ISSN:1738-6764, eISSN:2093-7504

SiCExtractor: A Retrieval-Augmented Generation Pipeline for PVT-Based SiC Crystal Growth Research

Fatemeh Pishkool [1], Hyoseok Oh [2], Ga-Ae Ryu [3], Jinhwa Park [4] and Kwan-Hee Yoo [5,*]

1 School of Computer Science, Chungbuk National University; Pishkool.Fateme@gmail.com
2 School of Computer Science, Chungbuk National University; hyosk5@chungbuk.ac.kr
3 Korea Institute of Ceramic Engineering and Technology; garyu@kicet.re.kr
4 Korea Institute of Ceramic Engineering and Technology; pjh@kicet.re.kr
5 School of Computer Science, Chungbuk National University; khyoo@chungbuk.ac.kr
* Correspondence

https://doi.org/10.5392/IJoC.2025.21.4.107

Manuscript Received 11 July 2025; Received 28 October 2025; Accepted 28 October 2025

Abstract: *In this study, we present a Retrieval-Augmented Generation (RAG)-based pipeline designed to extract key values from scientific literature on Silicon Carbide (SiC) crystal growth using the Physical Vapor Transport (PVT) method. To improve the relevance and completeness of the retrieved context, we implemented a hybrid retrieval strategy that combines dense retrieval via FAISS with sparse retrieval using BM25. We employed two distinct prompting approaches for key value extraction. The first approach addresses interactive user queries by utilizing the retrieved context to generate informed responses. The second approach, intended for bulk extraction, follows a two-step process: a binary classification prompt first checks for the presence of relevant information related to a query. If relevant information is confirmed, a subsequent prompt extracts the value under strict constraints—requiring exact phrasing without guessing or explanation. This binary pre-check significantly enhances the identification of true negative cases, thereby reducing irrelevant or missing data. For the generative component of our pipeline, we evaluated three large language models (LLMs): Llama 8B, Gemma 7B, and Mistral 7B, all operating on a local multi-GPU environment using FP16 precision. The results reveal differences in the efficiency of these models within our customized RAG system, particularly in their performance in extracting over 156 targeted technical key-value pairs from 13 benchmark papers.*

Keywords: Large Language Model; Retrieval Augmented Generation; Key-value Extraction; Prompt; SiC Crystal Growth Research

1. Introduction

Silicon Carbide (SiC) has gained significant attention in materials science research and industry because of its unique properties, including high thermal stability, excellent electrical conductivity, and remarkable mechanical hardness. These characteristics make SiC an ideal candidate for power electronics, optoelectronic devices, and systems operating in harsh environments. Among the various research areas related to SiC, crystal growth has emerged as a particularly critical field, as it directly influences the material's quality and performance. Within crystal growth techniques, Physical Vapor Transport (PVT) stands out as one of the most widely studied and commercially adopted methods for growing high-quality bulk SiC crystals. PVT growth parameters such as temperature, pressure, growth rate, and doping concentration play a pivotal role in determining the quality and characteristics of the resulting crystals [1]. As a result, many scientific papers have been published focusing specifically on the PVT method, each containing valuable data on growth conditions and outcomes.

Given the vast volume of these publications, manually extracting key values using traditional methods is time-consuming and challenging. To address this issue, we developed an intelligent automated pipeline for processing PDFs and extracting data from materials science papers related to SiC crystal growth, focusing on PVT-based research. Subsequently, large language models (LLMs) like GPT and Llama have demonstrated strong performance in understanding technical language and answering complex questions. With advancements in artificial intelligence, LLMs have become powerful tools for organizations and research institutions seeking to analyze large volumes of unstructured text and develop customized intelligent systems. To extract key values from PDF papers, we utilized the RAG (Retrieval Augmented Generation) architecture to extract approximately 156 key values from 115 research papers on PVT-based crystal growth, among which 13 papers were reserved for evaluation to assess extraction accuracy.

