Automating Knowledge Discovery in Material Science with RAG Framework

***Abstract*—The rapid growth of material science research has led to an unprecedented surge in the volume and complexity of scientific data, encompassing experimental results, computational simulations, theoretical models, and industrial applications. Traditional keyword-based search methods, while widely used, are increasingly insufficient for extracting meaningful insights from such heterogeneous and semantically rich datasets. Researchers often face challenges in locating relevant information hidden within vast repositories of publications, patents, and databases, slowing down the discovery of novel materials and the advancement of sustainable technologies. Recent developments in Artificial Intelligence (AI) and Natural Language Processing (NLP) offer promising avenues to overcome these challenges. In particular, Retrieval-Augmented Generation (RAG) frameworks combine the strengths of Large Language Models (LLMs) with powerful information retrieval mechanisms to deliver context-aware, semantically accurate responses. Unlike conventional retrieval systems that rely on surface-level keyword matches, RAG-based approaches leverage embeddings and source information to capture deeper semantic relationships across scientific literature and databases. To support this study, a domain-specific dataset was curated from unstructured PDF documents and preprocessed into structured text segments for training and evaluation. Experimental results show that the proposed RAG pipeline, powered by MatSciBERT embeddings, achieves strong performance with an overall accuracy of 0.784 and an F1-score of 0.858. These results demonstrate the effectiveness of the approach in automating semantic data retrieval in the materials science domain.**

***Index Terms*—RAG, Natural Language Processing, Semantic Search, Material Science, MatSciBERT, FAISS, Data Extraction**

I. Introduction

Material science is an multidisciplinary field dedicated for analyzing the structure, composition, properties, and applications of materials for functional and behavioral needs. It integrates concepts from chemistry, physics, and engineering to understand the intrinsic behavior of matter and to drive the development of new materials [1]. The field encompasses a wide range of materials, including metals, ceramics, polymers, composites, and semiconductors, which are central to designing and advancing modern technologies. With increasing industrial demands, there is a growing need for stable, high-performance, and advanced materials in sectors such as electronics, battery technologies, aerospace, and healthcare. The discovery and optimization of such materials often depend on analyzing large volumes of experimental data, computational simulations, and scientific literature [2].

However, much of this information exists in unstructured formats such as PDFs, which are difficult to analyze and manage automatically because they lack a consistent schema or organization. Unstructured documents typically contain lengthy paragraphs, irregular formatting, dense technical language, and varied use of headings or tables, making it challenging to search, index, and extract relevant content. In contrast, organized data such as databases, spreadsheets, or tabular datasets follows a rigid format with well-identified columns and therefore is easy to retrieve, store, and analyze. In a bid to fill this vacuum, higher-level methods from NLP and Machine Learning (ML) are increasingly becoming popular in a quest to transform unstructured information into structured datasets [3]. Such transformation is the cornerstone in constructing intelligent systems for efficient storage, retrieval, and analysis of scientific knowledge [4]. Traditional knowledge retrieval approaches using LLMs often fall short under the current circumstances. These models are dependent largely on the pre-knowledge bases and cannot consistently deal with highly specialized, unstructured documents such as material science PDFs. As a result, they may produce unverifiable or hallucinated responses.

To overcome these challenges, a RAG pipeline is proposed, and the general architecture of this approach is illustrated in Figure 1. It employs domain-specific embeddings, efficient retrieval, and document-backed context. At its core, the pipeline leverages MatSciBERT, a model pre-trained for materials science, to represent both user queries and document chunks as semantic vector embeddings. These embeddings encompass the sense of scientific terms and expressions, making sure of proper comprehension of specialist terms (e.g., ”synthesis,” ”element,” ”aspect,” and ”method”). The content of the PDFs is broken down into smaller, more manageable pieces of text, indexed by Facebook AI Similarity Search (FAISS), a library that allows for scalable and efficient similarity search through approximate nearest neighbor algorithms. At retrieval time, the query embedding is compared with the indexed document embeddings using the assistance of distance metrics such as cosine similarity or Euclidean distance, and it returns the most relevant chunks. This procedure provides fine-grained, context-aware matching even in big, complicated data sets. When the appropriate excerpts are identified, they are integrated with the user query and submitted to Gemini, a large language model, which produces a coherent and context-dependent output. In this process, the essay is not only flowing but is also aided with evidence from domain-related papers. Moreover, all the questions, retrieved excerpts, and outputs are logged for transparency and future analysis, and the system is reliable and auditable [5].

