Luca Rettenberger*, Marc F. Münker, Mark Schutera, Christof M. Niemeyer, Kersten S. Rabe, and Markus Reischl

Using Large Language Models for Extracting Structured Information From Scientific Texts

https://doi.org/10.1515/cdbme-2024-2129

Abstract: Extracting structured information from scientific works is challenging as sought parameters or properties are often scattered across lengthy texts. We introduce a novel iterative approach using Large Language Models (LLMs) to automate this process. Our method first condenses scientific literature, preserving essential information in a dense format, then retrieves predefined attributes. As a biomedical application example, our concept is employed to extract experimental parameters for preparing Metal-Organic Frameworks (MOFs) from scientific work to enable complex and information-rich applications in the biotechnology-oriented life sciences. Our open-source method automates extracting information from verbose texts, converting them into structured and easily navigable data. This considerably improves scientific literature research by utilizing the power of LLMs and paves the way for enhanced and faster information extraction from extensive scientific texts.

Keywords: Metal-Organic Frameworks, Large Language Models, Information Extraction, Automation

1 Introduction

Rapid advancements in Deep Learning (DL) during the last decades have transformed the modern science landscape. In particular, recent developments in Large Language Models (LLMs) revealed extensive capabilities that gain far-reaching recognition in all research domains, biomedical engineering being no exception [8].

LLMs excel in clinical and educational applications, aiding experts in answering patient queries and generating medical tests automatically [9]. They are also widely utilized in predictive chemistry [5], synthesis prediction [6], and drug design [4]. In addition to the multitude of applications for education and experiment design, LLMs are predestined for analyzing scientific works due to their text-based nature. Extracting knowledge from these texts is particularly valuable, given the scattered nature of information such as synthesis conditions and experiment designs across lengthy documents.

The automated extraction of experimental attributes for Metal-Organic Frameworks (MOFs) from scientific texts with LLMs is particularly interesting, as they are ideal candidates for applications ranging from drug delivery and bioimaging to biosensing and biocatalysis [13]. MOFs offer properties that are useful in biomedical engineering, such as their large surface area, tunable pore size, biocompatibility, and ability to create controlled microenvironments that allow precise manipulation of biochemical reactions and interactions in biological systems. MOFs consist of metal ions and linker molecules that form compact, regular crystals through ionic interactions. Enzymes can also be incorporated into these as biological cargo, resulting in biohybrid enzyme-MOF composites. All this makes the experiment design with MOFs parameter-rich and complex. Hence, several approaches use LLMs to accelerate and automate literature research concerning experiment design with MOFs. These include, among others, building knowledge graphs for MOFs [1], extracting experimental details from abstracts of scientific papers [3], and distilling synthesis information from scientific literature [6, 14]. All these approaches simplify and speed up work with MOFs remarkably, but can only be employed for short texts, as LLMs are generally limited in the amount of input they can process at once. A usual limit is 4.096 tokens (roughly 3.000 words) [10]. This shortcoming is critical if the experimental parameters for MOFs are to be identified automatically, as they are usually distributed throughout long scientific texts and supplementary information. While there are LLMs designed to process vast inputs simultaneously [7], such solutions require far more resources and often struggle with maintaining coherence and consistency across lengthy texts [11].

Our latest research introduces a concept that uses LLMs to analyze lengthy and complex scientific texts effectively, preserving all important information by iterative condensing and summarizing. Our concept is illustrated through the automated extraction of MOF attributes. With our approach, we strive to relieve domain experts by automating the process of extracting specific information from lengthy scientific texts. Our open-source method, is available at https://osf.io/nhegv/.

***Corresponding author: Luca Rettenberger,** Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany, e-mail: luca.rettenberger@kit.edu
Mark Schutera, Markus Reischl, Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Germany
Marc F. Münker, Christof M. Niemeyer, Kersten S. Rabe, Institute for Biological Interfaces, Karlsruhe Institute of Technology, Germany

