Received: 11 December 2024 | Accepted: 27 January 2025

DOI: 10.1002/mgea.88

RESEARCH ARTICLE

MGE ADVANCES
Materials Genome Engineering
Open Access

Local large language model-assisted literature mining for on-surface reactions

Juan Xiang | Yizhang Li | Xinyi Zhang | Yu He | Qiang Sun

Materials Genome Institute, Shanghai Engineering Research Center for Integrated Circuits and Advanced Display Materials, Shanghai University, Shanghai, China

Correspondence
Qiang Sun.
Email: qiangsun@shu.edu.cn

Abstract

Large language models (LLMs) excel at extracting information from literatures. However, deploying LLMs necessitates substantial computational resources, and security concerns with online LLMs pose a challenge to their wider applications. Herein, we introduce a method for extracting scientific data from unstructured texts using a local LLM, exemplifying its applications to scientific literatures on the topic of on-surface reactions. By combining prompt engineering and multi-step text preprocessing, we show that the local LLM can effectively extract scientific information, achieving a recall rate of 91% and a precision rate of 70%. Moreover, despite significant differences in model parameter size, the performance of the local LLM is comparable to that of GPT-3.5 turbo (81% recall, 84% precision) and GPT-4o (85% recall, 87% precision). The simplicity, versatility, reduced computational requirements, and enhanced privacy of the local LLM makes it highly promising for data mining, with the potential to accelerate the application and development of LLMs across various fields.

KEYWORDS
data mining, large language models, on-surface synthesis, prompt engineering

1 | INTRODUCTION

The development of AI for science has aroused widespread interest in accelerating the discovery of new materials,[1–6] while also driving the growing demand for large-scale data. Scientific publications, as the primary way of communication within the scientific community, offer comprehensive and detailed scientific knowledge and advancements. Therefore, the automatic extraction of data and knowledge from the ever-growing body of scientific literature holds immense potential for different applications.[7] However, the main challenge in applying traditional data techniques to text mining lies in the representation, preprocessing, and handling of linguistic features in text data, which poses unique challenges to data mining algorithms.[8] Additionally, research publications are often unstructured or highly heterogeneous in format, presenting obstacles to the analysis of large-scale text.

Natural language processing (NLP) techniques integrate statistical analysis and machine learning with linguistic knowledge, allowing for extracting high-quality information from unstructured text.[9,10] Early "bag-of-words" models ignored the continuity of words within text and performed poorly on complex tasks. In contrast, sequence-to-sequence[11] (seq2seq) models account for the order/contextual information of each word and generate outputs of arbitrary length, which are widely used in applications of machine translation,[12,13] text summarization,[14] and chatbots.[15] However, one drawback of seq2seq models is their limited memory capacity when processing longer sequences. To address this issue, the models based on transformer architecture with attention mechanisms[16] are better equipped to handle longer sequences. Google's introduction of BERT[17] (bidirectional encoder representations from transformers) enhanced the transformer architecture by incorporating optimized algorithms such as the masked language model (MLM) and next sentence prediction (NSP). These advancements have enabled BERT to excel across various tasks and become a mainstream technology in natural language processing.

The development of natural language processing (NLP) has significantly advanced text extraction and data mining across various scientific publications.[18–23] Cole et al.[24] were among the pioneers in developing a materials science text extraction tool called ChemDataExtractor, which can automatically extract structured data for magnetic and battery materials. Subsequently, other widely used chemical and material tools emerged, including ChemicalTagger,[25] ChemListem,[26] ChemSpot,[27] and OSCAR4.[28] Li et al. combined regular expressions (RE) with machine learning (ML) to extract synthesis routes for Pd-based catalysts.[29] More recently, Huang et al.[30] employed a "question–answering" (Q/A) approach to fine-tune BERT, adapting it to a wide range of complex texts. Although these tools range from regular expression-based systems to smaller language models such as BERT, large language models (LLMs) offer more versatile and user-friendly solutions for information extraction. Unlike previous tools, which typically require pre-set parsing rules or model fine-tuning and often need to be retrained when applied to new subfields, LLMs demonstrate strong generalization and text comprehension capabilities.[31] They do not require specialized training for specific domains and are effective across various text structures. For example, Dagdelen et al.[32] introduced a straightforward method to fine-tune large language models (GPT-3, Llama-2) to jointly identify named entities and extract relations. Xie et al.[33] demonstrated the effectiveness of using ChatGPT for zero-shot information extraction in named entity recognition (NER). Park et al.[34] proposed a new method to enhance information extraction from mental health data by integrating medical knowledge graphs and large language models (LLMs). Proper prompt engineering often leads to effective enhancements in extraction outcomes without specific model fine-tuning or deep understanding of the principles of language models.[35] For instance, Polak et al.[36] proposed using ChatGPT and prompt engineering to extract accurate material data, achieving an precision of over 90% in testing. Researchers found that appropriate prompts could improve the precision of LLMs in answering specialized medical questions.[37] Without the "tricks" of extensive prompt engineering, ChatGPT performed poorly on several simple tasks in chemistry.[38] However, it is important to note that the majority of LLMs are commercial products that run in cloud environments. Recent research[39] shows that popular models such as ChatGPT with over a trillion parameters would require about 100 million GPUs to process at a rate of 50 tokens per second using GPT-4, each GPU processing at 60 TFLOPs/s. This scale of computation (excluding communication and data transmission costs) is equivalent to 160 large-scale companies, posing severe environmental concerns such as substantial energy consumption and carbon emissions. Moreover, LLMs running in cloud environments also face challenges related to user privacy protection and network latency.[40,41]

The full potential of LLM has not been fully explored in most of the fields in chemistry. In this work, we focus on the literatures of on-surface synthesis.[42,43] As a relatively emerging field of surface chemistry, there is a noticeable lack of specific training data for on-surface synthesis.[44–49] Compared to other disciplines, on-surface synthesis lags behind due to fewer specialized databases and related data, posing the challenge of representing these scarce data in text formats suitable for LLMs.

To avoid high costs, usage limitations of external services, and the potential leakage of private data, this study demonstrates a workflow for literature mining using a local LLM that can run on a personal PC. The workflow was exemplified by extracting entities and attributes of the synthetic data in the field of on-surface synthesis (OSS) or on-surface reaction. We achieved three objectives in this paper: (1) utilizing smaller-scale LLMs deployed on local computers for text extraction. (2) Achieving accurate and efficient extraction of entities and attributes through specially designed workflow prompts. (3) The accuracy and data refinement ability of the local LLM in literature text mining were evaluated compared with other prevalent language models. This work highlights a new strategy for text mining in scientific literatures using local LLMs, emphasizing its versatility and accessibility.