The following sections outline the key contributions of this paper:

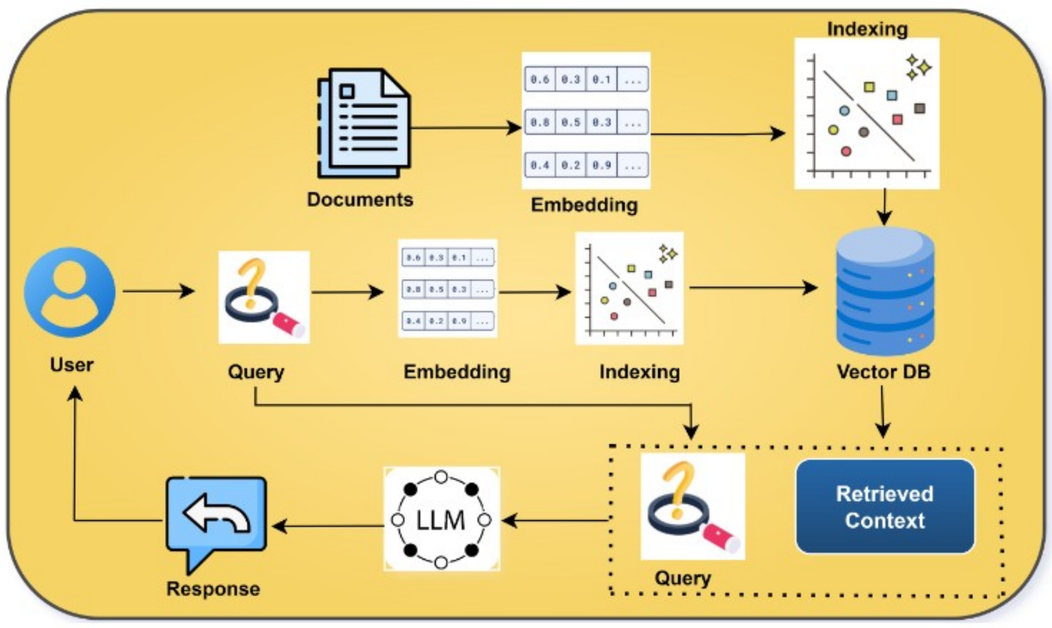

Fig. 1. RAG framework

• Unlike traditional LLM-only retrieval, which is limited by static pre-training and lacks document-supported knowledge, the proposed pipeline automates the process of searching, structuring, and reasoning over unstructured material science data [6].

• By combining semantic embeddings, vector indexing, and RAG, the system provides accurate, scalable, and domain-tailored knowledge retrieval.

• This makes it particularly effective for high-volume research tasks where scientists must navigate detailed information such as structural data, processing parameters, and performance characteristics. It also supports for downstream tasks by storing and organizing extracted information for long-term accessibility and trend analysis.

• Ultimately, the framework is capable of handling a wide variety of queries, from factual properties to conceptual insights which significantly reduces manual effort, accelerates knowledge discovery, and supports advanced research in material science.

This paper is organized as follows: Section 2 gives a comprehensive literature survey of NLP in the field of material science. Section 3 describes the research methodology proposed for retrieving contextually relevant information. The findings are presented in Section 4, which emphasizes the efficiency of the system. Lastly, Section 5 concludes with a summary of findings and recommendations for future work in RAG applications.

