JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

Single and Multi-Hop Question-Answering Datasets for Reticular Chemistry with GPT-4-Turbo

Nakul Rampal, Kaiyu Wang, Matthew Burigana, Lingxiang Hou, Juri Al-Johani, Anna Sackmann, Hanan S. Murayshid, Walaa A. AlSumari, Arwa M. AlAbdulkarim, Nahla E. Alhazmi, Majed O. Alawad, Christian Borgs,* Jennifer T. Chayes,* and Omar M. Yaghi*

Cite This: *J. Chem. Theory Comput.* 2024, 20, 9128–9137 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The rapid advancement in artificial intelligence and natural language processing has led to the development of large-scale datasets aimed at benchmarking the performance of machine learning models. Herein, we introduce "RetChemQA", a comprehensive benchmark dataset designed to evaluate the capabilities of such models in the domain of reticular chemistry. This dataset includes both single-hop and multi-hop question-answer pairs, encompassing approximately 45,000 question and answers (Q&As) for each type. The questions have been extracted from an extensive corpus of literature containing about 2,530 research papers from publishers including NAS, ACS, RSC, Elsevier, and Nature Publishing Group, among others. The dataset has been generated using OpenAI's GPT-4 Turbo, a cutting-edge model known for its exceptional language understanding and generation capabilities. In addition to the Q&A dataset, we also release a dataset of synthesis conditions extracted from the corpus of literature used in this study. The aim of RetChemQA is to provide a robust platform for the development and evaluation of advanced machine learning algorithms, particularly for the reticular chemistry community. The dataset is structured to reflect the complexities and nuances of real-world scientific discourse, thereby enabling nuanced performance assessments across a variety of tasks.

Given the increasing application of large language models (LLMs) across various scientific domains, including reticular chemistry,[1−5] the development of benchmark datasets for evaluating their performance is crucial. While benchmark datasets for many tasks across different subjects already exist—such as PubMedQA[6] for biomedical questions, HotPotQA[7] for complex question answering, and SQuAD[8,9] for reading comprehension—there is a noticeable lack of datasets for tasks specific to reticular chemistry. This study aims to bridge this gap by introducing a question and answer (Q&A) dataset, which we have named RetChemQA, tailored to the unique demands of reticular chemistry.[10,11] For the researchers working at the intersection of computer science and reticular chemistry, this dataset provides a standard against which new LLMs and methodologies can be benchmarked, allowing for the development of new algorithms and hardware; we believe this work also holds importance for those who have limited or no background in computer science. For reticular chemists working in the wet lab, the development of models/methodologies that provide more accurate predictions and hypotheses will in turn lead to a reduction in the time and resources required for empirical testing in the lab. Additionally, since the output provided by an LLM is strongly correlated with the nature of the prompt provided,[12] it would be useful to have a method where the prompt optimization is automated. This avoids the (i) inconsistency that might result from variation of a prompt seeking a particular output and (ii) subjective human judgment leading to arduous, time-consuming, and impractical approaches to evaluate LLM outputs and prompts for tasks that require thousands of repeating LLM calls.

The introduction of the RetChemQA dataset will enable the development and application of automated prompt optimization and evaluation frameworks like DSPy,[13] which can be useful when using LLMs for tasks such as refining experimental designs and optimizing synthesis conditions. Furthermore, building chat-based user interfaces for these prompt optimization frameworks will further lower the barriers to entry for scientists with limited or no knowledge of computer science.

In this contribution, we report RetChemQA, a dataset containing a total of about ∼90,000 Q&As automatically generated using GPT-4-Turbo from a corpus of about 2,530 papers including both the manuscript (MS) and Supporting Information (SI) where available. The dataset contains questions of 3 types: (i) *Factual questions*: these types of questions are those for which an answer is a stated fact, (ii) *Reasoning questions*: these types of questions require an understanding of the text to answer, and (iii) *True/False*

Received: June 20, 2024
Revised: September 15, 2024
Accepted: September 20, 2024
Published: October 8, 2024

https://doi.org/10.1021/acs.jctc.4c00805
J. Chem. Theory Comput. 2024, 20, 9128−9137

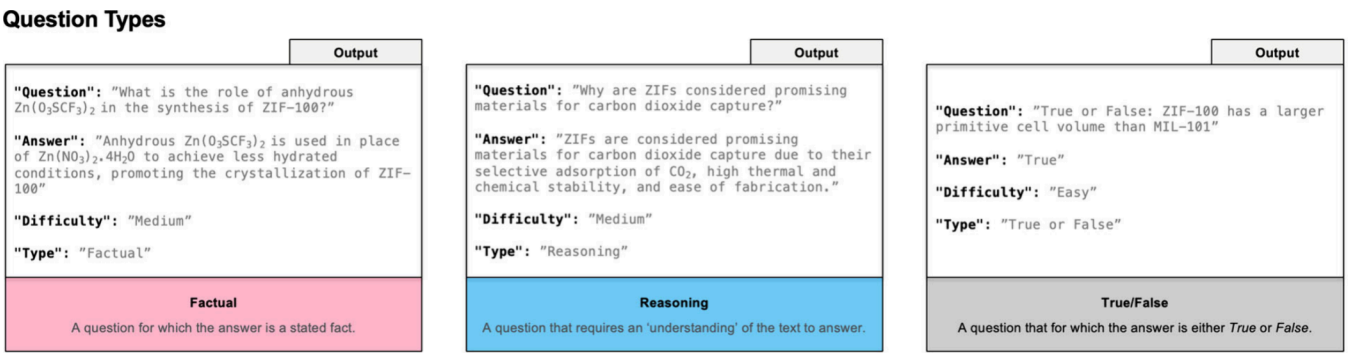

Figure 1. Type of questions in RetChemQA. The dataset consists of three main types of questions, from left to right, (i) Factual, (ii) Reasoning, and (iii) True/False. In the example shown, the questions have been generated using GPT-4-Turbo using the prompt shown in Figure 4, from ref 14.

questions: these are categorical questions that have a True/False answer. An example of each type of question is shown in Figure 1.