Motivated by the need for automated key-value extraction from research papers on SiC crystal growth, we have developed a web-based chatbot system. This tool enables researchers to upload newly published documents and extract key values through a Question Answering (QA) framework. Furthermore, the system supports batch processing by allowing users to download a CSV file containing the model's responses to a predefined set of keyword-related queries. This functionality facilitates efficient, large-scale extraction of relevant information without requiring users to input individual queries manually. At the core of our system is the RAG architecture, which combines the strengths of information retrieval and language generation. This architecture enables the integration of domain-specific knowledge from scientific publications without requiring retraining of the language model. Retrieved passages from relevant documents are used as contextual input to guide the LLM in generating precise and context-aware responses. We incorporated dense (FAISS-based) and keyword-based (BM25) retrieval methods to ensure comprehensive retrieval of all relevant chunks, ensuring that a wide range of relevant information is captured. Furthermore, prompts are carefully engineered to align with the scientific context of SiC crystal growth via PVT, improving the quality and specificity of the extracted values.

2. Related Works

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm that combines external knowledge retrieval with language model generation, providing a flexible alternative to traditional fine-tuning. The work by [RAG vs Fine-Tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture] highlights RAG's practical strengths in domains where data is dynamic, sparse, or not well represented in pretraining corpora. Their study emphasizes the modularity and cost-effectiveness of RAG, particularly in domain-specific applications like agriculture, where fine-tuning can be costly and less generalizable [2].

Inspired by this approach, our system SiCExtractor adopts a RAG framework to manage key value extraction from SiC research papers. We enable context-aware information extraction without fine-tuning.

2.2 Retriever

The paper [Predicting Efficiency/Effectiveness Trade-offs for Dense vs. Sparse Retrieval Strategy Selection] examines the key differences between dense and sparse retrieval methods. Dense retrieval uses semantic embeddings to capture contextual meaning more effectively, while sparse retrieval techniques, such as BM25, are better suited for keyword-based queries and the structured language often found in domain-specific texts. The authors also propose a hybrid retrieval strategy that combines both dense and sparse methods, incorporating a classifier that selects the best retrieval approach; sparse, dense, or hybrid, on a per-query basis. This dynamic selection framework aims to optimize computational efficiency and retrieval performance, reducing latency and costs while ensuring effectiveness [3].

Inspired by this approach, our system, SiCExtractor, employs a hybrid retrieval strategy that integrates dense retrieval (using FAISS) with sparse retrieval (utilizing BM25). This fixed hybrid method allows SiCExtractor to effectively manage the diverse and complex language of scientific literature in the SiC domain. By combining both semantic and keyword-based matching, the system enhances its ability to retrieve and process relevant information from unstructured scientific documents without relying solely on one retrieval method.

2.2 Prompt

The study [Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering] demonstrates the critical role of prompt design in enabling large language models (LLMs) to extract structured scientific information. Their work focuses on retrieving well-defined material properties such as temperature, pressure, and synthesis conditions, showing that carefully engineered prompts can significantly improve extraction accuracy in specialized domains [4].

While their approach effectively targets standard parameter extraction, the goal of our system, SiCExtractor, extends beyond retrieving isolated physical values. Our task involves extracting instrument-specific terms, compound names, variations in materials, and context-dependent qualifiers, which require a deeper understanding of domain-specific language and document structure. Due to the complexity of this broader objective, applying a fully optimized prompt strategy remains a challenge. As a result, we are currently adopting a simplified prompting approach, with plans to refine and expand it in future work to better handle the nuanced demands of SiC-related scientific literature.

3. Materials and Methods

In this work, we utilized the RAG (Retrieval-Augmented Generation) architecture to develop a pipeline for extracting data from approximately 115 English-language research papers focused on SiC crystal growth using the PVT method, with 13 of these papers designated as the evaluation set for performance assessment. Figure 1 illustrates the overall architecture of our RAG-based pipeline. In the initial stage of our pipeline, we followed the RAG framework by extracting text from PDF documents and dividing it into smaller segments, referred to as chunks. For each chunk, we generated embeddings using the hidden state representations from a pre-trained large language model (LLM), capturing rich semantic information. These embeddings were then stored in a FAISS vector database to enable efficient semantic retrieval during generation. We deliberately chose not to use external or dedicated embedding models, as our goal was to evaluate and compare the performance and efficiency of three different LLMs within our customized RAG setup, focusing on their suitability for our specific domain. To evaluate the performance of LLMs in our system, we selected 13 papers as ground truth [5-18]. We examined around 156 entities in these papers, and if the data existed, we manually extracted it, using these values as our ground truth for validation purposes.

