This progress will make LLMs valuable tools for building comprehensive scientific datasets that support efficient research and discovery.

Error analysis of LLMs

We analyzed the main causes of errors for each of the three LLMs. Fig. 3 presents some representative examples of the types of mistakes made by the models, highlighting common patterns in their performance.

For GPT-4, the primary cause of errors in completeness was missing key details, such as the amounts of chemicals used (Fig. 3a). Additionally, GPT-4 often missed gas reagents like H_2S or HCl. Moreover, in terms of correctness, GPT-4 struggled with accurately reporting temperatures during multi-step synthesis, which suggests difficulties in effectively extracting intricate details throughout complex synthesis processes.

For Claude, the main issue affecting completeness was missing certain conditions or concentrations, and regarding correctness, Claude made mistakes related to misinterpreting methods or temperature values. For instance, as shown in Fig. 3b, Claude incorrectly identified the synthesis method, mistaking liquid–liquid diffusion (used for synthesizing single crystals) for a method applicable to polycrystals, indicating that LLMs can confuse similar locations or content, leading to 'cross-contamination' in their outputs. As a result, although Claude generally captured more synthesis details, it occasionally struggled to accurately interpret specific experimental procedures.

For Gemini, the primary issue in completeness was missing amounts of chemicals. Fig. 3c shows how Gemini incorrectly interpreted "DMF–H_2O mixture (v/v 1/1, 1 mL)" as 1 mL each of DMF and H_2O instead of a total volume of 1 mL. In terms of correctness, Gemini also faced challenges with incorrect quantities, often due to ambiguous measurements. These errors indicate that Gemini's thorough approach sometimes led to incorrect extrapolation of details.

For the characterization-free compliance evaluation, as shown in Fig. 3d, GPT-4 frequently included characterization data, while Claude and Gemini were more effective in excluding such information, which aligns with the earlier quantitative results (Fig. 1). Therefore, Claude and Gemini demonstrated a better ability to distinguish synthesis-specific data from characterization details, which is crucial for creating a focused and relevant dataset.

Overall, the error analysis highlights specific areas where each LLM excelled or struggled. We found that LLMs often lost or misinterpreted numerical information during extraction, such as reagent amounts, temperatures, and concentrations, whereas errors related to reagent or product names were rare. This discrepancy may be due to the inherent mechanisms of LLM generation. Reagent and product names typically appear only once in synthesis descriptions, enabling LLMs to store them accurately without confusion. In contrast, similar numerical data is presented in multiple contexts. This can lead to confusion or overwriting of previously stored information, resulting in cross-contamination of details. Understanding this pattern allows us to better determine which types of information should be prioritized in databases to enhance overall accuracy and reliability.

Q&A dataset generation

For the randomly selected 50 DOIs, we require each model to generate 20 Q&A pairs for each DOI. Thus, the total size of the database is approximately 6000 Q&A pairs (50 DOIs $\times$ 20 Q&A pairs $\times$ 3 models (GPT-4, Claude, and Gemini) $\times$ 2 tasks (single-hop and multi-hop)). Every Q&A pair is evaluated by subject matter experts. It is worth mentioning that the evaluation results of GPT-4 have already been published in the RetChemQA dataset program.[12]

For the single-hop task, we evaluated the three LLMs—GPT-4, Claude, and Gemini—based on accuracy (to assess the quality of the answers), groundedness (to evaluate whether the questions were derived from the paper), precision (to determine whether questions were derived from the main text and answered correctly), and hallucination capture rate (to assess the self-correction ability of LLMs). As shown in Fig. 4a, both Claude and Gemini achieved scores exceeding 99% in accuracy, groundedness and precision. And the hallucination capture rates of both were 50.0%. In contrast, GPT-4's accuracy was 90.7%, while groundedness was slightly higher at 92.6%, precision was 89.8%, and hallucination capture rate was only 11.4%. These results indicate that Claude and Gemini are highly reliable for generating single-hop Q&A datasets, making them suitable as benchmarks for future studies. The consistently high performance of these two models suggests a strong ability to extract relevant information accurately while maintaining a high standard of quality across diverse question types and categories.