Moreover, the Q&A pairs generated are also categorized based on the difficulty levels: Easy, Medium, and Hard. Building on the categorization framework further, questions can also be classified on the number of reasoning steps required to answer them. Questions that require a single step of reasoning are termed as single-hop questions, and those that require multiple steps of reasoning are termed as multi-hop questions. When working with scientific literature, a single-hop question can often be answered by consulting only a single sentence provided in the MS or SI. On the other hand, a multi-hop question will often require information from multiple places in the MS and SI to answer. Examples of both the single-hop and multi-hop question types are shown in Figure 2. For the single-hop example question: "At what temperature range was the solvent-exchanged and evacuated ZIF-11 heated for gas-sorption analysis preparation?", we see that the answer generated is from a single contiguous piece of text taken from the MS, while for the multi-hop example question: "What temperature range was used for the solvothermal synthesis of ZIFs?", we see that the answer generated "The solvothermal synthesis was carried out at temperatures ranging from 85−150 °C" includes information from multiple parts of both the MS and SI of the paper. Interestingly, if the question were to be answered as a single-hop question, the question would have probably been answered incorrectly, as in the MS under the section "Typical ZIF" synthesis, the temperature given is "140 °C", while in the SI, where the individual synthesis conditions of each ZIF are provided, the temperature mentioned is different for each ZIF; so, any answer generated would have not included a range of temperatures as this information is not explicitly given anywhere in the paper.

METHODS

To build a corpus of text, we started with the CSD MOF Subset (April 2023)[16] that contains information on about 122,738 MOFs in 51,046 DOIs. Of the 122,738 MOFs present in the subset, we found that 8,089 MOF entries did not have an associated DOI—these MOF entries were removed. Next, we decided to consider only mainstream publishers: RSC, ACS, Wiley, Elsevier, AAAS, AIP, APS, Beilstein, CCS, De Gruyter, Frontiers, IOP, IUCr, NAS, Nature, Royal Society Publishing, T&F, and University Press (Oxford, Cambridge, Tsinghua). Full names for the acronyms are given in the SI in Table S1. After applying this criterion, we had 49,044 DOIs to work with. Finally, we further narrowed our corpus of text by working with only specific journals for each publisher (For more details, please see Table S2). In total, 2,530 DOIs were processed: Nature (220), RSC (215), Elsevier (82), ACS (1,283), AAAS (46), Wiley (653), NAS (10), CCS (10), AIP (5), and APS (6). To minimize any bias in the selection of the DOIs, they are randomly selected. It is important to clarify what the dataset might be biased against. While biasing the dataset against the information provided in the literature can be an issue, it does not impede the application or use of the dataset. If the dataset is intended to serve as a benchmark, the primary concern is ensuring that the question-answer pairs generated for a given DOI are correct. However, if the goal is to build a comprehensive database of Q&A pairs, such as a question bank for reticular chemistry, understanding potential biases becomes more crucial, in addition to ensuring the correctness of the question-answer pairs. Ideally, the goal here would then be to create a question bank that is representative of all the available literature on reticular chemistry. For each publisher, all the text and data mining were performed keeping in mind the contractual agreements the University of California, Berkeley, Library has with the individual publishers. To generate the Q&A + synthesis conditions datasets, the latest model from OpenAI GPT-4-Turbo (*gpt-4-0125-preview*) was used. In total, 337,577,236 tokens were processed at a cost of $3,600 for the whole project (including development and testing). The cost of generating a dataset (Q&A or synthesis conditions) is ∼$1,000; this translates to a cost of $0.40 per DOI.

We started with generating the single-hop Q&A dataset according to the workflow shown in Figure 3. To begin with, the processing environment is initialized. Next, for each *document_dir* given in each *publisher_dir*, the files are parsed, and the combined text is then tokenized and passed to the LLM for processing. A more detailed description of the data processing workflow algorithm is given in Figure S2. In the prompt provided, we explicitly specified that (i) the total of number of Q&As we want, in this case 20, and (ii) the number of different question types we want: 6 Factual, 7 True/False, and 7 Reasoning. We also mentioned the labels we want to include in the dataset: the question, the answer, the difficulty level, and the type of question. Here, a deliberate attempt was made to strike the right balance in that the prompt needed to be sufficiently open-ended to encourage creativity yet sufficiently structured to provide clear direction. To generate

https://doi.org/10.1021/acs.jctc.4c00805
J. Chem. Theory Comput. 2024, 20, 9128−9137