2 | RESULTS AND DISCUSSION

Knowledge distillation[50] and model quantization[51] are techniques that help to reduce the memory and computational demands of LLMs. The general distillation pipeline for LLMs involves a structured process to transfer knowledge from a complex teacher model to a simpler student model.[52] Model quantization reduces the number of bits needed to store data, decreasing the application's reliance on storage resources. Therefore, this offer solutions to deploy large-scale language models of smaller size on private computers. In this study, we use nous-hermes-llama2-13b. Q4_0,[53] a quantized 4-bit knowledge distillation trained model based on the llama2-13b model.[54]

Figure 1 illustrates the conceptual workflow powered by the local LLM. We begin with data preparation. Our system primarily collects scientific literatures in PDF and HTML formats. Given the need of a standardized format for subsequent data extraction, our system includes basic preprocessing for HTML and PDF files. All texts are then embedded to facilitate subsequent steps, where we employ strategy of prompt engineering to guide the local LLM to extract materials properties or synthetic parameters. To

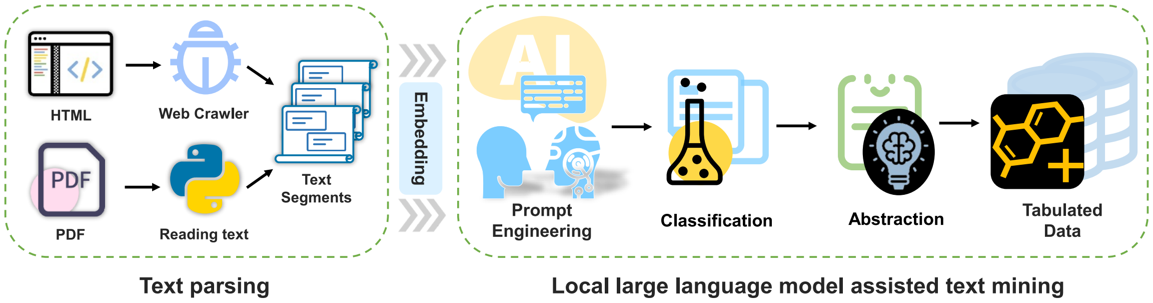
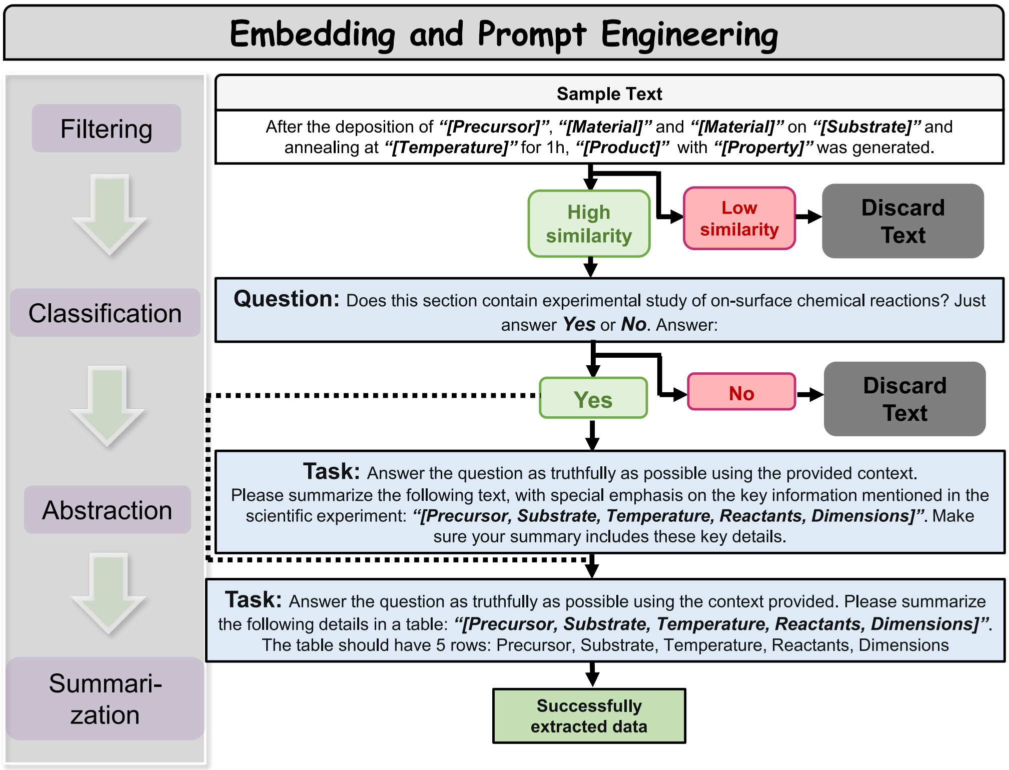


FIGURE 1 1 A simplified workflow describing the approach to extracting structured data. "Text parsing" process divides published articles into text segments. "local large language model assisted text mining" process consists of customized prompts to perform a series of processing for classifying, abstracting, and summarizing information into tabular data.



After the deposition of ***"[Precursor]"***, ***"[Material]"*** and ***"[Material]"*** on ***"[Substrate]"*** and annealing at ***"[Temperature]"*** for 1h, ***"[Product]"*** with ***"[Property]"*** was generated.

Question: Does this section contain experimental study of on-surface chemical reactions? Just answer ***Yes*** or ***No***. Answer:

Task: Answer the question as truthfully as possible using the provided context. Please summarize the following text, with special emphasis on the key information mentioned in the scientific experiment: ***"[Precursor, Substrate, Temperature, Reactants, Dimensions]"***. Make sure your summary includes these key details.

Task: Answer the question as truthfully as possible using the context provided. Please summarize the following details in a table: ***"[Precursor, Substrate, Temperature, Reactants, Dimensions]"***. The table should have 5 rows: Precursor, Substrate, Temperature, Reactants, Dimensions

FIGURE 2 A flowchart illustrating the data extraction from literatures of on-surface synthesis. It extracts structured data using a local large language model. Light blue boxes denote prompts provided to the model. Bold text within brackets "[]" signifies the values of extracted parameters.

maintain consistency with the multi-step process, we adopt a stepwise-guided prompting strategy.[55] In the filtering process, the prompts are designed to label useful content as "Yes". In subsequent abstraction and summarization processes, the design of the prompts adheres to three key principles: (1) provide accurate responses: the model is required to provide precise output, ensuring that key information is extracted from the literature with as much precision as possible, avoiding speculation or the generation of incorrect content. (2) Clarify task objectives: The prompts are designed to precisely guide the local LLM in identifying the key information to be extracted, such as synthesis conditions. (3) Data format: It is essential to ensure that the model's output is in a tabular data format.

In Figure 2, we provide a detailed depiction of extracting structured data. Texts are initially searched and compared to examples to identify their semantic similarities. The high-dimensional vector values of all text embeddings are compared with a small set of sample text embeddings using cosine similarity. The formula for cosine similarity is defined as follows:

$$\text{cos\_similartity} = \frac{A \bullet B}{\|A\|\|B\|},$$

where A and B are the vector embeddings of the two text samples, and ||A|| and ||B|| denote the magnitudes of the vectors A and B, respectively. This measure quantifies the cosine of the angle between the vectors, providing a metric of their similarity. The relevance scores derived are used as the basis for judgment; texts with high similarity are retained to enhance the quality of texts processed in subsequent steps, whereas texts with low relevance are filtered out.