Fig. 4b indicates that the results for the multi-hop task were similar to those for the single-hop task. Claude and Gemini

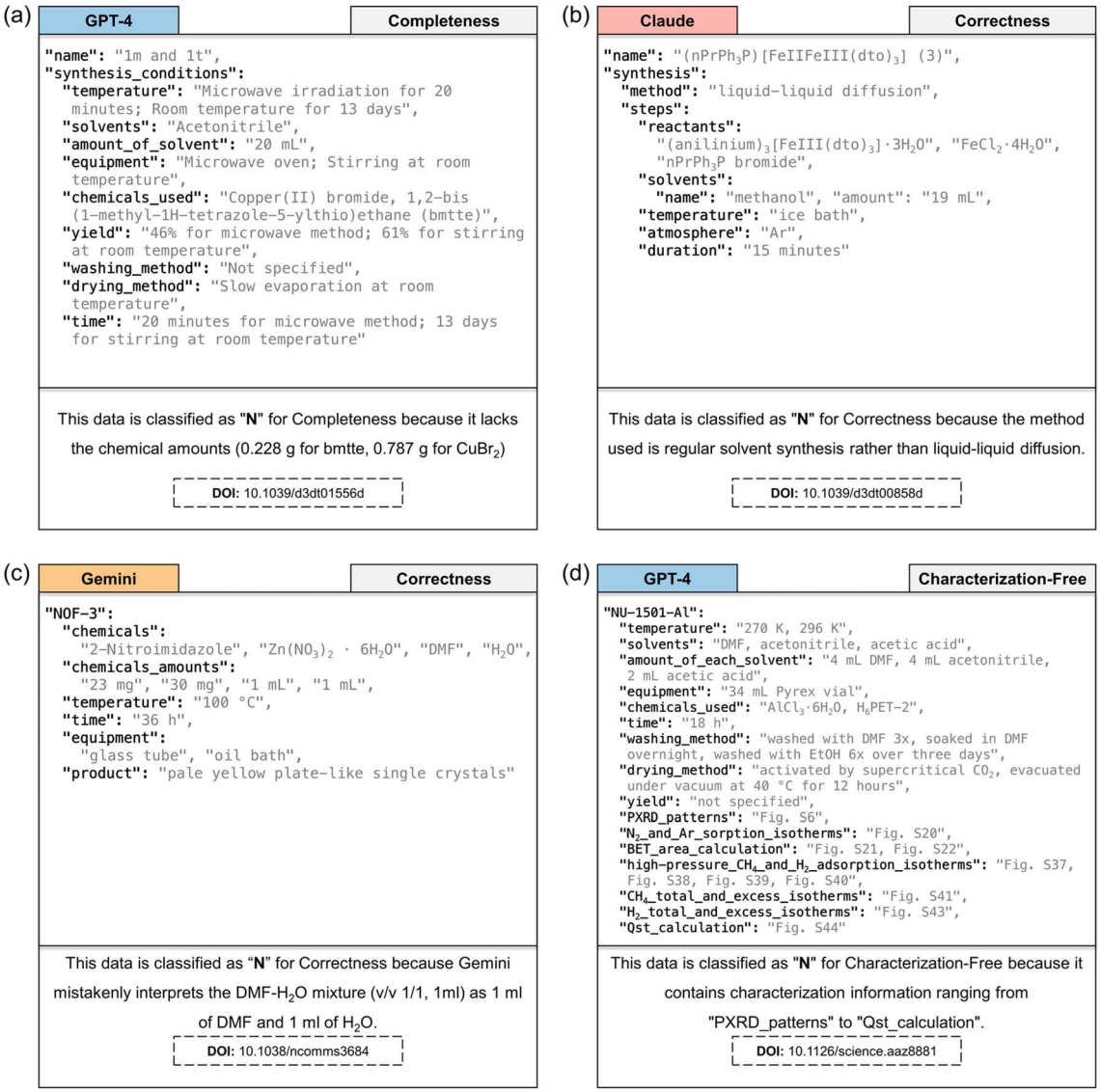


Fig. 3 Examples of wrong responses of LLMs in the synthesis condition extraction task. (a) GPT-4's mistake in completeness; (b) Claude's mistake in correctness; (c) Gemini's mistake in correctness; (d) GPT-4's mistake in characterization-free compliance.

continued to perform exceptionally well. Notably, GPT-4 showed slight improvements in both groundedness and precision, with significant increases in accuracy and hallucination capture rate, bringing them closer to the levels achieved by Claude and Gemini. As previously reported,[12] this improvement may be attributed to GPT-4 "thinking" more thoroughly when responding to the revised prompt, leading to better self-correction when faced with hallucinated questions in the multi-hop task compared to the single-hop task. This highlights the importance of prompt engineering in leveraging the full capabilities of LLMs, especially when dealing with nuanced, multi-step reasoning tasks. Moreover, recent work[12] has shown that prompt engineering can help encourage the generation of Q&A pairs with multi-hop reasoning, which helps better differentiate the capabilities of various LLMs and push the boundaries of their understanding.