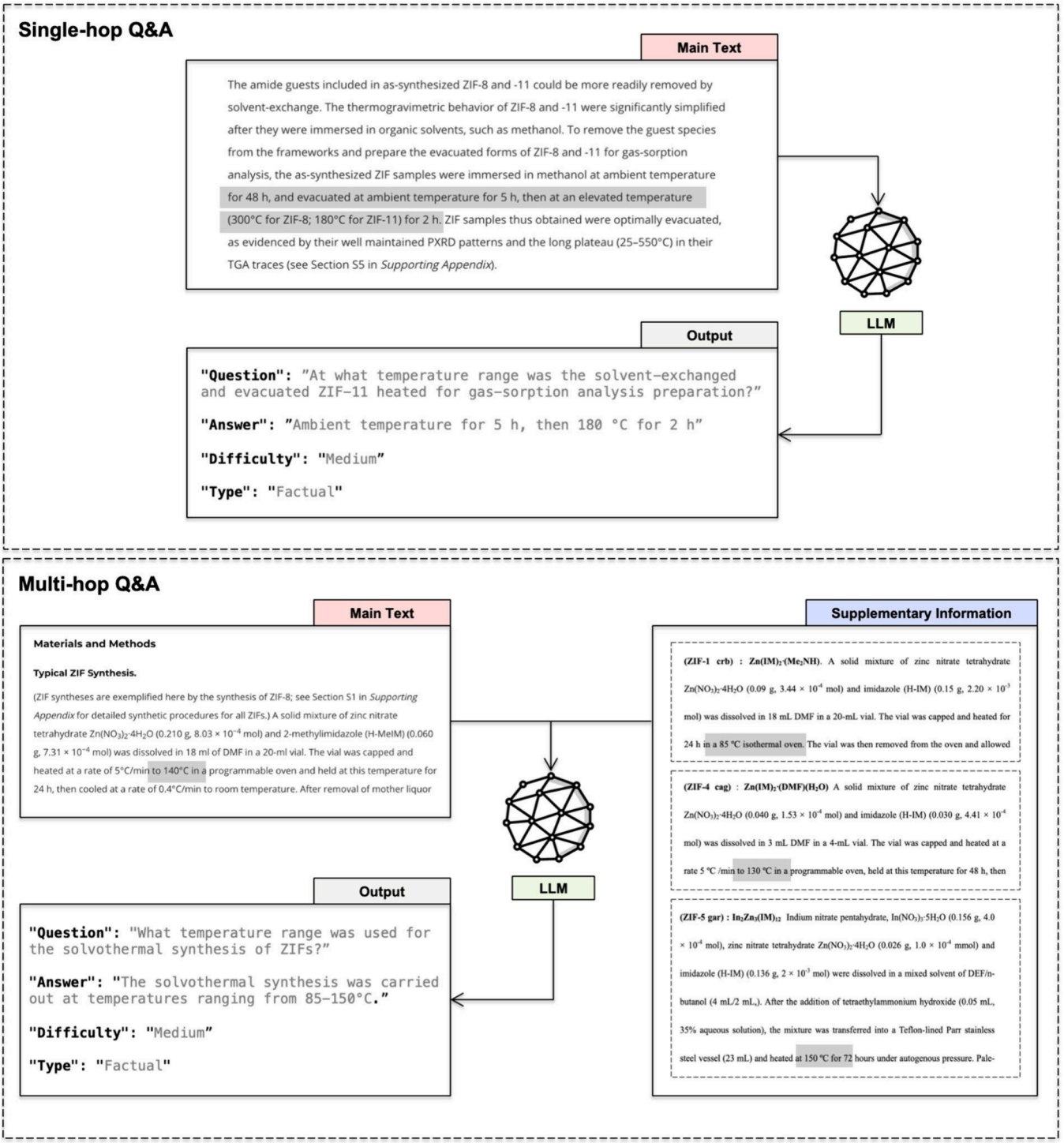

Figure 2. Single-hop vs multi-hop Q&A. Flow diagram of the information in a single-hop (top) and multi-hop (bottom) Q&A generation task. A single-hop Q&A is defined as one that requires only a single step of reasoning to answer; often, this involves retrieving information from a single sentence of a given paper. A multi-hop Q&A is defined as one that requires multiple steps of reasoning to answer; often, this involves retrieving information from multiple different parts of a MS. In the example shown, data must be collected from both the MS and SI to answer the question. In the example shown above, the questions have been generated using GPT-4-Turbo (gpt-4-0125-preview) using (i) the prompt shown in Figure 4 (left) for the single hop Q&A and (ii) the prompt shown in Figure 4 (right) for the multi-hop Q&A, from ref 15.

the multi-hop Q&A dataset, we initially did a simple modification to the prompt used to generate the single-hop Q&A dataset. We replaced the word "single-hop" with "multi-hop" in the whole prompt. Interestingly, this did not significantly change the output generated. There were many instances where both the number and type of single-hop and multi-hop questions generated for each DOI were very similar. Our goal therefore was then to develop a prompt that would diversify the types of questions generated. By providing additional context and including details like "A multi-hop Q&A is one that requires multi-step reasoning to come to an answer (this information can come from any part of the paper,

https://doi.org/10.1021/acs.jctc.4c00805
J. Chem. Theory Comput. 2024, 20, 9128−9137

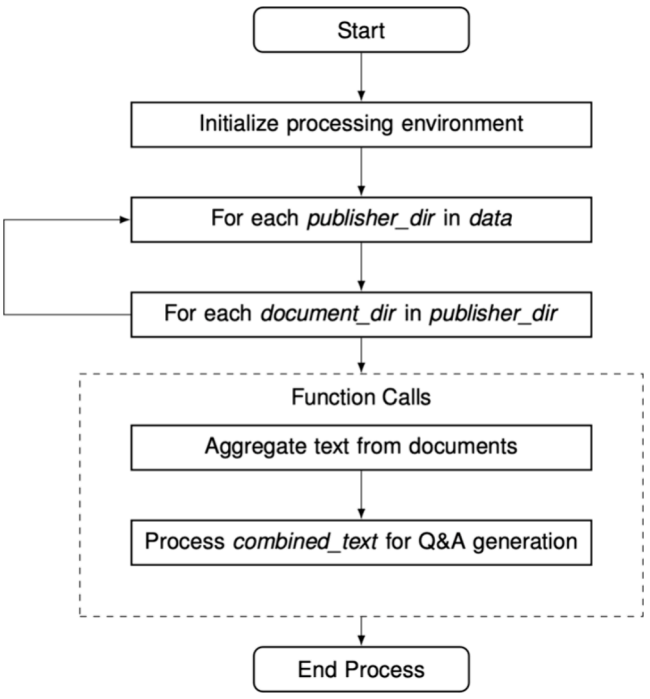

Figure 3. Dataset generation workflow. The figure depicts the dataset generation workflow with arrows indicating the order in which the steps are performed. For each *publisher_dir* (Nature, RSC, Elsevier, ACS, AAAS, Wiley, NAS, CCS, AIP, and APS) and every *document_dir* in a given *publisher_dir*, the following function calls are made: Aggregation of text from the documents (both the MS and SI, where available) followed by processing of the *combined_text* using the LLM, in this case GPT-4-Turbo, to generate the Q&A pairs. A more detailed dataset generation workflow/algorithm is given in Figure S2.

both MS and SI). To give you more details: A multi-hop Q&A will always involve going through multiple parts of the paper to come to an answer. This may include different paragraphs, different pages, and also different documents (i.e., the MS and SI)" in the prompt, we were able to significantly reduce but not completely eliminate the similar question types. The final prompts used to generate the single-hop and multi-hop Q&A datasets are shown in Figure 4. The generation of synthesis conditions is far from trivial, given that each paper uses different variables and formats for presenting information. This variability makes it difficult to establish a fixed set of variables to provide to the LLM. Moreover, each paper may contain synthesis conditions for more than one material, with the maximum number of materials being unknown. Given these uncertainties, devising the right prompt was challenging. Initial attempts were made to keep the prompt as open-ended as possible, allowing the LLM to decide on the number and type of variables that were deemed necessary. However, without sampling all the text provided in the dataset, it would be impossible to identify all the variables. Keeping the prompt open-ended resulted in the extraction of a lot of unnecessary information, including experimental characterization data such as crystal structure data and NMR peaks, which is often included under the "synthesis conditions" section of a paper. To minimize this, we explicitly added a statement to the prompt instructing the LLM not to extract any experimental characterization data. Although this approach significantly improved the nature of the outputs generated, it could not completely eliminate the extraction of experimental characterization data. It is important to note that, for DOIs associated with Wiley, we could process only the MS and not the SI, as we could not automate the downloading of the SI files. The final prompt used to generate the synthesis conditions dataset is shown in Figure S1.

For the multi-hop dataset, the data processing failed for 56 DOIs while for the single-hop dataset, the data processing failed for 34 DOIs. For the single-hop dataset, we came across an example (DOI: 10.1021/ja3073512) where the .json file mentioned "Add more questions as needed"—this was classified as an incomplete/failed generation. For some, the output generated was not in the format of a .json file, and such outcomes were counted as a failed DOI also. A summary of the errors/outputs generated for each of the DOIs for both the multi-hop and single-hop datasets is given in the supplementary files. For the synthesis conditions dataset, the data processing failed for 62 DOIs. A summary of the errors and the outputs generated for these 62 DOIs is given in the supplementary files.