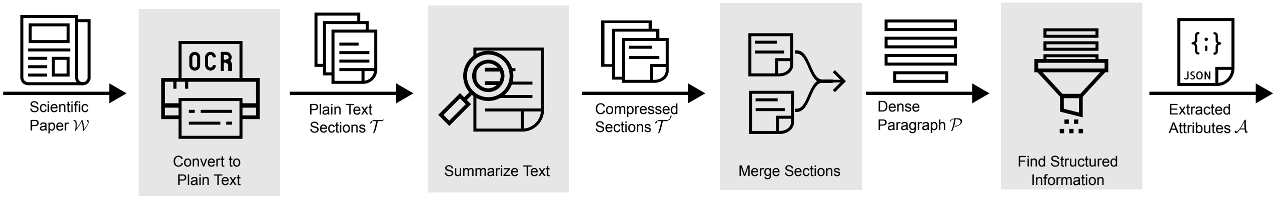
Scientific Paper $\mathcal{W}$
OCR
Convert to Plain Text
Plain Text Sections $\mathcal{T}$
Summarize Text
Compressed Sections $\mathcal{T}'$
Merge Sections
Dense Paragraph $\mathcal{P}$
Find Structured Information
{;}
JSON
Extracted Attributes $\mathcal{A}$

Fig. 1: Overview of our concept: A scientific work (paper) $\mathcal{W}$ is converted to plain text with Optical Character Recognition (OCR), split into sections $\mathcal{T}$, and summarized by an LLM, keeping all relevant information. The shortened sections $\mathcal{T}'$ are then merged into one, redundancy-free and information-dense paragraph $\mathcal{P}$. This paragraph is then used to extract predefined relevant information (attributes) $\mathcal{A}$ and converted into a structured format for output.

Tab. 1: The *Attributes* $\mathcal{A}$ to be extracted from the scientific works $\mathcal{W}$, sorted by *Complexity*.

Attribute $\mathcal{A}$	**Complexity**
Name of the formulated MOF composite	Intermediate
Enclosing MOF name	Intermediate
Name of the linker salt used for incorporation	Intermediate
Name of the metal salt used for incorporation	Intermediate
Name of the incorporated enzyme	Intermediate
Organism name of the incorporated enzyme	Advanced
Concentration of the used metal salt solution	Advanced
Concentration of the used linker salt solution	Advanced
Concentration of the used enzyme solution	Advanced
Educts of the enzymatic reaction	Hard
Products of the enzymatic reaction	Hard
Sequence of components for the composite	Hard

2 Concept

Fig. 1 provides an overview of the presented concept. We extract and condense the information of a scientific work $\mathcal{W}$ (typically a paper) to extract a set of pre-defined attributes $\mathcal{A}$, specified by the human expert. Since LLMs cannot process texts of arbitrary lengths, finding a solution to process texts longer than the maximum input size is particularly essential. Hence, $\mathcal{W}$ is converted into plain text and split at section headings to obtain a list of self-contained, shorter texts $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_n\}$, where n is the number of sections in $\mathcal{W}$. Each element $\mathcal{T}_i$ is then compressed by an LLM only to retain information related to the attributes $\mathcal{A}$, resulting in a list of shortened sections $\mathcal{T}' = \{\mathcal{T}_1', \mathcal{T}_2', ..., \mathcal{T}_n'\}$, where $\mathcal{T}_i'$ represents a condensed version of $\mathcal{T}_i$ that is reduced to the sought information: $|\mathcal{T}_i'| < |\mathcal{T}_i|$, where $|\cdot|$ denotes the length. The compressed sections $\mathcal{T}'$, until now only considered in isolation, are now iteratively combined into one short and redundancy-free paragraph $\mathcal{P}$, that contains all relevant information regarding the defined attributes $\mathcal{A}$. $\mathcal{P}$ is expanded iteratively, putting the individual sections of $\mathcal{T}'$ in context with each other. First, an initial summary $\mathcal{P}_1$ is generated from the first section $\mathcal{T}_1'$. For each subsequent section $\mathcal{T}_i'$, the previously generated summary $\mathcal{P}_{i-1}$ is used as a knowledge base and expanded with $\mathcal{T}_i'$ to obtain $\mathcal{P}_i$. With this approach, the last paragraph $\mathcal{P}_n$ encompasses a summary of the entire content of $\mathcal{W}$. Using our approach, the final paragraph $\mathcal{P}_n$ effectively summarizes the entirety of $\mathcal{W}$, with a focus on the predefined attributes $\mathcal{A}$ and processible at once by common LLMs. In the last step, $\mathcal{P}_n$ is used to first extract the attributes $\mathcal{A}$ and then output them in a structured format so that they can be processed automatically by other applications.

3 MOF Parameter Extraction

Our concept is exemplified for the extraction of MOF parameters but is applicable to arbitrary applications (e.g. the extraction of small molecule potencies in oncology).