It is important to note that the generation of embedding vectors and the calculation of similarity may be influenced by inherent search limitations, which could lead to unidentified but relevant text paragraphs. Therefore, to reduce the possibility of erroneous filtering and ensure more accurate operations of data in the workflow, we only filter out the texts with the lowest scores, retaining those with medium to high relevance and passing them to the classification process for the following categorization.

In the classification step, searching with only simple keywords have been proven to be ineffective.[56] The writing styles across different journals or papers do not adhere to a uniform standard. Descriptions of the synthetic data often vary in style, sometimes found within the “experiments” or “methods” sections, or described under “results and discussion,” or even dispersed across other sections of the paper. To address this challenge, we employ an incremental reading approach[55] to paper segmentations, focusing on one or two paragraphs at a time and asking the LLM to classify each part as either “yes” (paragraphs with target data, synthetic parameters in this case) or “no” (paragraphs without target data). The local LLM only passes paragraphs marked as “Yes” to the next step for abstraction (Figure 2).

The classification prompt request the LLM to determine if the provided context includes synthetic data, answering only “Yes” or “No”. The context is a text paragraph parsed from a complete research article, iteratively merged with the prompt and sent to the model to obtain a response. Each prompt represents a dialog instance, and the language model cannot view the answers to previous prompts, preventing potential biases in the decision-making for the current dialog.

In the step of data abstraction, the LLM utilizes its language understanding and generation capabilities to autonomously produce fluent contextually coherent abstracts containing the synthetic data. Notably, the abstraction process of the local LLM includes text cleaning and preprocessing, text normalization (such as correcting “150 8C” to “150°C” due to optical character recognition (OCR) technology flaws), and refining and emphasizing connections between information. In addition, it involves removing related but potentially redundant information (such as preprocessing steps for substrates), retaining only the most central data, thereby making the extraction process more efficient and focused. The prompt design follows a rigorous scientific approach to restore the authentic text information and specifically request the retention of key parameters of the on-surface reactions and the details of the reaction products. It ensures that these primary synthesis parameters are not lost during the abstraction process.

The final prompt of the abstraction process includes three parts: (i) ensuring that the local LLM uses the provided context to answer questions as truthfully as possible to counteract the inherent “hallucination” of language models; (ii) to prevent the loss of information during the abstraction, it specifically emphasizes the key information related to synthesis details mentioned in the paragraphs, listing the required categories of synthesis parameters; (iii) the context, consisting of paragraphs detailing experimental parts of on-surface synthesis conditions. The combined prompt produces a single question–answer interaction, thereby allowing the local LLM to generate an abstract of the given synthesis conditions as output.

One aspect where LLMs surpass traditional NLP methods is in their extensive pre-trained textual corpus. Consequently, LLMs can comprehend chemical nomenclature and reaction conditions, such as identifying and associating compound abbreviations (e.g., BPDSC) with their full names ([1,1′-biphenyl]-4,4′-disulfonyl dichloride). Traditional NLP models typically struggle in recognizing that BPDSC and [1,1′-biphenyl]-4,4′-disulfonyl dichloride refer to the same compound without a pre-compiled dictionary. In the case of extracting synthetic data from the literatures of on-surface synthesis, we select five key data considered most crucial for each on-surface synthesis experiment. Specifically, these parameters include precursors, products and their dimensions (0D, 1D, and 2D structures), substrates, and reaction temperatures, which are critical in controlling the kinetics and thermodynamics of each experiment. If no information is provided in a section, such as some literatures that might not involve dimensions of the products formed on the surface, we expect the LLM to respond with “N/A”.

The final prompt in summarization process consists of three parts: (i) requiring the LLM to summarize and tabulate the reaction conditions using human-provided text or information, answering as truthfully as possible to minimize hallucination; (ii) detailing the structure of the output table, listing expected categories and processing instructions; (iii) the context, consisting of paragraphs containing on-surface synthesis parameters abstracted from the previous step. It should be noted that we summarize and categorize most reaction types occurring on surfaces based on expert knowledge (see supporting information S1 for different types of on-surface reaction), and use the categorized reaction types as prompts for the LLM to choose the most fitting reaction type for the synthesis paragraphs. Given that the prompts for reaction types focus on categorization, and those for synthesis parameters focus on summarizing and extracting parameters, there is some conflict between the two. Therefore, we iteratively extract synthesis parameters and reaction types separately.

Following the aforementioned series of fully automated processes for text processing, 2034 text segments from 70 publications were screened and processed to parse and extract tabulated data (references to the literatures are provided in the supporting information S1).

To assess the accuracy of the local LLM in text mining, we conducted a comprehensive analysis of the entire dataset, rather than sampling. We first evaluated the hit rates for the reaction types, substrate types, and dimensions of molecular products extracted from the text paragraphs. Hit rates are used to determine whether key information is lost, indicating whether the text paragraphs containing the synthesis information are consistent with the synthetic reaction parameters described in the whole literature. For example, if Yan et al.[57] completed an imine formation reaction on Ag(111) with the products on the surface appearing as chains and rings (i.e., 1D and 2D), and the local LLM's extraction of the synthesis parameters from the segmented text paragraphs of this article includes and matches the above information, it is considered a hit; conversely, if it does not, it is a miss. As with Cai et al.,[58] who synthesized graphene nanoribbons via cyclodehydrogenation on both Ag(111) and Au(111), if the extraction results do not include Ag(111)—since the description in the literature primarily detailed the experiments on Au(111) while summarily describing Ag(111), leading to information loss—it represents a miss for Ag (111). In Figure 3, we display the hit rates and frequencies for reaction types, substrate, and dimensions of molecular products. The analysis shows that parameters with higher frequencies generally have higher hit rates, whereas those with lower frequencies tend to have lower hit rates. This positive-correlation trend suggests that text segmentation has a minor impact on the complete extraction of parameters. Specifically, out of 32 synthesis parameters, only 2 with lower frequencies were not accurately extracted (hit rate of 0), and these parameters have a low occurrence rate across the entire dataset, thus having limited impact on the overall extracted dataset. This study, by comparing hit rates and data frequencies, validates the effect of text segmentation on the data extraction process. The results indicate that the proportion of information loss caused by text segmentation is small and does not significantly impact the overall data extraction.

In Figure 4, we present an example of extraction in which the local LLM streamlines the original text paragraphs through the abstraction process to retain the five reaction parameters, and subsequently summarizes the data into a tabulated format. The results demonstrate that the abstraction process, while simplifying and standardizing the text, fully and accurately preserves the critical synthesis information required by the prompts. We have also compared the performances of the local LLM used in this work with two popular LLMs (GPT-3.5 and GPT-4). The local LLM achieved extraction results comparable to those of GPT-3.5 and GPT-4, demonstrating the effectiveness of the smaller-scale local LLM. In the following, we will evaluate the performance of the local LLM's extraction capabilities.