In addition, for both the Q&A (single-hop and multi-hop) and synthesis conditions datasets, we have omitted the "context" label that is often included to avoid any copyright issues or concerns with the publishers. We would recommend the readers to use the entire text of a particular DOI including SI, where available, as the "context" for a given Q&A. Importantly, each file is named as follows: [DOI]_single-hop.json or [DOI]_multi-hop.json or [DOI]_synthesis-conditions.json; this should make working with the dataset easier as each file will only contain information specific to the DOI specified in the prefix of the filename.

Existing Q&A dataset evaluation criteria such as accuracy, precision, etc. are based on the premise that the question for which the answer is being evaluated is itself correct. This may not always hold true when the set of question-answer pairs is generated using an LLM. It is important to keep in mind that LLMs may also "hallucinate questions"—generate a question-answer pair from information *not* provided as "context" in the prompt to the LLM—and therefore, the answer to that question may also be incorrect. Hence, it was required that we come up with an evaluation metric that takes into account such outcomes. In addition, existing evaluation frameworks in literature do not generalize well across different question types. For example, for a Factual/Reasoning question there is no "negative" answer, and therefore, classifying an answer type as "False Negative (FN)" is not possible. This required the development of a new evaluation framework that, although similar to the framework used previously in literature, is tuned for our particular Q&A generation approach.

In the evaluation approach considered in this paper, the question-answer pair is first assessed based on whether it has been generated from the context provided in the prompt or not. If the question has been generated from the context provided, we next evaluate whether the answer to the question is correct or not. If the answer to the question is correct, the question-answer pair is classified as "True Positive (TP)"; else, the question answer pair is classified as "False Positive (FP)". On the other hand, if the question-answer pair generated is *not* from the context provided to the LLM and the answer to the question generated is correct, for example, the answer to a hallucinated question is "I cannot answer this question from the information provided in the context/prompt"—the LLM has itself identified that this is an hallucinated question; it is

https://doi.org/10.1021/acs.jctc.4c00805
J. Chem. Theory Comput. 2024, 20, 9128−9137

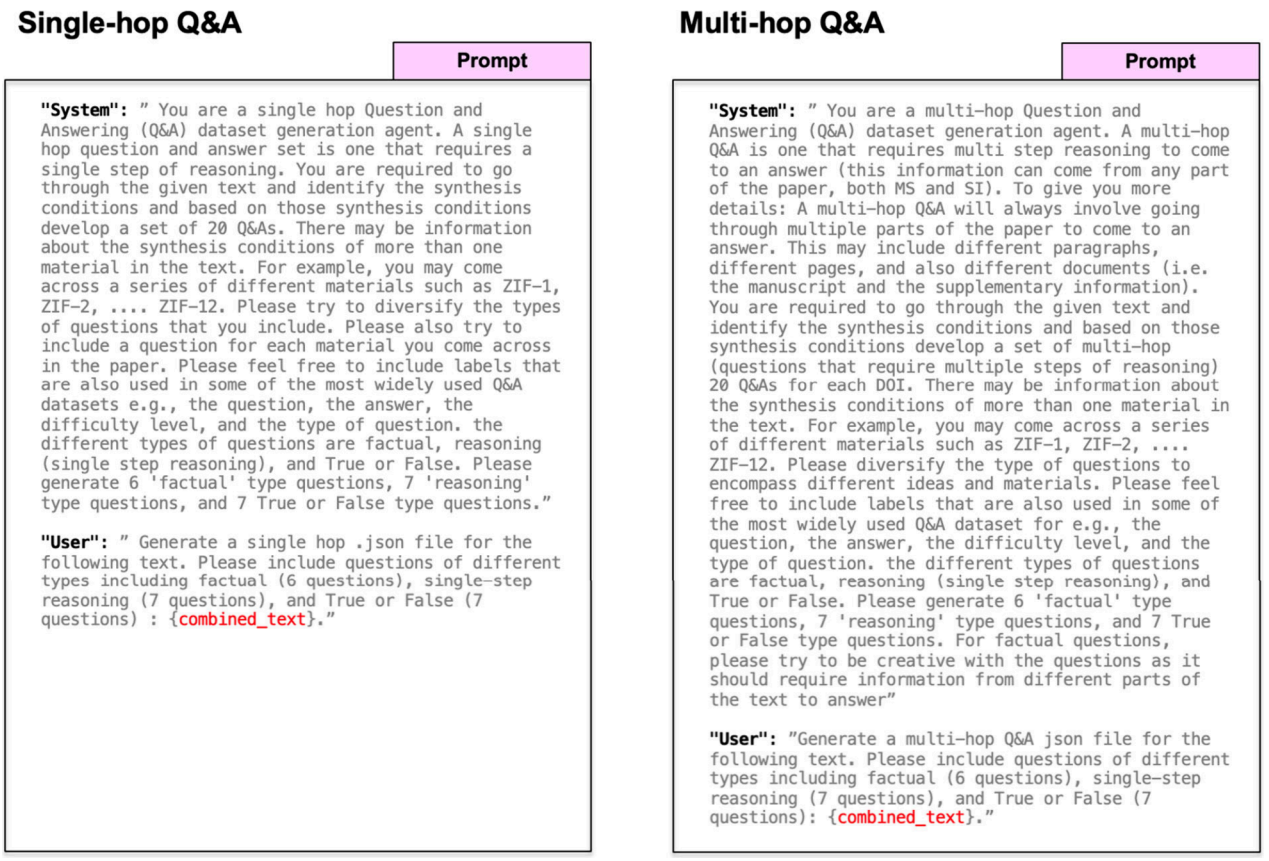

Figure 4. Prompts used to generate the set of Q&As. The prompt used to generate the single-hop Q&As is shown on the left, and the prompt used to generate the multi-hop Q&As is shown on the right. Each prompt consists of messages that are adopted to specific "roles" to guide the model's response. The "system" role provides the high-level instructions, while the "user" role provides the query. The "combined_text" variable holds all the text information contained in the MS and SI (where available). This information is provided as part of the prompt to GPT-4-Turbo.

classified as "True Negative (TN)"; else, the question-answer pair is classified as "FN". The evaluation framework described above is also shown as a flowchart in Figure S3.