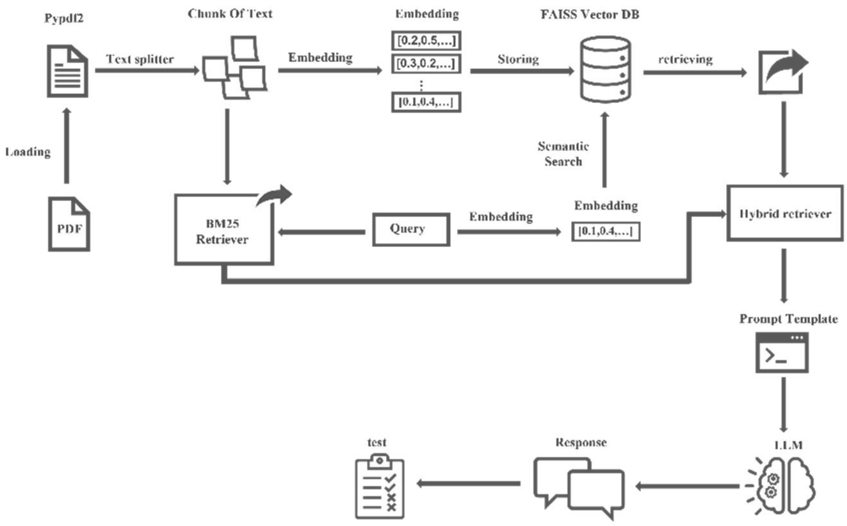


Figure 1. Architecture of SiCExtractor based on a RAG pipeline. Parsed PDF text is chunked, embedded, and stored in FAISS (dense) and BM25 (sparse) indexes. Retrieved chunks from both methods are combined and used by the LLM to generate context-aware, domain-specific responses.

3.1 Retrievers

We implemented a hybrid query-based retriever to extract relevant information in response to user queries. First, we employed vector-based retrieval using FAISS to identify the most semantically similar chunks. Second, to enhance recall and ensure that critical information was not overlooked, we performed parallel keyword-based retrieval using the BM25 algorithm to retrieve additional textually relevant chunks. This hybrid query-based retriever strategy produced a combined set of candidate chunks in a hybrid query-based retriever. The union of these chunks was then provided to a language model, along with a domain-specific prompt designed to facilitate accurate data extraction from scientific articles focused on SiC crystal growth via the PVT method.

To define the hybrid retrieval process formally, let $C_{FAISS} = \{c_1, c_2, c_3, ...\}$ denote the set of chunks retrieved using the FAISS-based dense retrieval approach, where semantic similarity between the query and document embeddings is used. Similarly, let $C_{BM25} = \{c_1, c_2, c_3, ...\}$ represents the set of chunks retrieved using the BM25 sparse retrieval method, which relies on keyword matching and term frequency-based scoring.

The final set of context chunks used for downstream response generation is computed as the union of these two sets:

$$C_{hybrid} = (C_{FAISS} \cup C_{BM25})[:k]$$
$$C_{hybrid} \leq k(k = 5) \quad (1)$$

This process ensures that the most semantically relevant and textually matched chunks are selected for context, effectively minimizing redundancy while maximizing the quality and diversity of information available for model inference. By merging the results from both retrieval methods, FAISS for dense semantic matching and BM25 for keyword-based retrieval, we ensure that the most relevant chunks are captured by the retriever, providing a comprehensive and well-grounded context for the generation stage.

The combined set of context chunks is denoted as C_{hybrid} and this final set is then input into the language model. This approach allows the model to accurately extract specific data points related to SiC crystal growth within the PVT method, ensuring high-quality, context-aware responses.