For each output synthesis parameter, we define three types of metrics: true positive (TP, correctly identifying synthesis parameters), false positive (FP, incorrectly assigning to the wrong synthesis parameters or extracting irrelevant information), and false negative (FN, failing to extract certain synthesis parameters). The primary metrics

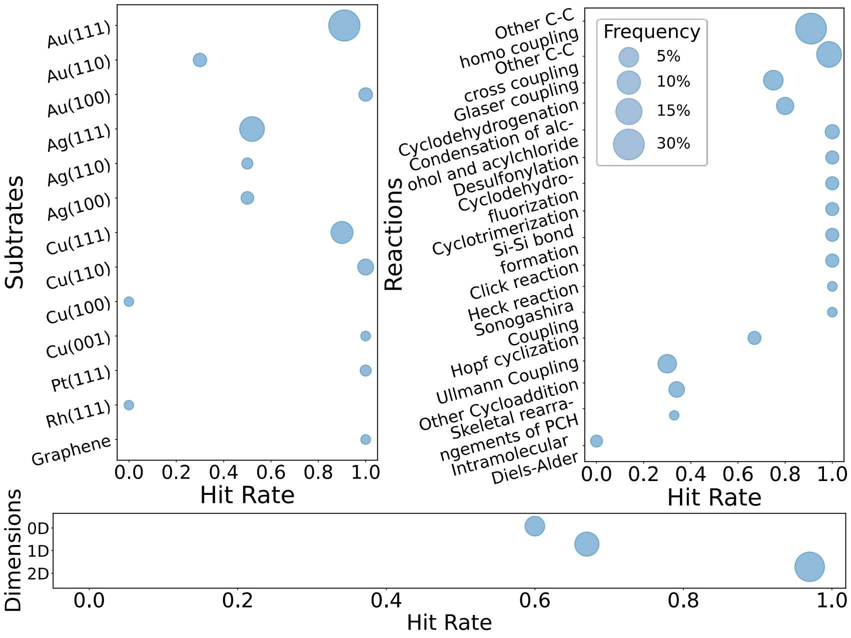
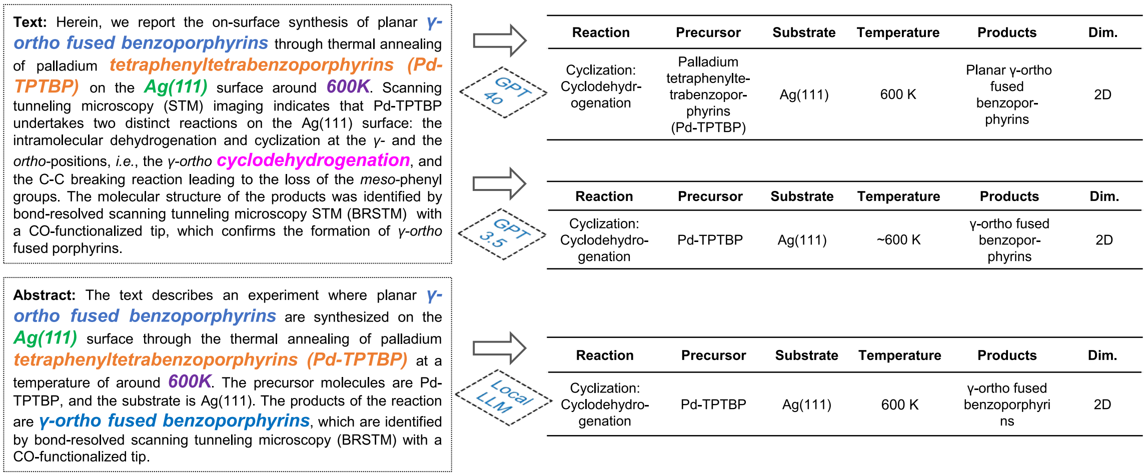


FIGURE 3 Hit rate of text segments on key information within the whole article. The size of each circle represents the frequency at which each attribute occurs.



Text: Herein, we report the on-surface synthesis of planar *γ-ortho fused benzoporphyrins* through thermal annealing of palladium *tetraphenyltetrabenzoporphyrins (Pd-TPTBP)* on the *Ag(111)* surface around *600K*. Scanning tunneling microscopy (STM) imaging indicates that Pd-TPTBP undertakes two distinct reactions on the Ag(111) surface: the intramolecular dehydrogenation and cyclization at the γ- and the *ortho*-positions, *i.e.*, the *γ-ortho* *cyclodehydrogenation*, and the C-C breaking reaction leading to the loss of the *meso*-phenyl groups. The molecular structure of the products was identified by bond-resolved scanning tunneling microscopy STM (BRSTM) with a CO-functionalized tip, which confirms the formation of *γ-ortho* fused porphyrins.

Abstract: The text describes an experiment where planar *γ-ortho fused benzoporphyrins* are synthesized on the *Ag(111)* surface through the thermal annealing of palladium *tetraphenyltetrabenzoporphyrins (Pd-TPTBP)* at a temperature of around *600K*. The precursor molecules are Pd-TPTBP, and the substrate is Ag(111). The products of the reaction are *γ-ortho fused benzoporphyrins*, which are identified by bond-resolved scanning tunneling microscopy (BRSTM) with a CO-functionalized tip.

GPT 4o

Reaction	Precursor	Substrate	Temperature	Products	Dim.
Cyclization: Cyclodehydrogenation	Palladium tetraphenyltetrabenzoporphyrins (Pd-TPTBP)	Ag(111)	600 K	Planar γ-ortho fused benzoporphyrins	2D

GPT 3.5

Reaction	Precursor	Substrate	Temperature	Products	Dim.
Cyclization: Cyclodehydrogenation	Pd-TPTBP	Ag(111)	~600 K	γ-ortho fused benzoporphyrins	2D

Local LLM

Reaction	Precursor	Substrate	Temperature	Products	Dim.
Cyclization: Cyclodehydrogenation	Pd-TPTBP	Ag(111)	600 K	γ-ortho fused benzoporphyrins	2D

FIGURE 4 Extract results. An example of extracting parameters for on-surface synthesis using different large language models, and comparisons of their results.