Moreover, the evaluation framework can also handle question-answer pairs that have been incorrectly classified by the LLM: for example, a Reasoning/Factual question has been classified as a True/False question; the question-answer pair is classified as "out of context", allowing us to penalize the LLM for the incorrect categorization of the question-answer pair. We introduce a similar penalty if the LLM generates a single-hop question-answer pair when in the prompt provided it is explicitly stated to generate a multi-hop question-answer pair. This evaluation framework is broadly applicable to all the different question types considered in this dataset and therefore allows for comparison of the performance of the LLM in the generation of the different question types. Examples of question-answer pairs classified as TP, FP, TN, and FN in the single-hop and multi-hop datasets are shown in Figures S4 and S5. Following the classification of each Q&A pair, the performance of the LLM is assessed based on the following metrics and is specific to the evaluation framework described above:

(1) Accuracy: This is a measure of the ability of the LLM to correctly answer questions that have been generated both in or out of context—here, a penalty is introduced for answers that are *incomplete* or wrong, whether the question is in or out of context. It is defined as the ratio of the sum of the correctly answered questions, (TP + TN) to the total number of possible outcomes (TP + TN + FP + FN). A high accuracy score indicates better performance while a low accuracy indicates otherwise.

(2) Precision: This is a measure of the ability of the LLM to accurately answer questions that have been generated only in context—In addition to the penalties introduced above, here, a penalty is also introduced for (i) hallucinated questions even if answered correctly, (ii) incorrectly generated questions, and (iii) incorrectly categorized questions. It is defined as the ratio of accurately answered in context questions (TP) to the total number of possible outcomes (TP + FN + FP + FN). A high precision score is desired as it indicates better performance; a low precision score indicates otherwise.

(3) Hallucination Rate: This is a measure of proportion of Q&A pairs hallucinated by the LLM. It is defined as the ratio of the sum of hallucinated Q&A pairs (TN + FN) to the total number of possible outcomes (TP + TN + FP + FN). A low hallucination rate indicates better performance, while a high hallucination rate indicates otherwise.

(4) Hallucination Capture Rate: This is a measure of the LLM's ability to identify and correct a hallucinated (out-of-context) question it has generated itself. It is defined as the ratio of hallucinated questions generated but answered correctly (TN) to the total number of hallucinated questions generated (TN + FN). A high

https://doi.org/10.1021/acs.jctc.4c00805
J. Chem. Theory Comput. 2024, 20, 9128−9137

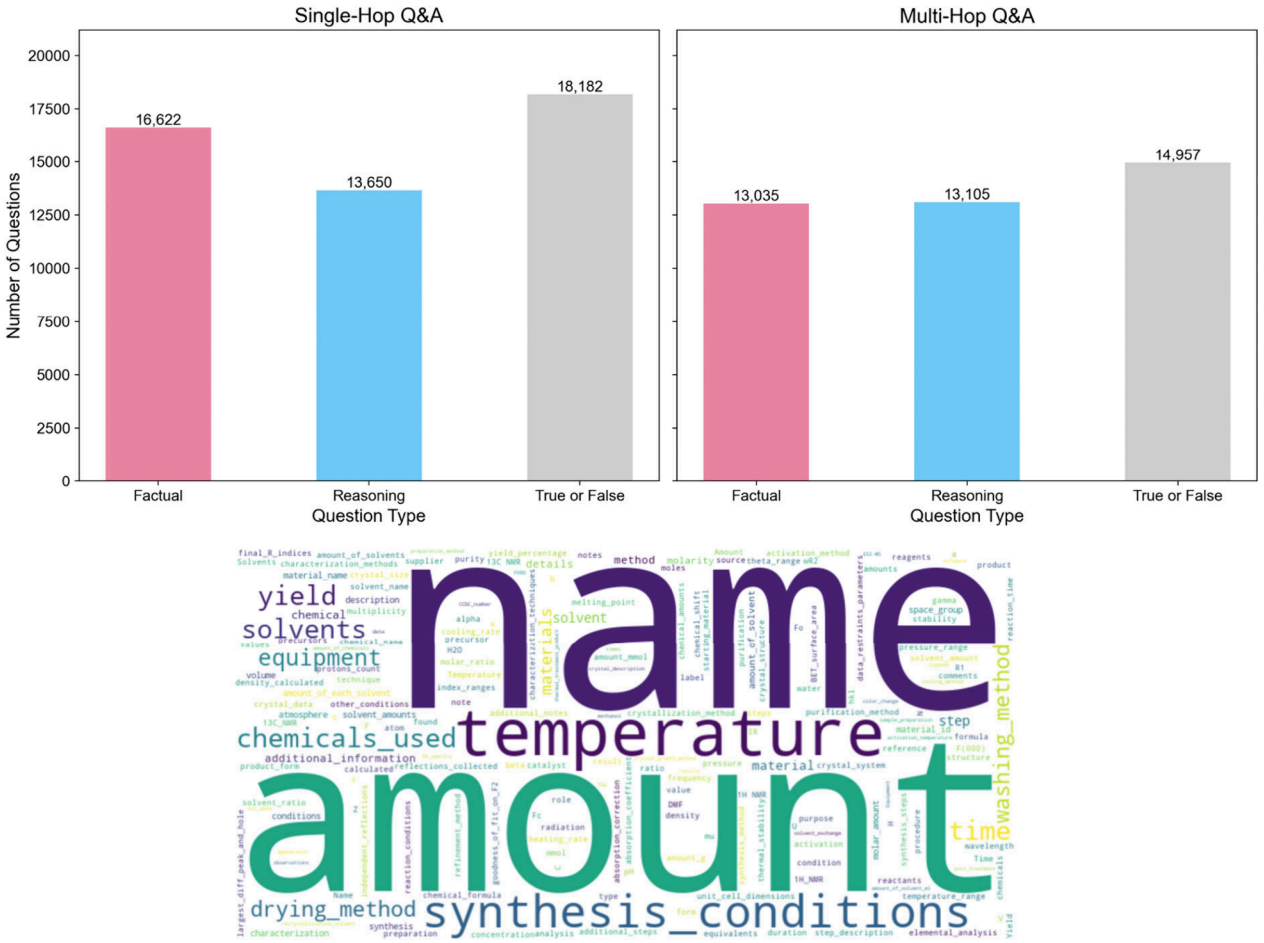

Figure 5. Results for the RetChemQA dataset. Total number of questions generated (and the type: Factual, Reasoning, and True/False) in the single-hop and multi-hop datasets (top), and a word cloud showing the different synthesis related keywords generated in the synthesis conditions dataset (bottom). The size of a word indicates its frequency or importance in the text being analyzed, where larger words represent those that appear more frequently and smaller words represent those that appear less frequently.