3.1.1 FAISS

In this research, we utilize the FAISS (Facebook AI Similarity Search) library to perform efficient similarity-based retrieval of dense text embeddings, which forms our Retrieval-Augmented Generation (RAG) pipeline for PVT-based SiC research papers. FAISS is an open-source, powerful tool that facilitates k-nearest neighbor search (KNN) in high-dimensional vector spaces. It provides a variety of indexing structures optimized for balancing retrieval speed, accuracy, and memory consumption. Some of these indexing structures include:

- Brute-force search using the IndexFlat index
- Approximate Nearest Neighbor (ANN) indexes
- Search-time tuning and optimization (e.g., IndexRefine)
- Customizable distance metrics (L2, Inner Product, etc.)

In this project, we selected the IndexFlatL2 index, which performs an exhaustive brute-force search based on the L2 (Euclidean) distance metric. In addition to FAISS, we transform embeddings into the cosine similarity metric for enhanced performance [18]. The embeddings for x and y are transformed using the following equations:

$$x' = x/||x||, \; y' = y/||y|| \quad (2)$$

Where:

- x, y denotes original embedding vectors
- $||x||, ||y||$ are L2 norms (Euclidean norms) of the vectors
- x', y' are normalized vectors

These transformations aid in modifying the vectors for a more precise cosine-based comparison, enhancing retrieval quality by aligning them with the geometric characteristics of the data.

3.1.2 BM25

In our system, BM25 (Best Matching 25) serves as a sparse retrieval mechanism that selects relevant text segments based on keyword-level matching. BM25 calculates a relevance score for each chunk as a probabilistic ranking function by combining term frequency (TF), inverse document frequency (IDF), and a length normalization factor [18]. The BM25 score for a document d and a query q is computed as:

$$BM25(d,q) = \sum_{t \in q} \mathrm{IDF}(t) \cdot \frac{\mathrm{TF}(t,d) \cdot (k_1 + 1)}{\mathrm{TF}(t,d) + k_1 \cdot (1 - b + b \cdot \frac{\mathrm{len}(d)}{\mathrm{avgdl}})} \quad (3)$$

$$\text{where } \mathrm{IDF}(t) = \log \frac{N - n(t) + 0.5}{n(t) + 0.5}$$

Where:

- t denotes a lexicon term in q
- $\mathrm{TF}(t, d)$ is the term frequency of t in document d
- $\mathrm{IDF}(t)$ is the inverse document frequency of term t
- k_1 controls term frequency saturation
- b controls document length normalization
- N is the total number of documents
- $n(t)$ is the number of documents containing term t

This method is well-suited for capturing exact or near-exact term matches from user queries. By leveraging the statistical properties of word occurrences within and across documents, BM25 enables efficient retrieval of text chunks that are highly relevant in terms of lexical overlap, making it a strong baseline for identifying explicitly stated information.

3.2 Prompt Design

Our pipeline employs two distinct prompt strategies to interact with a large language model for different purposes: first, interactive user queries; second, bulk automated extraction. Each strategy is optimized to ensure precision, context-awareness, and adherence to the document's content.

3.2.1 Interactive Prompt for User Queries

For direct user interaction, we designed a conversational prompt that guides the model to extract and return information clearly from a given PVT-based SiC crystal growth document context. This context is composed of the most relevant text chunks retrieved using a hybrid retrieval approach. The prompt includes the user's question, the retrieved context, and conversation history (if any), and instructs the model to answer using only this information as mentioned in Figure 2. It emphasizes clarity, completeness (listing multiple values when available), and caution, asking the model to admit when the answer is not found. This helps ensure trustworthy, context-grounded responses without hallucination.

```
prompt = (
"You're a helpful assistant trained to extract technical details from scientific documents about SiC material research.\n"
"Your job is to help the user by answering their question clearly and accurately using only the information from the provided context.\n"
"• Include units if available.\n"
"• If there are multiple relevant values, mention all of them clearly, separated by commas.\n"
"• If the answer is not found in the context, politely say you couldn't find it.\n\n"
f"Conversation History:\n{conversation_history_str}\n\n"
f"Context:\n{context_str}\n\n"
f"User Question:\n{user_message}\n\n"
```