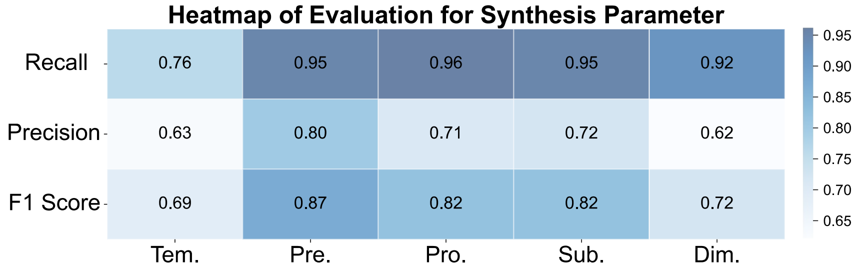


FIGURE 5 Performance evaluation. Recall, precision and F1 score of the synthesis parameters of temperature (Tem.), precursor (Pre.), product (Pro.), substrate (Sub.) and dimension (Dim.).

used to evaluate the performance are precision and recall, defined as follows:

$$\text{Precision} = \frac{\text{Ture Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

$$\text{Recall} = \frac{\text{Ture Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

In this study, the performance of extracting synthesis parameters from literatures of on-surface synthesis are shown in Figure 5. Overall, the local LLM exhibited commendable performance in extracting various synthesis parameters, particularly in recall, with four synthesis parameters achieving a recall rate of over 0.9. Among all the text segments tested, precursors achieved the highest F1 score (0.87). Products achieved high F1 scores (0.82). Precursors and products that follow chemical naming conventions have relatively standardized and clear chemical names and formulas. Therefore, both precursors and products achieved high precision, with the precision of precursors slightly higher and the precision of products both above 0.7. The F1 scores of dimensions and size are relatively low, but also close to 0.7.

We found that the recall and precision of substrates were high (Figure 5), reaching 0.95 and 0.72, respectively. This phenomenon can be attributed to the fact that the substrate parameters mentioned above usually follow a fixed pattern (such as single chemical element + crystallographic orientation), and the types of substrates used are relatively limited and standardized within the field, as indicated by the probability distribution of substrates in dataset from Figure 3. Compared to substrates, the extractions of precursors and products are more complex, involving not only named entity recognition but also the extraction of “relationship pairs” from the text. For example, “The first SCOF-1 is obtained by the molecular dehydration of 1,4-benzenediboronic acid (BDBA), … The second SCOF-2 network is formed by the condensation reaction of BDBA and 2,3,6,7,10,11-hexahydroxytriphenylene (HHTP) …” where the precursor for SCOF-1 is BDBA, and for SCOF-2, it is BDBA and

HHTP.[59] Despite dealing with relationship pairs, the local LLM's extraction performance for precursors and products is satisfactory, achieving recall scores of 0.95 and 0.96, and precision of 0.80 and 0.71, respectively. A typical example is the text: "We designed the compound 1,6-di-2-naphthylhex-3-ene-1,5-diyne (DNHD), which has an enediyne-moiety and two naphthyl groups. According to previous studies of the Bergman reaction, the intramolecular cyclization takes place upon stimulation by heat, forming the diradical intermediate, which then grows into a linear polyphenylene through radical polymerization."[60] In this text, the extraction results for precursors and products are "1,6-di-2-naphthylhex-3-ene-1,5-diyne (DNHD)" and "Linear polyphenylene," showing our framework's ability to extract complete synthetic precursors and products, rather than intermediate products such as "diradical intermediate."

Additionally, we find that the local LLM seems better at handling word-class data. For data involving numerical data, that is, temperature and dimension, precision is slightly lower. Further analysis of FP samples for temperature revealed that when the text does not specify the reaction temperature of on-surface synthesis, the local LLM defaults the reaction to "occurring at room temperature," which may stem from the "basic knowledge" formed by LLMs' extensive pre-trained corpus. The fundamental reason is the model's lack of extensive training data on scientific corpus. Similarly, dimension also exhibited "basic knowledge," with the model defaulting the unspecified dimensions of on-surface synthesis products to "2D". We thus look forward to the emergence of language models that incorporate training data of scientific literatures with chemical knowledge to enhance the extraction performance.

We found that the local LLM seems to be better at handling word class data. For numerical data, namely temperature and dimension, the precision is slightly lower. Overall, temperature and dimension perform worse than precursors, products and substrates. This may be due to the lack of numerical training data during the model training process. In addition, further analysis of the FP sample of temperature revealed that when the paragraph does not specify the reaction temperature of the on-surface synthesis, the local LLM will default the reaction to "occur at room temperature". This may be due to the fact that there are no clear values or descriptions in the synthesis paragraph. For example, "Here we demonstrate what is to our knowledge the first example showing how the Bergman cyclization reaction can be employed on a surface to explore the feasibility of forming covalently linked conjugated carbon nanostructures. In this work, we designed the compound 1,6-di-2-naphthylhex-3-ene-1,5-diyne (DNHD), which has an enediyne moiety and two naphthyl groups (Scheme 1). According to previous studies of the Bergman reaction, the intra- molecular cyclization takes place upon stimulation by heat, forming the diradical intermediate, which then grows into a linear polyphenylene through radical polymerization (Scheme1). The Cu(110) surface was chosen as the substrate because of its well-known one-dimensional (1D) templating effect to facilitate the formation of unidirectional linear structures."[60] In this case, the temperature is not explicitly given and maybe mentioned in a specific context (other synthesis conditions paragraph in the literature); the model defaults to assuming that the temperature is room temperature without contextual support, which leads to lower precision of the temperature. For dimension, the expression in the literature is often not so clear; for example, in "In the overview image of these two precursors, the islands consisting …",[61] the model infers the dimension to be 2D based on "islands", but when there is no clear word in the text to infer, the model may default to 2D based on words such as "surface"; for example, in "Most importantly, such an on-surface chemistry strategy provides us a relatively facile and efficient method for the stereoselective synthesis of specific diene moieties through dehydrogenative/dehalogenative homocoupling reactions of alkenes/alkenyl bromides on the same surface (i.e., Cu(110) here) (Scheme 1), which demonstrates the great potential of on-surface synthesis for exploring new chemistry and con-structing novel surface nanostructures. After the deposition of BVBP molecules on Cu(110) held at low temperatures (~170 K), the molecules already de-brominate on the surface as shown in …",[62] the dimension of the product was not explicitly specified, but the model made incorrect judgments based on words such as "on-surface synthesis", "on the surface" and "on the same surface", and recalled the wrong answer 2D, which resulted in a high recall rate but a low precision rate, even with the models of GPT-3.5-turbo and GPT-4o at times.

Moreover, we required the local LLM to classify all the extracted reactions by their reaction types. The expert-defined reaction types, along with the prompts used (both listed in the supporting information S1), are input into the local LLM, which then selects the most appropriate reaction type for each synthesis paragraph. The experimental results revealed limitations of the local LLM in processing the classification of reaction types. In particular, when the classification results of reaction types were input in the form of prompts that occupy many tokens, the model exhibited lower precision, despite a good recall rate (listed in the supporting information S1). Through detailed analysis of FP, we found that "hallucination" accounted for 35% of the total number of FP. This phenomenon may be attributed to two factors: firstly, the inherent "hallucination" issue of large language models; secondly, the increased number of tokens limits the model's ability to process more information.