Definition of Structured Information

The experimental attributes $\mathcal{A}$ to be extracted from the scientific papers $\mathcal{W}$ are displayed in Tab. 1. We have identified these parameters as important since they are highly relevant for the classification, formulation, or characterization of the described biohybrid enzyme-MOF composites. The first five parameters are the most accessible due to their key roles and prominent semantic positions in the text. The *organism name* and *concentrations* are more complex as they require not only the localization and extraction of the corresponding parameter but also the associated attributes. The final three parameters are challenging due to their complex definition. The LLM must make multiple distinctions for the parameters *educts* and *products*. This includes identifying reactants or products and their linkage to enzyme catalysis reactions while excluding metal ions and linker molecules of the enzyme incorporation into MOF reaction that is also linked to the same enzymes. The *sequence of components* requires an understanding of the time and process of enzyme incorporation into MOF.

Dataset

To evaluate our concept, we generate a dataset in which we ask a domain expert to manually extract the defined MOF parameters $\mathcal{A}$ from four scientific papers conducting experiments involving MOFs. Each paper stems from a different journal to ensure diversity in the data. Further, the papers contain a main text and may have Supplementary Information. Not every sought attribute is necessarily mentioned in every paper. Further details regarding the dataset are provided at our repository: https://osf.io/nhegv/.

Experiment Design

We evaluate our compression approach regarding extraction quality of the defined attributes against an LLM model capable of processing very long inputs [7]. The comparison model receives the same prompt for generating the attributes as our method in the last step. Since there may be deviations in the generated answers, which are not necessarily incorrect, a human expert is consulted to assess whether the found parameters are valid or not. Further, we report the compression rate of our model from the initial scientific work parsed with OCR to the most compressed version to evaluate how much the text has been condensed.

4 Results

Architecture and Implementation

We employ nougat [2] for OCR parsing. We do not modify the generated text, except by removing the references and author lists by hand. Sections too long to be fed into our LLM model are split in half. A Llama-2 [10] model with 70 billion parameters is used for both the scientific text compression, as well as for the extraction of the attributes from the compressed paragraph $\mathcal{P}_n$. The long-input model we compare our approach to is Yarn-Llama-2-13b-128k [7], a modified Llama-2 trained for long inputs. We optimized the output of the models via prompt-engineering. The final prompts used are accessible at https://osf.io/nhegv/. All models are implemented in PyTorch with the use of the HuggingFace [12] platform. The text generation is performed on a system with two NVIDIA GeForce RTX A6000 GPUs.

Experiments

Tab. 2: The accuracy of our *Compression* approach compared to a *Direct* [7] evaluation of the scientific texts. The accuracy is calculated as the number of correct predictions against the number of scientific papers.

Attribute $\mathcal{A}$	**Ours [%]**	**Direct [%]**
Name of the formulated MOF composite	100	75
Enclosing MOF Name	100	75
Name of the linker salt used for incorporation	100	0
Name of the metal salt used for incorporation	75	0
Name of the incorporated enzyme	100	75
Organism name of the incorporated enzyme	100	100
Concentration of the used metal salt solution	75	25
Concentration of the used linker salt solution	50	50
Concentration of the used enzyme solution	50	50
Educts of the enzymatic reaction	50	25
Products of the enzymatic reaction	100	0
Sequence of components for the composite	50	0

Tab. 3: The lengths of the *Intial* texts, the *Compressed Sections*, and the *Dense Paragrahs* for each scientific paper. All lengths are given as the sum of words.

Paper Num.	**Length Original**	**Length Compressed Sections**	**Length Dense Paragraphs**
1	3.767	2.335	538
2	6.229	2.921	474
3	4.920	3.124	897
4	5.043	3.527	756

Tab. 2 lists the accuracy for the sought attributes for both our compression method (*Ours*) and the comparison model [7] (*Direct*). Our model achieves high accuracy for *intermediate* attributes, with only *metal salt* presenting occasional errors. The result for the comparative model regarding the *intermediate* attributes is strikingly less precise. No attribute was always found and neither the the *linker salts* nor the *metal salts* could be determined in any paper. The name of the *incorporated enzymes* could always be found by our method and in 75% of cases for the comparison model. However, the other *advanced* attributes, particularly *concentration* determinations, pose challenges for both methods. Our approach demonstrates superiority, achieving 75% accuracy for *metal salt solutions* and 50% for *linker salt solutions* and *enzyme concentrations*. In contrast, the comparison model determines 25% of the *metal salt solutions* correctly, and 50% of the *enzyme* and *linker salt solutions*. The *educts* are also difficult to determine, but here our method with 50% accuracy still has an advantage over the comparison method with 25%. Remarkably, our approach consistently identifies *enzymatic reaction products*, a task the comparison model could never do correctly. Finally, our model successfully identifies the *sequence of components* in 50% of cases, whereas the comparison model again fails entirely.