Figure 2. This is a figure that shows the Interactive Prompt for User Queries

3.2.2 Bulk Extraction Prompt

We implement a two-step prompting strategy, when determining if a valid answer is absent. In the first step, a binary classification prompt asks the LLM to decide whether the retrieved context contains relevant information for the given query. This pre-check improves the system's ability to identify true negative (TN) cases by filtering out empty or irrelevant results (the prompt is shown in Figure 3.a). If the model responds with "yes," it proceeds to a second prompt focused strictly on value extraction. This step enforces strict rules: no guessing, no explanation, and exact phrasing (the prompt is shown in Figure 3.b). Both steps operate solely on the hybrid-retrieved chunks, ensuring that the final answers are grounded in the most relevant content from the documents.

3.3 Generator

```
yesno_prompt = (
"You are given a document context and a technical key.\n"
"Determine whether a value or specific information related to the following key exists specifically in the category.\n"
f"Key: {key}\n"
f"Category: {category}\n"
f"Context: {context}\n\n"
f"Answer only 'yes' or 'no'.\n"
f"Answer:"
)
```

(a)

```
extract_prompt = (
    "You are given a document context and a technical question.\n"
    "Answer using ONLY the exact values or names explicitly written in the context.\n\n"
    "Requirements:\n"
    "- DO NOT explain or rephrase.\n"
    "- DO NOT add anything that is not in the context.\n"
    "- DO NOT guess.\n"
    "- If you find multiple relevant values, separate them with commas.\n"
    "- If no answer is present, reply with 'no'.\n\n"
    f"Context: {context}\n"
    f"User Question: {user_question}\n"
    "Answer:"
)
```

(b)

Figure 3. This is a figure that shows the bulk automated extraction prompt of SiCExtractor. This prompt includes two steps: **(a)** binary classification; **(b)** value extraction.

We deployed multiple large language models (LLMs), including Llama3.1 8B, Gemma 7B, and Mistral 7B, locally on a system equipped with three NVIDIA GeForce RTX™ 2080 Ti GPUs. These models were integrated into our pipeline as both generators and embedding models. Specifically, we utilized their hidden state representations to generate semantically rich embeddings for document chunks, which were then used in the retrieval component of our architecture. Due to hardware limitations and restricted GPU memory (VRAM), all models were operated in half-precision (FP16) mode instead of full-precision (FP32) or more advanced quantization techniques. Operating in FP16 allowed us to balance model fidelity and memory efficiency, enabling deployment on our available three NVIDIA GeForce RTX™ 2080 Ti GPUs. This consistent precision and hardware configuration ensured a fair and comparable evaluation of each model's performance across the pipeline.

4. Results

The performance of three LLMs, Gemma 7B, Llama 8B, and Mistral 7B, was evaluated for their ability to determine whether a key value is present in the retrieved text and to extract the corresponding value accurately. The evaluation was conducted exclusively using the bulk extraction prompt, as the user interaction prompt was found unsuitable for automated extraction tasks due to its reliance on dynamic user input and less structured response behavior. The focus of the evaluation was on the accuracy and reliability of key value extraction, utilizing metrics derived from confusion matrix elements, including true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), along with standard evaluation metrics such as accuracy, precision, recall, specificity, and error rate.