Table 1. Average Number of Different Question Types (Factual, True/False, Reasoning) and Difficulty Levels (Easy, Medium, or Hard) Generated per DOI[a]

	Factual			True/False			Reasoning			
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	Total
Single-hop	2	4	1	4	2	1	0	3	3	19
Multi-hop	1	3	1	4	2	0	0	2	3	17

[a]For the single-hop Q&A dataset, a total of 48,454 question-answer pairs were successfully generated from 2,496 DOIs. For the multi-hop Q&A dataset, a total of 41,097 question-answer pairs were successfully generated from 2,474 DOIs. Each value in the table has been rounded to the nearest whole number.

hallucination capture rate is desired as it means that the LLM is able to identify its mistake, while a low hallucination capture rate indicates otherwise.

As discussed previously, the evaluation of the synthesis conditions dataset is far less straightforward. This is because, for each DOI, the number of variables and the format in which the synthesis conditions are reported is different. In addition, no single set of synthesis conditions follows a standard format. This complicates the development of a performance assessment metric that encompasses all of the different synthesis conditions. Therefore, we propose a binary Yes (Y) or No (N) assessment metric. If for a particular material the entire set of the synthesis conditions extracted is correct, we assign a "**Y**"; if it is incorrect or incomplete, we assign an "**N**"; this is criterion 1. If the conditions *do not* include experimental characterization data, we assign an "**N**"; otherwise, we assign it a "**Y**"—this is criterion 2. To assess the performance of the LLM in extracting synthesis conditions, we then calculate the ratio of the number of **Y**s generated to the total number of outcomes (no. of **N**s + no. of **Y**s) for criterion 1 and the ratio of the number of **N**s generated to the total number of outcomes (no. of **N**s + no. of **Y**s) for criterion 2. A product of the two ratios then allows us to determine the performance of the LLM in extracting the synthesis conditions—we term this the "Obedience" score. This is because a high "Obedience" score would imply that the LLM is strictly "obeying" the two main instructions provided in the prompt: (i) to extract all the synthesis conditions for a given a material and (ii) to not

https://doi.org/10.1021/acs.jctc.4c00805
J. Chem. Theory Comput. 2024, 20, 9128−9137

include any experimental characterization data. On the other hand, a low "Obedience" score would mean that the LLM fails to adequately follow the instructions provided in the prompt.

RESULTS AND DISCUSSION

In total, 89,551 question-answer pairs were generated with 54% (48,454) being single-hop Q&As and 46% (41,097), multi-hop Q&As. For both the single-hop (2,496 DOIs) and multi-hop (2,474 DOIs) datasets, an approximately equal number of DOIs were processed. The distribution of the different question types in both the single-hop and multi-hop datasets is shown in Figure 5. For the single-hop Q&A dataset, for each DOI, ∼19 questions are generated of which on average 7 are Factual, 7 are True/False, and 6 are Reasoning. On the other hand, for the multi-hop Q&A dataset, for each DOI, ∼17 questions are generated of which on average 5 are Factual, 6 are True/False, and 5 are Reasoning. A detailed summary of the distribution of the different question types along with their difficulty levels is given in Table 1.

It is interesting to note here that, in the prompt provided to the LLM, for both the single-hop and multi-hop dataset generation tasks, the total number of questions specified to be generated (20 Q&As), along with the distribution of the question types to be generated, is the same, i.e., 6 Factual, 7 True/False, and 7 Reasoning. For the single-hop dataset generation, this instruction was followed more closely by the LLM as compared to that for the multi-hop dataset generation task. We believe this is due to the complexity arising from the multistep reasoning required to generate the questions for the multi-hop dataset. This compromise in performance is also seen in the ability of the LLM to generate answers for the multi-hop dataset. For the multi-hop dataset, the precision—a measure of the proportion of Q&A pairs generated that are from the context provided and correctly answered—is lower as compared to the single-hop dataset. In addition, the hallucination rate—a measure of the proportion of Q&A pairs generated out of context—is higher for the multi-hop dataset as compared to the single-hop dataset. On the other hand, the accuracy—a measure of proportion of Q&As generated that are (i) from the context provided and correctly answered and (ii) if out of context (i.e., hallucinated) then correctly identified as such—is higher for the multi-hop dataset as compared to the single-hop dataset; naturally, the question then is why? This is because the LLM, while hallucinating more when generating the multi-hop questions, is also more "cautious" and, hence, is able to correctly identify and correct course when generating the answers; i.e., the hallucination capture rate is much higher for the multi-hop dataset (84%) than the single-hop dataset (22%). Our hypothesis here is that, since the multi-hop Q&A generation task is more complicated, the LLM approaches it with more caution and is more careful when generating the answers. This hypothesis is further corroborated by the latency—a measure of the time required from the initiation of the request to the completion of the response—which is found to be higher for the multi-hop dataset than the single-hop dataset for the same input provided. While the performance of the LLM in the Q&A generation task is impressive, the quality of the Q&A pairs generated is more impressive. The types of Q&A pairs generated are often of the quality as those asked by graduate students or even senior chemists. For example, the question: "What experimental techniques were used to characterize the spin transition mechanism in Fe2Cl2(bbta)?" generated for ref 17 requires one to read the entire text to understand what methods were used to characterize the spin transition mechanism. The paper discusses multiple methods, including gas adsorption measurements, Powder X-ray diffraction, differential scanning calorimetry, measurements of dc magnetic susceptibility, Mössbauer spectroscopy, and infrared spectroscopy. The answer generated by the LLM: "Powder X-ray diffraction, infrared spectroscopy, Mössbauer spectroscopy, and measurements of dc magnetic susceptibility were used to characterize the spin transition mechanism" only includes the correct 4 out of the 6 methods discussed in the paper.

A summary of the performance assessment of the single-hop and multi-hop Q&A datasets is given in Table 2, while a more detailed performance assessment by question type is shown in Figure 6. For the generation of the synthesis conditions task, the ratio of the number of **Y**s to the total number of outcomes (no. of **N**s + no. of **Y**s) was determined to be 0.794 for criterion 1, indicating that, ∼80% of the time, the LLM was able to correctly extract all the synthesis conditions for a given material. The ratio of the number of **N**s to the total number of outcomes (no. of **N**s + no. of **Y**s) for criterion 2 was determined to be 0.893, indicating that ∼90% of the time, the LLM did not extract details related to the experimental characterization. The obedience score was determined to be 0.708, indicating that ∼70% of the time the LLM followed both the instructions provided in the prompt. In total, the synthesis conditions generated for 238 DOIs were checked manually. The evaluations for all the datasets (including the single-hop and multi-hop datasets) are given in the Supporting Information.