II. Literature Review

Researchers across the world have increasingly adopted RAG techniques for information retrieval from research papers, with the aim of accelerating scientific discovery. One such approach involves a specialized LLM-based text generator designed specifically to handle queries related to metal additive manufacturing (AM), known as AMGPT [7]. This model helps users access and interpret a curated set of literature by dynamically incorporating information into responses using a RAG framework. Rather than building a model from the ground up, the authors incorporated a pre-trained LLaMA2-7B model from Hugging Face into a RAG framework that handles 50 AM-related papers and textbooks in PDF form, later converted to TeX through Mathpix.

In [8], the authors explored a novel technique to overcome the challenges encountered in traditional RAG approaches in LLMs, such as outdated information, hallucinations, limited interpretability due to context constraints, and inaccurate retrieval. To overcome these challenges, Graph RAG(G-Rag) combines graph databases to improve the retrieval process. The suggested approach processes material science papers by extracting salient entities (termed as MatIDs) from sentences, and these are subsequently utilized to query external Wikipedia knowledge bases (KBs) for supplementary relevant information. An agent-based parsing method is used to attain a more fine-grained representation of the documents. A modified version of G-RAG, utilizes graph database to store relationships among these entities, improving both retrieval precision and contextual comprehension. This new approach shows impressive improvements in performance for domains where accurate information retrieval is needed, example: in material science.

The authors of [9] carried out a comparative analysis between a number of NLP approaches, including the frequency-based approach(spaCy), transformer model (Simple T5), and RAG with LLM-(GPT3.5-turbo). SciTLDR dataset is selected for this experimental research and the authors employed three different methods to carry out three different auto-generating systems for literature reviews. The ROUGE scores are employed in the evaluation of all the three systems. From the evaluation, the highest ROUGE-1 score is obtained by the LLM-GPT-3.5-turbo, at 0.364. The transformer model ranks second and spaCy ranks last. Lastly, a graphical user interface is developed for the optimal system based on the LLM.

III. Methodology

The proposed RAG methodology enables structured information extraction from unstructured documents, which are typically challenging to analyze. By integrating document retrieval (SectionIII-A) with language generation (SectionIII-B), RAG produces responses that are both contextually relevant and present in the source material. The overall architecture of our model is illustrated in Figure 2 and further explained in the following subsections.

A. Retrieval

The retrieval phase of RAG model filters and isolates the most relevant information from a large collection of unstructured textual data, such as scientific PDFs. These documents are often not organized in a format suitable for direct querying. They typically contain lengthy paragraphs, varied formatting, inconsistent use of headings or tables, and dense technical language, which makes it difficult to pinpoint specific, context-relevant pieces of information directly. To address this, a series of steps are taken to semantically structure the content.

1) Data Ingestion: The pipeline begins by allowing users to upload multiple PDF documents. Each document is processed