After filtering and classifying 2034 text segments, 298 samples that containing surface synthesis information were retained for evaluation. Figure 6 presents a comparison of the extraction performance between the local LLM and other prevalent language models. The overall recall and precision of the local LLM are 91% and 70%, respectively. This is particularly impressive for a 13B parameter model that is quantized and running solely on a local PC without any fine-tuning. This proves that our series of processes successfully enable the local LLM to extract information from scientific

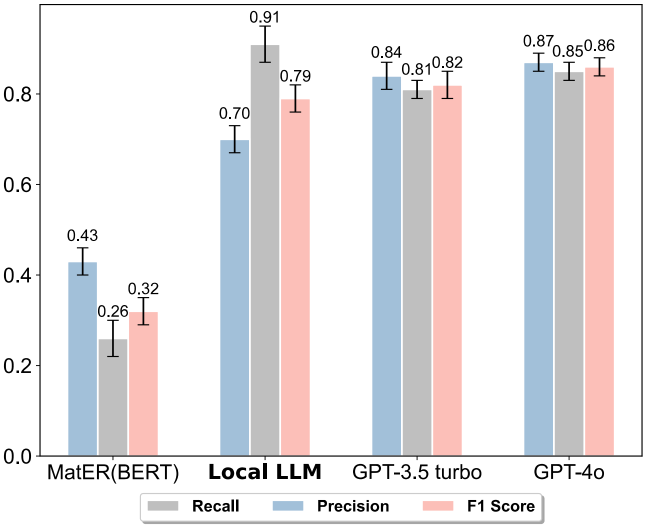


FIGURE 6 Comparison with other models. The performance of the local large language model compared to other prevalent language models.

literatures. In addition, it can be observed that local LLM has the characteristics of high recall and low precision, which is because more parameters and computing resources are needed to perform reasoning when dealing with tasks that require complex reasoning and detail description. For example, the wrong results are recalled when the extracted parameters are not obvious; on the other hand, the extracted results are incomplete when the parameters are described in detail in the text. These together lead to high recall and low precision of local LLM. We also evaluated the competitive performance of local LLM compared to other language models. MatEntityRecognition[63] is a BERT-based tool designed to extract materials, precursors, and targets within a paragraph (which is very consistent with our work). We tested its ability to extract precursors and products without any fine-tuning by inputting surface synthetic paragraphs into MatEntityRecognition. The results showed a recall rate of 29% and a precision rate of 47%, showing the limitations of the BERT model in terms of generality and accuracy. Because the MatER model is a small language model requiring optimization for specific datasets, it is evident that the recall rate of MatER is poor without any specific fine-tuning due to the shift in the target task. Its limited performance and lack of generalization capabilities restrict its ability to handle generalized tasks. In simpler tasks with easily extractable information, such as “phenylacetylene and iodobenzene react on smooth Au(111) under vacuum conditions to yield the homocoupling products diphenyldiacetylene and biphenyl,”[64] the precursors “phenylacetylene” and “iodobenzene” are recalled accurately, and “diphenyldiacetylene” is correctly identified as a product. However, “biphenyl” as a product was not recalled. Furthermore, it can be observed that GPT-3.5 turbo and GPT-4o, which have more parameters and require expensive computational resources, have a recall of 84% and 87% and a precision of 81% and 85%, respectively. This shows that local LLM performs well, but the recall and precision rates of GPT-3.5 turbo and GPT-4o are more balanced. The balanced recall and precision of the GPT models are attributable to their larger parameter sizes and high computational resources. These models can easily understand relationships in simpler tasks, enabling them to extract the correct answer. For more complex problems, they have the computational resources and sufficient parameters to perform reasoning and derive accurate results.

We provide a more detailed analysis in the following example[65]: “In method I, intact molecules are deposited onto the surface, where they are subsequently activated by dissociation of the substituent atoms upon heating of the sample. In method II, this activation has already occurred in the evaporator, and the activated molecules are subsequently deposited on the surface, kept at room temperature. In both cases, the activated building blocks are covalently connected directly on the surface upon thermal diffusion. The design of suitable molecules for the formation of such nanostructures requires the incorporation of latent reactive legs that can be activated selectively, without breaking the other bonds. For this purpose, carbon–halogen bonds, exhibiting much smaller binding energies compared to the central framework, were chosen. After selective thermal dissociation either on the surface (method I) or in the evaporator (method II), the resulting activated (most likely radical) fragments can combine in an addition reaction without producing further byproducts. Considering these requirements, we have chosen a porphyrin with four phenyl legs as a central building block.” It is important to note that the synthetic conditions in this example are not obvious and the model needs to spend computational resources to make reasonable inferences to get a very accurate answer. Therefore, it helps to understand the performance differences between models with different parameter sizes and computational resource requirements. In this example, we excluded the extraction of information

TABLE 1 The extracted results of local LLM, GPT-3.5-turbo, and GPT-4o models on example paragraphs.

	Precursor	Substrate	Temperature	Products	Dimensions
Local LLM	Activated molecules	Evaporator	N/A	Nanostructures	2D
GPT-3.5 turbo	Porphyrin molecules	N/A	N/A	Activated fragments (likely radicals)	N/A
GPT-4o	Porphyrin molecules	N/A	N/A	Covalently bonded nanostructures	N/A

Abbreviation: LLM, large language model.

from method I because it contains too few synthetic conditions to evaluate the model's ability, so we focused on the extraction of synthetic conditions from method II.

The ground truth of this example is that the precursor is “porphyrin with four phenyl legs”, the product is “covalently linked porphyrin-based nanostructure”, and other synthesis parameters should be “N/A”. The MatER model performed the worst, failing to recall both the precursor and the product (labeled as false negative, FN). The local LLM recalled some information, as shown in Table 1, but some of the recalled answers are inaccurate (the overall performance of the local LLM was high recall but low precision). First, the recall of the substrate was obviously incorrect and did not meet expectations. The reason why the recall result of the dimension is wrong is similar to the previous description, which is due to the limited computing resources and model parameters that lead to the insufficient reasoning ability. The recall of the precursor, “activated molecule”, was also incorrect (identified as false positive, FP). In contrast, the recall of the precursor with GPT-3.5 turbo and GPT-4o was more accurate (identified as true positive, TP). However, GPT-3.5 turbo's recall of the product was not as accurate as GPT-4o. The “activated fragment (probably a free radical)” is not the final product after surface binding; the final product should be the “covalently linked porphyrin-based nanostructure” mentioned in the passage. Therefore, we determined that the results provided by GPT-4o are consistent with the original text. This comparison clearly shows how the lack of reasoning ability affects the accuracy of the model output and whether more detailed and accurate answers can be extracted.

In summary, local LLM performs well overall, especially when computing resources and parameters are limited. It is undeniable that the GPT models have demonstrated stronger reasoning capabilities and accuracy advantages when faced with complex problems.

3 | CONCLUSION

In conclusion, we demonstrate the application of the local LLM, through a series of processes and appropriate prompt engineering, in extracting scientific data. By using a case study from the on-surface synthesis domain, we show that local LLMs, running solely on personal PC terminals, can effectively extract scientific information, and perform comparably in extraction tasks to GPT3.5 turbo and GPT-4o models. This contrasts with the currently prevalent LLMs that either run online or on clusters, requiring substantial computational resources and being costly. Another primary advantage of this method is its strong privacy. In terms of security, sensitive or confidential data must be sent for processing, our work presents a potential solution to this issue. This study showcases the potential of LLMs in assisting with literature data mining in the fields of chemistry.