Table 2. Performance Assessment of the Single-Hop and Multi-Hop Q&A Datasets[a]

	Accuracy	Precision	Hallucination Rate	Hallucination Capture Rate
Single-hop	0.948	**0.943**	**0.028**	0.217
Multi-hop	**0.983**	0.934	0.055	**0.841**

[a]The datasets are evaluated based on the following metrics: Accuracy = (TP+TN)/(TP+TN+FP+FN); Precision = TP/(TP+TN+FP+FN); Hallucination Rate = (TN+FN)/(TP+TN+FP+FN), and Hallucination Capture Rate = (TN)/(TN+FN). TP = True Positive; TN = True Negative; FP = False Positive; and FN = False Negative. For Accuracy, Precision, and Hallucination Capture Rate, higher values indicate better performance (1 indicates perfect performance), while for the hallucination rate, lower values indicate better performance (0 indicates perfect performance). For the single-hop dataset, a total of 265 DOIs were evaluated, while for the multi-hop dataset, a total of 233 DOIs were evaluated.

CONCLUDING REMARKS

In this study, we present a question-answering dataset specific to reticular chemistry. The dataset is generated automatically using an LLM, in this case GPT-4-Turbo. The LLM performs exceptionally well in generating both single-hop and multi-hop question-answer pairs with precision values of ∼94%, indicating that the majority of the time the Q&A pair generated is from the context provided (and not hallucinated). Interestingly, we find that, when the task at hand is more complex, for example, the task of generating a multi-hop Q&A pair, although the LLM hallucinates more, it is also more "careful" evaluating the answers it generates, and therefore, we

https://doi.org/10.1021/acs.jctc.4c00805
J. Chem. Theory Comput. 2024, 20, 9128−9137

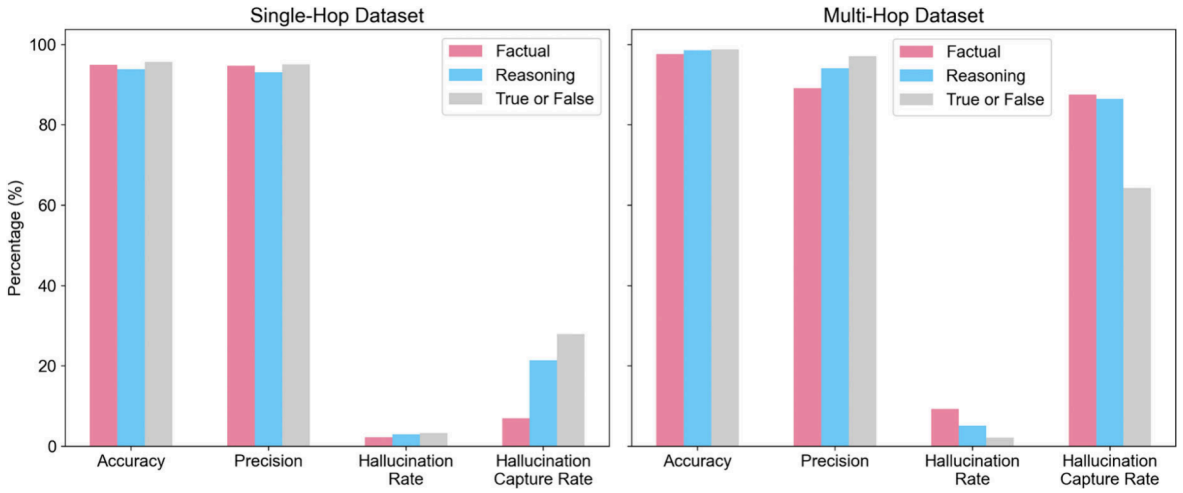

Figure 6. Performance assessment by question type for both the single-hop and multi-hop Q&A datasets. A comparison of the performance of the LLM as evaluated based on the accuracy, precision, hallucination rate, and hallucination capture rate for the single-hop dataset generation task (left) and the multi-hop dataset generation task (right).

hypothesize is able to rectify its mistake, as indicated by an exceptionally high hallucination capture rate of 84% for the multi-hop Q&A dataset. This eventually gave us a higher accuracy score for the multi-hop Q&A dataset. The Q&A pairs generated are of stunning quality often matching in quality with those expected from graduate students or even senior chemists. While these results are impressive, we acknowledge that general strategies such as better prompt engineering or model fine-tuning could further improve the performance of the LLMs. Our experience with prompt engineering has demonstrated that manually identifying the "right" prompt for consistently better performance across different DOIs is challenging. One of purposes of building this benchmark dataset is to facilitate the development of automated prompt optimization frameworks like DSPy, which can streamline the high-throughput evaluation of prompt performance across various DOIs. Even for fine-tuning purposes, the need for comprehensive templates that include the prompt, context, and expected output is crucial. RetChemQA serves this need by providing the benchmark dataset for each DOI, effectively filling in the "expected output" entry in templates that are required for such optimizations.

The extraction and classification of synthesis conditions is far more complex given that the format of the synthesis conditions reported is different for each DOI. There is not a "fixed" template or a given set of variables to follow for the LLM when extracting the synthesis conditions. The LLM exhibits an obedience score of ∼70%, indicating that approximately in 7 out of 10 instances the LLM is able to both extract all the given set of synthesis conditions and also make sure that no experimental characterization data is included. While improving the performance of the LLM by developing more accurate models tuned for data extraction tasks is one way to address the low obedience score, another way would be to address the format of the data funnel that enters the LLM. In this case, we propose the development of a synthesis condition information file (.sif) similar to its counterpart, the crystallographic information file (.cif), which standardizes the reporting of synthesis conditions. We envision that the RetChemQA dataset will (i) catalyze the development of large language models built for reticular chemistry and (ii) help democratize access to LLMs by enabling the development and application of automated prompt optimization frameworks, leading to an improvement in the reliability and accuracy of the outcomes generated by an LLM in the domain of reticular chemistry.

■ ASSOCIATED CONTENT

Data Availability Statement

All supplementary files can be accessed in the following GitHub repository: https://github.com/nakulrampal/RetChemQA.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c00805.