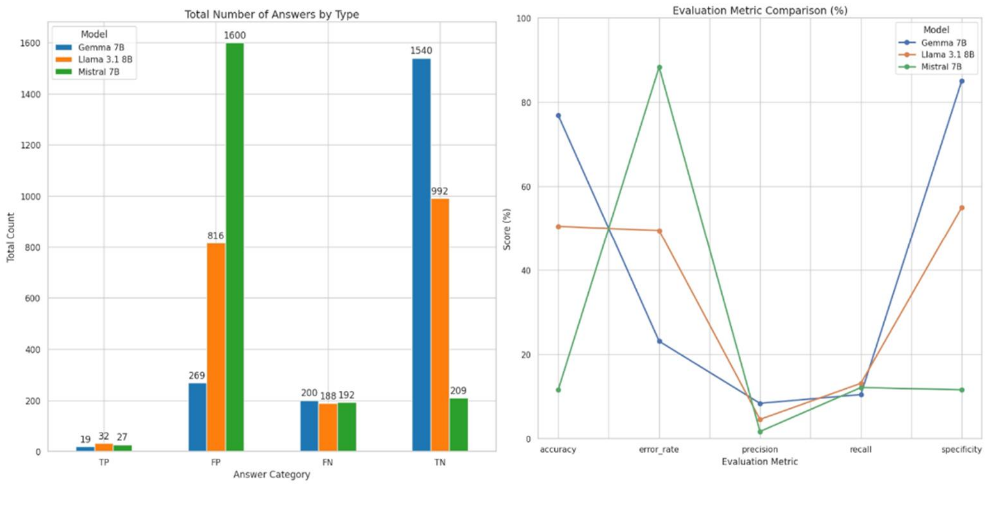
Across the 13 papers, Gemma 7B demonstrated the most conservative and accurate extraction behavior, as shown in Figure 4.a. It recorded the highest number of true negatives (1540), indicating strong performance in rejecting incorrect key values, and maintained the lowest number of false positives (19), reflecting a cautious approach in extracting only well-supported values. Llama 8B exhibited a more balanced behavior, with 992 true negatives and 816 false positives, indicating a moderate extraction tendency with a reasonable level of correctness. In contrast, Mistral 7B exhibited an overgeneration pattern, producing 1600 false positives, the highest among the models, and only 209 true negatives, revealing lower precision in its extraction strategy. Notably, all three models had similarly low true positives (19–32) and high false negatives (188–200), suggesting a shared challenge in accurately identifying valid key values during bulk extraction.

The evaluation metrics supported these observations. As detailed in Table 1, Gemma 7B achieved the highest accuracy at 76.9%, followed by Llama 8B at 50.5%, while Mistral 7B scored the lowest at only 11.6%. In terms of error rate, Mistral 7B recorded the highest at 88.4%, whereas Gemma 7B maintained the lowest at 23.1%, as illustrated in Figure 4.b. Precision was generally low across all models: 8.4% for Gemma 7B, 4.6% for Llama 8B, and 1.7% for Mistral 7B. Recall values were similarly low: 10.5% for Gemma 7B, 13.1% for Llama 8B, and 12.1% for Mistral 7B, which underlines the shared challenge in accurately identifying valid key values using the bulk prompt. Specificity, which measures the model's ability to reject incorrect extractions, was highest for Gemma 7B at 58.1%, followed by Llama 8B at 50.01%, and Mistral 7B at only 11.6%, reinforcing Gemma's strength in conservative yet reliable extraction.

In summary, when tested on key value extraction across 13 research papers, Gemma 7B outperformed the other models with its cautious and precise approach. Llama 8B offered moderate performance with a trade-off between extraction and correctness. Mistral 7B, while aggressive in identifying potential values, suffered significantly in precision and overall reliability due to a high rate of incorrect extractions.

Table 1. Evaluation metrics for three LLMs (Gemma 7B, Llama 8B, and Mistral 7B) on the bulk extraction task

	Gemma 7B	**Llama 8B**	**Mistral 7b**
Accuracy	**76.9%**	50.5%	11.6%
Error Rate	**23.1%**	49.5%	88.4%
Precision	**8.4%**	4.6%	1.7%
Recall	10.5%	**13.1%**	12.1%
Specificity	**58.1%**	50.01%	11.6%



(a) (b)

Figure 4. (a) Confusion matrix analysis of key value extraction for Gemma 7B, Llama 8B, and Mistral 7B across 13 scientific papers, showing True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). **(b)** Evaluation metrics include Accuracy, Error Rate, Precision, Recall, and Specificity.

4. Discussion and Future Works

Despite progress in our RAG pipeline, the True Positive (TP) rate remains low, primarily because the current model is not fine-tuned for the specific domain of SiC crystal growth. Domain-specific terminology, complex relationships between key-values, and hierarchical categorical structures make extraction challenging. The dataset includes five main categories: Chemical Information, Component, Process, Property, and

Performance, each with multiple sub-categories. Some keys, such as "Pressure," appear in multiple sub-categories but must be interpreted independently. As a result, the model occasionally produces correct content but assigns it to the wrong category, reducing TP rates and overall extraction accuracy.