Finally, we provide a pipeline written in Python code named OSSExtract (see “Data Availability” for more details). It leverages LLMs deployed on a personal PC, allowing users to transform complexly formatted scientific articles into tabulated data, forming databases within domains.

4 | METHODS

The main specifications of the computer platform for running the program are as follows: i5-13490F @ 2.5 GHz CPU, 16 GB RAM. This paper utilizes the GPT4All ecosystem (https://www.nomic.ai/gpt4all), selecting the quantized nous-hermes-llama2-13b.Q4_0.gguf model based on the llama2-13b model. The model all-MiniLM-L6-v2 is simple yet effective, used for converting each fragment into a 384-dimensional vector text embedding. The local LLM is scripted in Python 3.11.4, running on a local hardware specified above. For the local LLM model, the settings are temperature = 0.0, top_*p* = 0.6. For the GPT models, the GPT-3.5 turbo and GPT-4o from GPT-4 are used, with temperature = 0.0. No system prompts (empty string) are used in any of the models. To evaluate the model performances, for the local LLM, a comprehensive analysis of the entire dataset was conducted, whereas for other models, a 10% sampling of the dataset was used. To ensure the unbiasedness of the data, we used a simple random sampling method. This method ensures that all samples have the same probability of being selected, ensuring the randomness and unbiasedness of the selected samples. We adopted a multi-round evaluation strategy to randomly sample the model three times (10% each time) and take the average of the evaluation to ensure the reliability of the evaluation results.

AUTHOR CONTRIBUTIONS

Juan Xiang: Investigation; writing—original draft; writing—review & editing; data curation; software; formal analysis; visualization; validation; conceptualization; methodology. **Yizhang Li**: Writing—review & editing; data curation;

visualization. **Xinyi Zhang**: Writing—review & editing; data curation. **Yu He**: Writing—review & editing; data curation. **Qiang Sun**: Writing—original draft; writing—review & editing; conceptualization; funding acquisition; project administration; resources; supervision.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 22072086, 22302120).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The code used in this study is available at https://github.com/juanxiang-shu/OSSExtractor. The quantized llama2 13b model is available at https://huggingface.co/TheBloke/Nous-Hermes-Llama2-GGUF/blob/7ea9e8b35ded6f4842a331dd786619779fb20442/nous-hermes-llama2-13b.Q4_0.gguf. The data extracted by local LLM, MatEntityRecognition, GPT3.5_tubo, and GPT4o are available at https://github.com/juanxiang-shu/OSSExtract/Data.

ORCID

Qiang Sun https://orcid.org/0000-0003-4903-4570

REFERENCES

1. Meng K, Huang C, Wang Y, et al. BNM-CDGNN: batch normalization multilayer perceptron crystal distance graph neural network for excellent-performance crystal property prediction. *J Chem Inf Model*. 2023;63(19):6043-6052.
2. Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett*. 2018;120(14):145301.
3. Xie J. Prospects of materials genome engineering frontiers. *Mater Genome Eng Adv*. 2023;1(2):e17.
4. Tian Z, Yang Y, Zhou S, et al. High-dimensional Bayesian optimization for metamaterial design. *Mater Genome Eng Adv*. 2024;2(4).
5. Lin J, Ban T, Li T, et al. Machine-learning-assisted intelligent synthesis of UiO-66(Ce): balancing the trade-off between structural defects and thermal stability for efficient hydrogenation of Dicyclopentadiene. *Mater Genome Eng Adv*. 2024;2(3).
6. Zhao J, Lai J, Wang J, et al. Accelerating spin Hall conductivity predictions via machine learning. *Mater Genome Eng Adv*. 2024;2(4):e67.
7. Kononova O, He T, Huo H, Trewartha A, Olivetti EA, Ceder G. Opportunities and challenges of text mining in materials research. *iScience*. 2021;24(3):102155.
8. Moed, H; Glänzel, W; Schmoch, U Handbook of quantitative science and technology research: the use of publication and patent statistics in studies of S&T systems; 2005.
9. Lauriola I, Lavelli A, Aiolli F. An introduction to deep learning in natural language processing: models, techniques, and tools. *Neurocomputing*. 2022;470:443-456.
10. Olivetti EA, Cole JM, Kim E, et al. Data-driven materials research enabled by natural language processing and information extraction. *Appl Phys Rev*. 2020;7(4).
11. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. *ArXiv*. 2014:1409.
12. He X, Haffari G, Norouzi M. Sequence to sequence mixture model for diverse machine translation. *arXiv Prepr arXiv:1810.07391*. 2018:583-592.
13. Bahar P, Brix C, Ney H. Towards two-dimensional sequence to sequence model in neural machine translation. *arXiv Prepr arXiv:1810.03975*. 2018:3009-3015.
14. Shi T, Keneshloo Y, Ramakrishnan N, Reddy CK. Neural abstractive text summarization with sequence-to-sequence models. *ACM Trans Data Sci*. 2021;2(1):1-37.
15. Palasundram K, Sharef NM, Nasharuddin N, Kasmiran K, Azman A. Sequence to sequence model performance for education chatbot. *Int J Emerg Technol Learn (iJET)*. 2019;14(24):56-68.
16. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017.
17. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding; 2018.
18. Guo J, Ibanez Lopez A, Gao H, et al. Automated chemical reaction extraction from scientific literature. *J Chem Inf Model*. 2021;62(9):2035-2045.
19. Trewartha A, Walker N, Huo H, et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*. 2022;3(4):100488.
20. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*. 2021;38(8):2102-2110.
21. Hellert T, Montenegro J, Pollastro A. PhysBERT: a text embedding model for physics scientific literature. *arXiv Prepr arXiv:2408.09574*. 2024;2(4).
22. Dalla Torre H, Gonzalez L, Mendoza Revilla J, et al. The nucleotide transformer: building and evaluating robust foundation models for human genomics. *BioRxiv*. 2023;2023. 2001. 2011.523679.
23. Ostendorff M, Rethmeier N, Augenstein I, Gipp B, Rehm G. Neighborhood contrastive learning for scientific document representations with citation embeddings. *arXiv Prepr arXiv:2202.06671*. 2022.
24. Swain MC, Cole JM. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J Chem Inf Model*. 2016;56(10):1894-1904.
25. Hawizy L, Jessop D, Adams N, Murray Rust P. ChemicalTagger: a tool for semantic text-mining in chemistry. *J cheminformatics*. 2011;3(1):17.
26. Corbett P, Boyle J. Chemlistem: chemical named entity recognition using recurrent neural networks. *J Cheminformatics*. 2018;10(1):59.
27. Rocktäschel T, Weidlich M, Leser U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinforma Oxf Engl*. 2012;28(12):1633-1640.
28. Jessop D, Adams S, Willighagen E, Hawizy L, Murray Rust P. OSCAR4: a flexible architecture for chemical textmining. *J cheminformatics*. 2011;3(1):41.
29. Li S, Zhang Y, Fang Z, et al. Extracting the synthetic route of Pd-based catalysts in methanol steam reforming from the scientific literature. *J Chem Inf Model*. 2023;63(20):6249-6260.
30. Huang S, Cole J. BatteryBERT: a pretrained Language Model for battery database enhancement. *J Chem Inf Model*. 2022;62(24):6365-6377.
31. Bai X, Xie Y, Zhang X, Han H, Li JR. Evaluation of open-source Large Language models for metal–organic frameworks research. *J Chem Inf Model*. 2024;64(13):4958-4965.
32. Dagdelen J, Dunn A, Lee S, et al. Structured information extraction from scientific text with large language models. *Nat Commun*. 2024;15(1):1418.
33. Xie T, Li Q, Zhang J, Zhang Y, Liu Z, Wang H. Empirical study of zero-shot NER with ChatGPT. In: *Conference on Empirical Methods in Natural Language Processing*; 2023.
34. Park C, Lee H, Jeong Or. Leveraging medical knowledge graphs and Large Language models for enhanced mental disorder information extraction. *Future Internet*. 2024;16(8):260.
35. Khot T, Trivedi H, Finlayson M, et al. Decomposed prompting: a modular approach for solving complex tasks. *ArXiv*. 2022. abs/2210.02406.