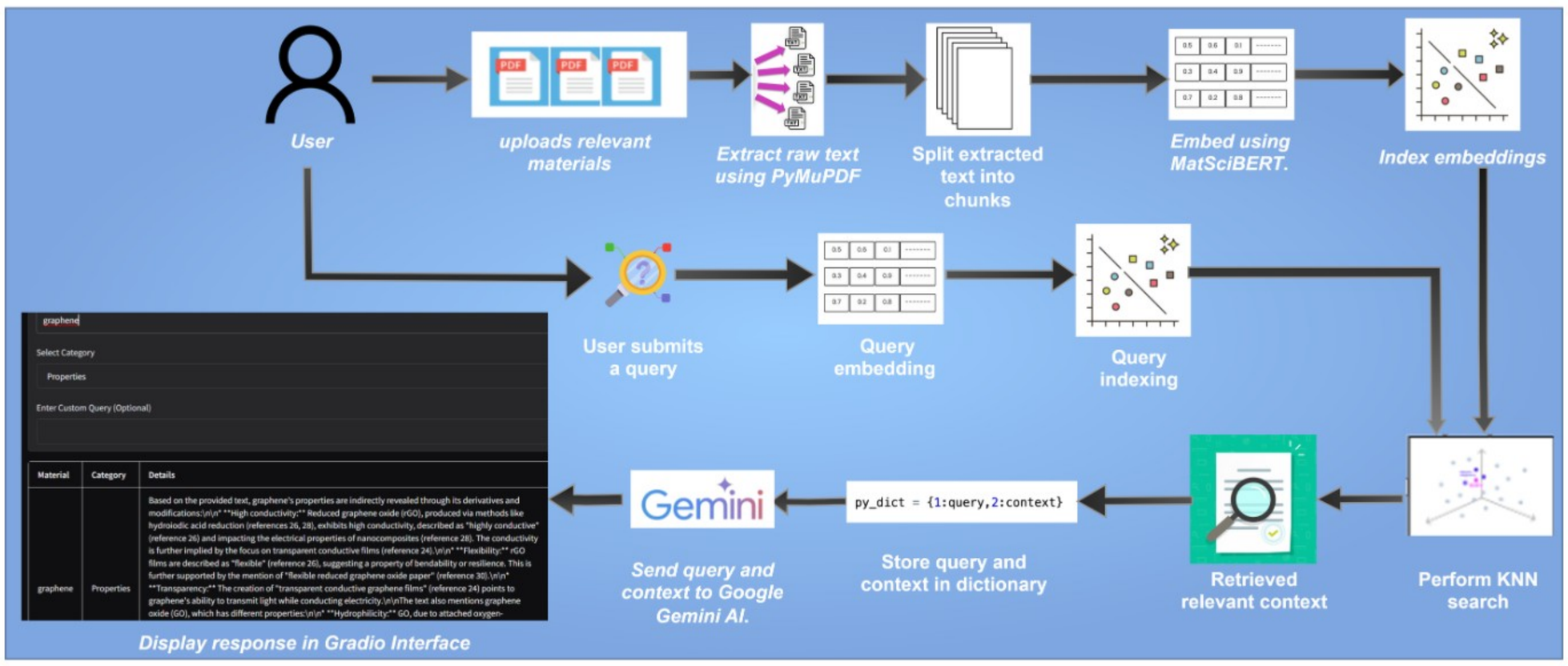

Fig. 2. RAG-Driven Framework for Extracting and Responding to Material Science Data Queries

page by page, and the textual content of each page is systematically extracted with *PyMuPDF* library. This ensures that the information from every section of the document is retained accurately. The extracted text serves as the foundational input for downstream processing, including chunking and embedding. This approach also enables accurate and scalable information retrieval from high-volume data sources.

2) Text Splitting: The proposed system splits the extracted text from each page into smaller and more manageable chunks based on sentence boundaries. The text is first broken into individual sentences, and these sentences are then grouped together incrementally until a pre-defined character limit (e.g., 500 characters) is reached. This method ensures each chunk stays within the size limits and keeps the sentence structure intact. It also makes the data easier to use for embedding and retrieval tasks. Splitting the text this way improves search efficiency and relevance because the system can match smaller, more focused segments instead of large, vague blocks.

3) Text Embedding: In this system, embeddings are essential for matching the user query to the unstructured texts extracted from PDFs. Both the query in natural language and the document chunks, are processed through the MatSciBERT model to convert them to a fixed-length vector. The embeddings encapsulates the semantics of scientific terms and the context surrounding it, enabling the system to correlate information based on the queries to the document-as-contents.

4) Indexing using FAISS: The system uses FAISS for fast and scalable similarity search that works by indexing high-dimensional vectors to support approximate nearest neighbor search. FAISS was designed to potentially handle billions of embeddings. It builds optimized indexes for similarity search, performing efficient distance calculations (such as Euclidean distance). This allows the system to efficiently retrieve the most relevant text chunks while ensuring strong performance, scalability, and support for millions of embeddings.