Although Gemma 7B achieved a relatively high overall accuracy of 76.9%, this primarily reflects correct rejection of incorrect key-values rather than complete identification of valid key-values. The low precision and recall indicate that the model often misses or misassigns valid key-values, highlighting an opportunity to enhance precision and recall in future work.

The small model size (7–8B parameters) further limits the ability to understand subtle, domain-specific patterns. Additionally, current chunking may split important contextual information across segments, weakening the retriever's and model's performance.

Future work will focus on structured, multi-step prompts, adaptive chunking to preserve context, and domain-specific fine-tuning. These improvements are expected to significantly enhance recall and precision, ensuring more accurate and complete extraction of information from complex scientific literature.

5. Conclusions

In this work, we presented a domain-specific API system for extracting key scientific values from SiC crystal growth papers using a Retrieval-Augmented Generation (RAG) architecture. The system was designed to handle complex technical content and focused on retrieving and generating answers to structured queries related to material properties, growth conditions, and experimental parameters. To enhance retrieval accuracy and ensure high-quality context for generation, we implemented a hybrid retriever that combines sparse (BM25) and dense (FAISS) methods. This dual approach enables the system to leverage both exact keyword matching and semantic similarity, thereby improving the relevance of retrieved passages from parsed text files. We evaluated our system using three different large language models, comparing their performance in accurately extracting structured information from SiC research papers. The comparative results helped identify the most effective model for our domain-specific extraction tasks based on consistency, contextual understanding, and value accuracy.

Acknowledgments: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2024-00450652, "Establishment of a Wide Bandgap Semiconductor Material HUB and Development of Innovative Material Technology Utilizing Process Informatics (PI)-based Database and AI Technologies").

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] A. Balaguer, V. Benara, R. Cunha, R. Estevão, T. Hendry, D. Holstein, J. Marsman, N. Mecklenburg, S. Malvar, L. O. Nunes, R. Padilha, M. Sharp, B. Silva, S. Sharma, V. Aski, and R. Chandra, "RAG vs Fine-Tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture," arXiv:2401.08406v3 [cs.CL], Jan. 2024. [Online] Available: https://arxiv.org/abs/2401.08406v3

[2] N. Arabzadeh, X. Yan, and C. L. A. Clarke, "Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection," in Proc. 30th ACM Int. Conf. on Information and Knowledge Management (CIKM '21), Virtual Event, QLD, Australia, pp. 1-5, ACM, New York, NY, USA, Nov. 1-5, 2021, doi: https://doi.org/10.1145/3459637.3482159.

[3] M. P. Polak and D. Morgan, "Extracting accurate materials data from research papers with conversational language models and prompt engineering," Nature Communication, vol. 15, no. 1, p. 1569, Feb. 2024, doi: https://doi.org/10.1038/s41467-024-45914-8.

[4] P. J. Wellmann, "Review of SiC crystal growth technology," Semicond. Sci. Technol., vol. 33, no. 10, p. 103001, 2018.

[5] C. Wang, Y. Zhang, Y. Zhang, Y. Wang, and D. Xu, "A compensation mechanism for semi insulating 6H-SiC doped with vanadium," Journal of Semiconductors, vol. 29, no. 2, pp. 206-209, Feb. 2008.

[6] J. Steiner, M. Roder, B. D. Nguyen, S. Sandfeld, A. Danilewsky, and P. J. Wellmann, "Analysis of the basal plane dislocation density and thermomechanical stress during 100 mm PVT growth of 4H-SiC," Materials, vol. 12, no. 13, pp. 1-14, Jul. 2019, doi: https://doi.org/10.3390/ma12132207.

[7] J. W. Choi, C. H. Son, J. M. Choi, G. S. Lee, W. J. Lee, I. S. Kim, B. C. Shin, and K. R. Ku, "Initial stage modification for 6H-SiC single crystal grown by the physical vapor transport (PVT) method," Mater. Sci. Forum, vol. 615-617, pp. 7-10, 2009, doi: https://doi.org/10.4028/www.scientific.net/MSF.615-617.7.