36. Polak MP, Morgan D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat Commun*. 2024;15(1):1569.
37. Filienko D, Wang Y, Jazmi CE, et al. Toward Large Language models as a therapeutic tool: comparing prompting techniques to improve GPT-delivered problem-solving therapy. *arXiv Prepr arXiv:2409.00112*. 2024.
38. White AD, Hocky GM, Gandhi HA, et al. Assessment of chemistry knowledge in large language models that generate code. *Digit Discov*. 2023;2(2):368-376.
39. Liu Z, Zhao C, Iandola F, et al. MobileLLM: optimizing sub-billion parameter language models for on-device use cases. *arXiv:2402.14905*. 2024.
40. Das BC, Amini MH, Wu Y. Security and privacy challenges of Large Language models: a survey. *ArXiv*. 2024. abs/2402.00888.
41. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (LLM) security and privacy: the Good, the Bad, and the Ugly. *High-Confidence Comput*. 2024;4(2):100211.
42. Clair S, de Oteyza DG. Controlling a chemical coupling reaction on a surface: tools and strategies for on-surface synthesis. *Chem Rev*. 2019;119(7):4717-4776.
43. Grill L, Hecht S. Covalent on-surface polymerization. *Nat Chem*. 2020;12(2):115-130.
44. Qie B, Wang Z, Jiang J, et al. Synthesis and characterization of low-dimensional N-heterocyclic carbene lattices. *Science*. 2024;384(6698):895-901.
45. Chenxiao Z, Bhagwandin D, Xu W, et al. Dramatic acceleration of the hopf cyclization on gold(111): from enediynes to peri-fused diindenochrysene graphene nanoribbons. *J Am Chem Soc*. 2024;146(4):2474-2483.
46. Chi L, Wang L, Han Y, et al. Synthesis of hexabenzocoronene-cored graphdiyne nanosheets through dehydrogenative coupling on Au(111) surface. *Angew Chem*. 2024;136(45):e202411722.
47. Bonifazi D, Deyerling J, Berna BB, et al. Solution vs on-surface synthesis of peripherally oxygen-annulated porphyrins through C-O bond formation. *Angew Chem Int Ed*. 2024;n/a(n/a):e202412978.
48. Jiang H, He Y, Lu J, et al. Unraveling the mechanisms of on-surface photoinduced reaction with polarized light excitations. *ACS Nano*. 2024;18(1):1118-1125.
49. Jiang H, Lu J, Zheng F, Zhu Z, Yan Y, Sun Q. Steering on-surface polymerization through coordination with a bidentate ligand. *Chem Commun*. 2023;59(52):8067-8070.
50. Gu Y, Dong L, Wei F, Huang M. Knowledge distillation of Large Language models. *ArXiv*. 2023. abs/2306.08543.
51. Zhu X, Li J, Liu Y, Ma C, Wang W. A survey on model compression for Large Language models. *ArXiv*. 2023. abs/2308.07633.
52. Xu X, Li M, Tao C, et al. A survey on knowledge distillation of Large Language models. *ArXiv*. 2024. abs/2402.13116.
53. https://huggingface.co/TheBloke/Nous-Hermes-Llama2-GGUF
54. https://huggingface.co/meta-llama/Llama-2-13b
55. Zheng Z, Zhang O, Borgs C, Chayes JT, Yaghi OM. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *J Am Chem Soc*. 2023;145(32):18048-18062.
56. Smeaton AF. Progress in the application of natural language processing to information retrieval tasks. *Comput J*. 1992;35(3):268-278.
57. Yan Y, Zheng F, Zhu Z, Lu J, Jiang H, Sun Q. On-surface synthesis of ethers through dehydrative coupling of hydroxymethyl substituents. *Phys Chem Chem Phys*. 2022;24(36):22122-22128.
58. Cai J, Ruffieux P, Jaafar R, et al. Atomically precise bottom-up fabrication of graphene nanoribbons. *Nature*. 2010;466(7305):470-473.
59. Zwaneveld NAA, Pawlak R, Abel M, et al. Organized Formation of 2D extended covalent organic frameworks at surfaces. *J Am Chem Soc*. 2008;130(21):6678-6679.
60. Sun Q, Zhang C, Li Z, et al. On-surface formation of one-dimensional polyphenylene through bergman cyclization. *J Am Chem Soc*. 2013;135(23):8448-8451.
61. Gao HY, Held PA, Amirjalayer S, et al. Intermolecular On-Surface σ-Bond Metathesis. *J Am Chem Soc*. 2017;139(20):7012-7019.
62. Kanuru VK, Kyriakou G, Beaumont SK, Papageorgiou AC, Watson DJ, Lambert RM. Sonogashira Coupling on an Extended Gold Surface in Vacuo: reaction of Phenylacetylene with Iodobenzene on Au(111). *J Am Chem Soc*. 2010;132(23):8081-8086.
63. https://github.com/CederGroupHub/MatEntityRecognition
64. Grill L, Dyer M, Lafferentz L, Persson M, Peters M, Hecht S. Nanoarchitectures by covalent assembly of molecular building blocks. *Nat Nanotechnol*. 2007;2(11):687-691.
65. Sun Q, Cai L, Ma H, Yuan C, Xu W. The stereoselective synthesis of dienes through dehalogenative homocoupling of terminal alkenyl bromides on Cu(110). *Chem Commun*. 2016;52(35):6009-6012.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Xiang J, Li Y, Zhang X, He Y, Sun Q. Local large language model-assisted literature mining for on-surface reactions. *MGE Advances*. 2025;3(1):e88. https://doi.org/10.1002/mgea.88