Prompt used to extract the synthesis conditions from a paper; algorithm used to run the dataset generation; evaluation criteria for each Q&A in the single-hop and multi-hop datasets; examples of questions classified as TP, FP, TN, and FN in the single-hop and multi-hop datasets; acronyms defined; journal summary (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Christian Borgs − *Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society and Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720, United States*; Email: borgs@berkeley.edu

Jennifer T. Chayes − *Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, Department of Electrical Engineering and Computer Sciences, Department of Mathematics, Department of Statistics, and School of Information, University of California, Berkeley, California 94720, United States*; Email: jchayes@berkeley.edu

Omar M. Yaghi − *Department of Chemistry, University of California, Berkeley, California 94720, United States; Kavli Energy Nanoscience Institute and Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, University of California, Berkeley, California 94720, United States; KACST−UC Berkeley Center of Excellence for Nanomaterials for Clean Energy Applications,*

https://doi.org/10.1021/acs.jctc.4c00805
J. Chem. Theory Comput. 2024, 20, 9128−9137

King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia; orcid.org/0000-0002-5611-3325; Email: yaghi@berkeley.edu

Authors

Nakul Rampal − *Department of Chemistry, University of California, Berkeley, California 94720, United States; Kavli Energy Nanoscience Institute and Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society, University of California, Berkeley, California 94720, United States;* orcid.org/0000-0002-6187-5631

Kaiyu Wang − *Department of Chemistry, University of California, Berkeley, California 94720, United States; Kavli Energy Nanoscience Institute, University of California, Berkeley, California 94720, United States*

Matthew Burigana − *Department of Chemistry, University of California, Berkeley, California 94720, United States; Kavli Energy Nanoscience Institute, University of California, Berkeley, California 94720, United States*

Lingxiang Hou − *Department of Chemistry, University of California, Berkeley, California 94720, United States; Kavli Energy Nanoscience Institute, University of California, Berkeley, California 94720, United States*

Juri Al-Johani − *Bakar Institute of Digital Materials for the Planet, College of Computing, Data Science, and Society and Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720, United States*

Anna Sackmann − *Data Services Librarian, University of California, Berkeley, California 94720, United States*

Hanan S. Murayshid − *Artificial Intelligence & Robotics Institute, King Abdulaziz City for Science and Technology (KACST), Riyadh 11442, Saudi Arabia*

Walaa A. AlSumari − *Artificial Intelligence & Robotics Institute, King Abdulaziz City for Science and Technology (KACST), Riyadh 11442, Saudi Arabia*

Arwa M. AlAbdulkarim − *Artificial Intelligence & Robotics Institute, King Abdulaziz City for Science and Technology (KACST), Riyadh 11442, Saudi Arabia*

Nahla E. Alhazmi − *Hydrogen Technologies Institute, King Abdulaziz City for Science and Technology (KACST), Riyadh 11442, Saudi Arabia*

Majed O. Alawad − *KACST−UC Berkeley Center of Excellence for Nanomaterials for Clean Energy Applications, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c00805

Author Contributions

N.R., C.B., J.T.C., and O.M.Y. conceived the idea and drafted the outline. N.R. wrote the initial draft of the manuscript, including the design of the figures. K.W. led the evaluation of the multi-hop dataset. M.B. together with J.A.-J. led the evaluation of the single-hop dataset. L.H. led the evaluation of the synthesis conditions dataset. A.S. made sure that the data scraping followed the guidelines and permissions as outlined in the agreement between the UC Berkeley Library and the publishers. All authors contributed to the review and editing of the final manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

N.R. acknowledges the Bakar Institute of Digital Materials for the Planet (BIDMaP) Emerging Scholars Program for the funding that supports this work.

REFERENCES

(1) Shi, L.; Liu, Z.; Yang, Y.; Wu, W.; Zhang, Y.; Zhang, H.; Lin, J.; Wu, S.; Chen, Z.; Li, R.; Wang, N.; Liu, Z.; Tan, H.; Gao, H.; Zhang, Y.; Wang, G. LLM-Based MOFs Synthesis Condition Extraction Using Few-Shot Demonstrations. *arXiv* **2024**, 2408.04665.

(2) Zheng, Z.; Rong, Z.; Rampal, N.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. A GPT-4 Reticular Chemist for Guiding MOF Discovery. *Angew. Chem., Int. Ed.* **2023**, *62* (46), No. e202311983.

(3) Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *J. Am. Chem. Soc.* **2023**, *145* (32), 18048−18062.

(4) Zheng, Z.; Alawadhi, A. H.; Chheda, S.; Neumann, S. E.; Rampal, N.; Liu, S.; Nguyen, H. L.; Lin, Y. H.; Rong, Z.; Siepmann, J. I.; Gagliardi, L.; Anandkumar, A.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. 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) Zheng, Z.; Zhang, O.; Nguyen, H. L.; Rampal, N.; Alawadhi, A. H.; Rong, Z.; Head-Gordon, T.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs. *ACS Cent Sci.* **2023**, *9* (11), 2161−2170.

(6) Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* **2019**, 2567−2577.

(7) Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; Manning, C. D. HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* **2018**, 2369−2380.

(8) Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* **2016**, 2383−2392.

(9) Rajpurkar, P.; Jia, R.; Liang, P. Know What You Don't Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* **2018**, *2*, 784−789.

(10) Yaghi, O. M. Reticular Chemistry in All Dimensions. *ACS Cent Sci.* **2019**, *5* (8), 1295−1300.

(11) Yaghi, O. M. Reticular Chemistry—Construction, Properties, and Precision Reactions of Frameworks. *J. Am. Chem. Soc.* **2016**, *138* (48), 15507−15509.

(12) Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.-Y.; Wen, J.-R. A Survey of Large Language Models. *arXiv* **2023**, 2303.18223.

(13) Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; Miller, H.; Zaharia, M.; Potts, C. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *arXiv* **2023**, 2310.03714.

(14) Wang, B.; Côté, A. P.; Furukawa, H.; O'Keeffe, M.; Yaghi, O. M. Colossal Cages in Zeolitic Imidazolate Frameworks as Selective Carbon Dioxide Reservoirs. *Nature* **2008**, *453* (7192), 207−211.

(15) Park, K. S.; Ni, Z.; Côté, A. P.; Choi, J. Y.; Huang, R.; Uribe-Romo, F. J.; Chae, H. K.; O'Keeffe, M.; Yaghi, O. M. Exceptional Chemical and Thermal Stability of Zeolitic Imidazolate Frameworks. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103* (27), 10186−10191.

(16) Li, A.; Perez, R. B.; Wiggin, S.; Ward, S. C.; Wood, P. A.; Fairen-Jimenez, D. The Launch of a Freely Accessible MOF CIF Collection from the CSD. *Matter* **2021**, *4* (4), 1105−1106.

(17) Reed, D. A.; Keitz, B. K.; Oktawiec, J.; Mason, J. A.; Runcevski, T.; Xiao, D. J.; Darago, L. E.; Crocellà, V.; Bordiga, S.; Long, J. R. A Spin Transition Mechanism for Cooperative Adsorption in Metal-Organic Frameworks. *Nature* **2017**, *550* (7674), 96−100.

https://doi.org/10.1021/acs.jctc.4c00805
J. Chem. Theory Comput. 2024, 20, 9128−9137