5) KNN Search: K-Nearest Neighbors (KNN) Search is a popular approach for selecting the most similar data points to a given input data point in a dataset, based on locations in the vector space. In KNN, K is simply the number of nearest neighbors we want to retrieve. The KNN algorithm calculates the distance between vectors - using different measures such as Euclidean distance or cosine similarity - in order to find the K vectors that are closest to the chosen input vector, allowing for the efficient retrieval of similar items in a variety of applications.

Once the user's query is converted into an embedding, KNN is applied to search for the closest matches by calculating the distance (e.g., Euclidean distance) between the query embedding and the precomputed document embeddings in the vector space. Instead of comparing the query embedding with every document embedding individually, the process relies on an index that organizes similar embeddings together. This index allows the KNN algorithm to quickly narrow the search space and focus on the nearest embeddings, improving efficiency.

B. Generation

The top-K most relevant document chunks retrieved from the FAISS index are concatenated with the user's query and supplied as augmented context to the LLM (Google Gemini). The model leverages both the contextual excerpts and the query to generate a coherent, context-aware response tailored to the user's intent.

The retrieved excerpts act as supporting evidence, ensuring factual validity and domain relevance, while the generative model provides fluency and natural language coherence. This synergy allows the system to produce responses that are not only semantically accurate but also aligned with the underlying scientific context.

1) Data Logging: All interactions are logged for transparency and future analysis. Extracted materials and responses are stored in CSV files, enabling tracking of materials and user