[8] B. Gao and K. Kakimoto, "Effect of the inclusion of transparency on the thermal field and interface shape in long-term sublimation growth of SiC crystals," J. Jpn. Assoc. Crystal Growth, vol. 40, no. 1, pp. 20-24, 2013.

[9] K. Grasza and E. Tymicki, "Application of 6H to 4H polytype conversion to effective reduction of micropipes in 4H SiC crystals," Mater. Sci. Forum, vol. 600-603, pp. 31-34, 2009, doi: https://doi.org/10.4028/www.scientific.net/MSF.600-603.31.

[10] P. Gao, J. Xin, X. Liu, Y. Zheng, and E. Shi, "Control of 4H polytype of SiC crystals by moving up the crucible to adjust the temperature field of the growth interface," CrystEngComm, vol. 21, no. 44, pp. 6787-6793, 2019, doi: https://doi.org/10.1039/C9CE01363F.

[11] H. Tsuge, S. Sato, M. Katsuno, T. Fujimoto, and W. Ohashi, "Growth of high quality 4H-SiC crystals in controlled temperature distributions of seed crystals," Mater. Sci. Forum, vol. 717-720, pp. 13-16, 2012, doi: https://doi.org/10.4028/www.scientific.net/MSF.717-720.13.

[12] C. Basceri, I. Khlebnikov, Y. Khlebnikov, P. Muzykov, M. Sharma, G. Stratiy, M. Silan, and C. Balkas, "Growth of micropipe-free single crystal silicon carbide (SiC) ingots via physical vapor transport (PVT)," Mater. Sci. Forum, vol. 527-529, pp. 39-42, 2006, doi: https://doi.org/10.4028/www.scientific.net/MSF.527-529.39.

[13] X. Xie, J. Yu, X. Yang, X. Chen, X. Xu, X. Hu, X. Liu, and D. Liu, "Investigation of carbon inclusions in SiC crystals grown by PVT method," Mater. Sci. Forum, vol. 1004, pp. 20-25, 2020, doi: https://doi.org/10.4028/www.scientific.net/MSF.1004.20.

[14] J. D. Blevins, A. K. Gupta, I. Zwieback, E. Emorhokpor, A. Souzis, T. Anderson, and C. Avvisato, "Development of a manufacturing process for large diameter semi-insulating silicon carbide substrates," in Proc. CS MANTECH Conf., Palm Springs, CA, USA, pp. 1-4, May. 16-19, 2011.

[15] M. Arzig, U. Künecke, M. Salamon, N. Uhlmann, and P. J. Wellmann, "Analysis of the morphology of the growth interface as a function of the gas phase composition during the PVT growth of silicon carbide," Mater. Sci. Forum, vol. 1062, pp. 89-93, 2022, doi: https://doi.org/10.4028/p-f58944.

[16] J. W. Choi, J. G. Kim, B. K. Jang, S. K. Ko, M. O. Kyun, J. D. Seo, K. R. Ku, C. Y. Lee, and W. J. Lee, "Polytype control by pretreatment of SiC source powder for 4H-SiC single crystal growth," Materials Science Forum, vol. 963, pp. 38-41, Jul. 2019, doi: https://doi.org/10.4028/www.scientific.net/MSF.963.38.

[17] M. Douze, J. Johnson, M. Lomeli, A. Guzhva, G. Szilvasy, L. Hosseini, C. Deng, P. E. Mazaré, and H. Jégou, "The FAISS library," arXiv preprint arXiv:2401.08281v3, vol. [cs.LG], Feb. 2025. [Online] Available: https://arxiv.org/abs/2401.08281

[18] T. Shen, G. Long, X. Geng, C. Tao, T. Zhou, and D. Jiang, "Large language models are strong zero-shot retrievers," arXiv preprint arXiv:2304.14233v2, vol. [cs.CL], Aug. 2023. [Online] Available: https://arxiv.org/abs/2304.14233