During our evaluation, we found that although the prompt did not explicitly require it, Gemini proactively numbered the different types of questions (Fig. S3–S5†), resulting in more organized responses. This approach suggests a higher level of understanding of structured data generation. To quantify this behavior, we introduced an alignment parameter (Table 1) to compare how well different LLMs adhered to formatting requirements. As shown in Table 1, GPT-4's responses (Fig. S6†) rarely followed the expected format, approximately half of Claude's responses were organized, and nearly all of Gemini's responses met the formatting criteria. This difference in adherence reflects the models' varying abilities to interpret implicit organizational cues. We believe that adherence to proper formatting is crucial for creating structured datasets in the future, as it significantly enhances ease of use, reduces the need for manual adjustments, and facilitates subsequent analysis.

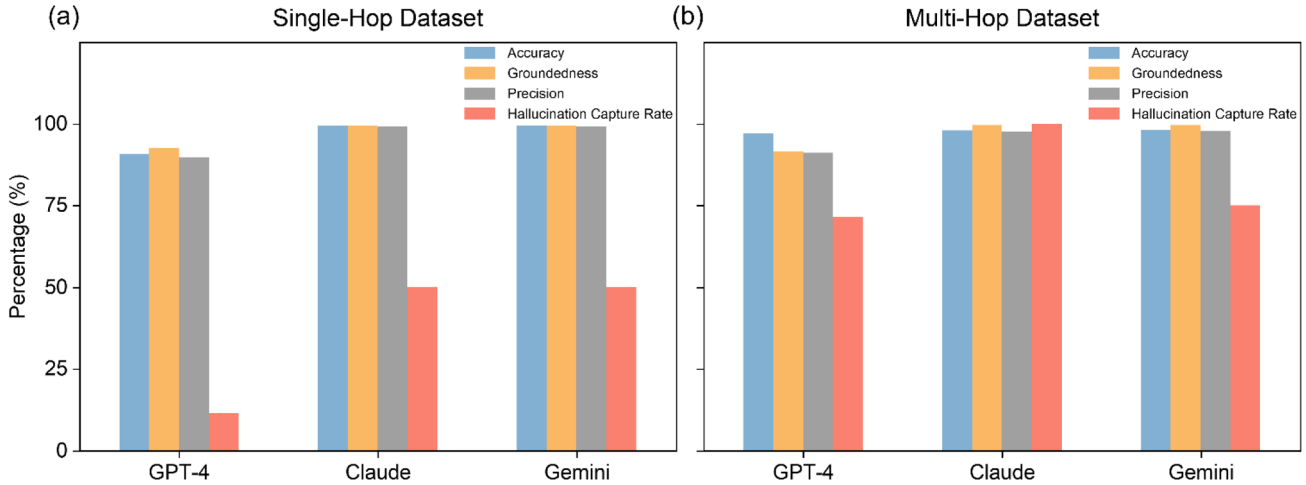


Fig. 4 Performance of LLMs in the (a) single-hop Q&A and (b) multi-hop Q&A generation task. Accuracy measures the correctness of answers; groundedness measures the quality of questions; precision measures overall performances; hallucination capture rate measures self-correction abilities.

Table 1 Alignment: number of DOIs meeting the format requirements in each dataset

	Single-hop Q&A	Multi-hop Q&A
GPT-4	3/50	2/50
Claude	29/50	20/50
Gemini	44/50	45/50

Overall, based on these findings, we conclude that Gemini offers several advantages in generating Q&A datasets: high accuracy, well-structured responses, and lower cost ($0.18 per DOI, Table S2†) compared to other models. Moreover, Gemini's ability to independently impose structure on its outputs suggests that it is particularly well-suited for applications where consistency and ease of integration are paramount.

Conclusions

In this study, we evaluated the capabilities of three LLMs—GPT-4, Claude, and Gemini—in reading MOF-related literature through two key tasks: synthesis condition extraction, and Q&A dataset generation. Claude performed exceptionally well in terms of completeness, effectively covering a wide range of synthesis details, while Gemini showed both outstanding correctness and completeness, making it the most reliable for accurate and complete data extraction. GPT-4, although less effective in quantitative metrics, exhibited strong logical reasoning and context inference abilities, as evidenced by its capacity to extrapolate synthesis conditions effectively. Our error analysis indicated common challenges across all LLMs, particularly in the interpretation of numerical data, highlighting the need for improvements to enhance data extraction consistency. In the Q&A dataset generation tasks, Claude and Gemini outperformed GPT-4 in accuracy and groundedness for both single-hop and multi-hop questions, making them strong candidates for use as benchmarks in future studies. Gemini's ability to impose a structured format on its responses further suggests its suitability for building organized datasets with minimal post-processing requirements.