Tab. 3 displays the compression rates for both steps of our concept. The results show a substantial reduction in the length of the texts. Initially, the papers had varying lengths, ranging from 3.767 to 6.229 words. After compression, the lengths of the sections were notably reduced, with the compressed sections containing between 2.335 and 3.527 words. Moreover, the dense paragraphs, the final most reduced text, demonstrate significant condensation, with lengths ranging from 474 to 897 words. The compression rate results combined with the superior identification of the relevant attributes compared to direct extraction show that our approach is capable of distilling the relevant information regarding the searched attributes in a minimum of words. In addition, our results illustrate that current LLMs are not suitable for processing very long texts to extract information, even if this is technically possible.

5 Discussion

Our presented concept is a leap in the direction of automated retrieval of structured information from scientific papers. This is especially valuable in the biomedical field, where domain experts have little time to spare, and searching for specific attributes in texts may require a considerable amount of time. The superiority of our method over an approach optimized for long texts that processes the whole scientific works at once, confirms the potential of our approach, even though some parameters are not identified correctly. However, this result is to be expected, as we deliberately chose the challenging task of MOF parameter retrieval including attributes of different difficulty levels to demonstrate the potential and limitations of our method in extracting structured information. Although our system cannot yet be used completely autonomously, the workload for domain experts searching for information in scientific texts is already significantly reduced using our approach, since most attributes can be found with substantially less effort, as we were able to show.

6 Conclusion

We recognize the problem that extracting structured information from scientific works is labor-intensive and error-prone. Hence, we build upon the recent developments in the field of LLMs and present a concept that automatically distills predefined attributes from lengthy scientific texts. In contrast to the evident solution to increase the amount of text processable by the LLM, we iteratively condense the contained information to break down complex texts to the most significant information regarding the respective challenge. Thus, we can employ the condensed text to well-established methods and exploit the full potential of state-of-the-art technologies. Our experiments confirm the superiority of our method over an approach optimized for long texts that processes the whole scientific work at once.

Potential enhancements to our prototypical concept include fine-tuning network parameters, employing human-in-the-loop approaches, or providing biomedical databases using techniques like Retrieval-Augmented Generation (RAG) when the model is uncertain. Expanding our method to more complex datasets will assess its effectiveness in handling a broader range of scientific literature. Further, subsequent steps may involve developing our experimental findings into a fully employable product. This may include engaging multiple domain experts in concept evaluation and determining the extent to which they find the extracted MOF attributes useful, helping to refine and optimize the information extraction process.

References

[1] An Y, et al. Exploring pre-trained language models to build knowledge graph for metal-organic frameworks (mofs). In: IEEE International Conference on Big Data IEEE2022, 3651–3658.

[2] Blecher L, et al. Nougat: Neural optical understanding for academic documents 2023.

[3] Dagdelen J, et al. Structured information extraction from scientific text with large language models. *Nature Communications* 2024;15:1418.

[4] Grisoni F. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology* 2023;79:102527.

[5] Jablonka KM, et al. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* 2024;:1–9.

[6] Luo Y, et al. Mof synthesis prediction enabled by automatic data mining and machine learning. *Angewandte Chemie International Edition* 2022;61:e202200242.

[7] Peng B, et al. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:230900071* 2023;.

[8] Thapa S, et al. Chatgpt, bard, and large language models for biomedical research: opportunities and pitfalls. *Annals of biomedical engineering* 2023;51:2647–2651.

[9] Thirunavukarasu AJ, et al. Large language models in medicine. *Nature medicine* 2023;29:1930–1940.

[10] Touvron H, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:230709288* 2023;.

[11] Wang X, et al. Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:240202244* 2024;.

[12] Wolf T, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:191003771* 2019;.

[13] Yang J, et al. Metal–organic frameworks for biomedical applications. *Small* 2020;16:1906846.

[14] Zheng Z, et al. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. *Journal of the American Chemical Society* 2023;145:18048–18062.