Overall, Claude strikes a balance between accuracy and coverage for generating synthesis conditions dataset, while Gemini's completeness and structured approach make it the best choice for creating comprehensive Q&A datasets. These findings show that LLMs can help build scientific databases, but improvements in prompt design and preprocessing are needed to make them truly effective.

Data availability

The prompts for extracting synthesis conditions and generating Q&A datasets; the evaluation flowchart for each product in the synthesis condition dataset; examples of Gemini and GPT-4 responses in Q&A generating task; the number of the selected DOIs for each task; and the cost analysis are available in the ESI.† The generated datasets along with their human evaluations, and the associated processing scripts used in this paper are available at **https://doi.org/10.5281/zenodo.15376525**.

Author contributions

Y. S., N. R. and O. M. Y. conceived the idea and drafted the outline. Y. S., N. R. wrote the initial draft of the manuscript, including the design of the figures. Y. S. led the evaluation of all datasets. C. Z. assisted with the evaluation of multi-hop dataset. All authors contributed to the review and editing of the final manuscript.

Conflicts of interest

The authors declare no competing financial interest.

Acknowledgements

This research was supported by the King Abdulaziz City for Science and Technology (Center of Excellence for Nanomaterials and Clean Energy Applications, KACST), and the Bakar Institute of Digital Materials for the Planet (BIDMaP). Y. S. and N. R. thank A. M. Alabdulkarim at KACST for her suggestions on the draft of the manuscript. N. R. acknowledges the BIDMaP Emerging Scholars Program for the funding that supports this work. Y. S. acknowledges the assistance from Gemini in designing the TOC graphic.

References

1 O. M. Yaghi, M. J. Kalmutzki and C. S. Diercks, *Introduction to Reticular Chemistry: Metal-Organic Frameworks and Covalent Organic Frameworks*, Wiley-VCH, Weinheim, 2019, DOI: **10.1002/9783527821099**.

2 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.

3 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**, 491–501.

4 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, L. Gagliardi, A. Anandkumar, C. Borgs, J. T. Chayes and O. M. Yaghi, Shaping the Water-Harvesting Behavior of Metal–Organic Frameworks Aided by Fine-Tuned GPT Models, *J. Am. Chem. Soc.*, 2023, **145**(51), 28284–28295.

5 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**, 4705.

6 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**, 4958–4965.

7 V. T. Silva, A. Rademaker, K. Lionti, R. Giro, G. Lima, S. Fiorini, M. Archanjo, B. W. Carvalho, R. Neumann, A. Souza, J. P. Souza, G. Valnisio, C. N. Paz, R. Cerqueira and M. Steiner, Automated, LLM Enabled Extraction of Synthesis Details for Reticular Materials from Scientific Literature, *arXiv*, 2024, preprint, arXiv:2411.03484, DOI: **10.48550/arXiv.2411.03484**.

8 L. Shi, Z. Liu, Y. Yang, W. Wu, Y. Zhang, H. Zhang, J. Lin, S. Wu, Z. Chen, R. Li, N. Wang, Z. Liu, H. Tan, H. Gao, Y. Zhang and G. Wang, LLM-based MOFs Synthesis Condition Extraction using Few-Shot Demonstrations, *arXiv*, 2024, preprint, arXiv:2408.04665, DOI: **10.48550/arXiv.2408.04665**.

9 OpenAI, GPT-4 Technical Report, *arXiv*, 2023, preprint, arXiv:2303.08774, DOI: **10.48550/arXiv.2303.08774**.

10 *Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku*, **https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf**, accessed December 2, 2024.

11 Gemini Team and Google, Gemini 1.5: unlocking multimodal understanding across millions of tokens of context, *arXiv*, 2024, preprint, arXiv.2403.05530, DOI: **10.48550/arXiv.2403.05530**.

12 N. Rampal, K. Wang, M. Burigana, L. Hou, J. Al-Johani, A. Sackmann, H. S. Murayshid, W. A. AlSumari, A. M. AlAbdulkarim, N. E. Alhazmi, M. O. Alawad, C. Borgs, J. T. Chayes and O. M. Yaghi, Single and Multi-Hop Question-Answering Datasets for Reticular Chemistry with GPT-4-Turbo, *J. Chem. Theory Comput.*, 2024, **20**, 9128–9137.