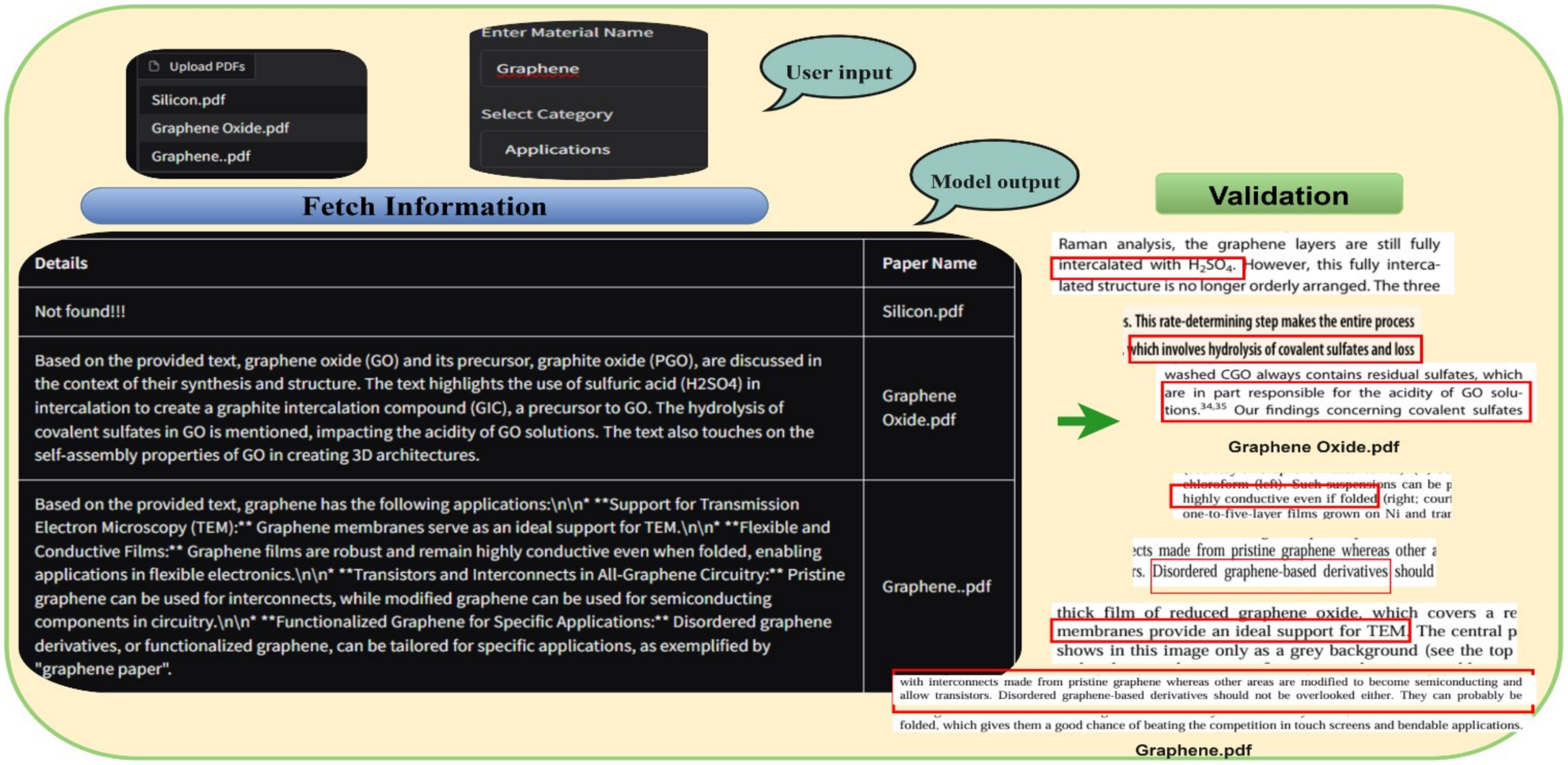

Fig. 3. A user interface enabling exploration of material applications, incorporating validation of the extracted content with respect to the original PDF.

queries. Results are displayed in a user-friendly format through the Gradio interface. This approach, inspired by RAG, ensures efficient, scalable, and informative outputs.

The proposed pipeline captures the complete workflow of the RAG framework incorporating PDF ingestion, embedding, retrieval, and the generation of final responses.

IV. Simulation

A. Dataset Creation

An evaluation dataset was systematically developed to benchmark retrieval and extraction performance in material science. The dataset was derived from 15 peer-reviewed research articles, yielding a total of 155 queries categorized into four classes: (i) Properties, (ii) Synthesis Methods, (iii) Applications, and (iv) Custom Queries. The *Custom Queries* class encompassed both *numerical tasks* (requiring extraction of scalar values such as structural dimensions, mechanical constants, or storage capacities) and *descriptive tasks* (requiring higher-order reasoning such as ranking, comparative analysis, or conceptual identification).

Numerical queries: Examples include:

- *"Inter-layer distance in graphene"*
- *"Young's modulus of the material"*
- *"Li-storage capacity of germanium-based materials"*

Descriptive queries: Examples include:

- *"Which is the best semiconductor used for photocatalytic hydrogen generation?"*
- *"Which material exhibited the highest hardness?"*
- *"What is the matrix material?"*

This dataset design explicitly stresses both factual precision and semantic generalization, thereby approximating the dual nature of practical information needs encountered in materials research.

B. Simulation Metrics

The model's performance was quantitatively assessed using the standard classification metrics. Each response of the query was annotated based on its correctness against the source paper. The annotation scheme follows standard parameters, defined below in accordance with the proposed methodology:

- True Positive (TP): Answer is present in the source and correctly retrieved/generated.
- True Negative (TN): No answer exists and the model correctly returns none.
- False Positive (FP): Model produces an answer that is present in the source but irrelevant or not supported by the query context.
- False Negative (FN): A valid answer exists, but was not captured by the model.

Based on the values of TP, TN, FP, FN values, the following simulation metrics are calculated.

- **Accuracy**: The proportion of total predictions (both positive and negative) that were correct.
- **Precision**: The proportion of retrieved answers that were relevant (i.e., correct and supported by the source).
- **Recall**: The proportion of all relevant answers in the source that the model was able to retrieve.
- **F1-Score**: The harmonic mean of precision and recall, representing the overall effectiveness of the model.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

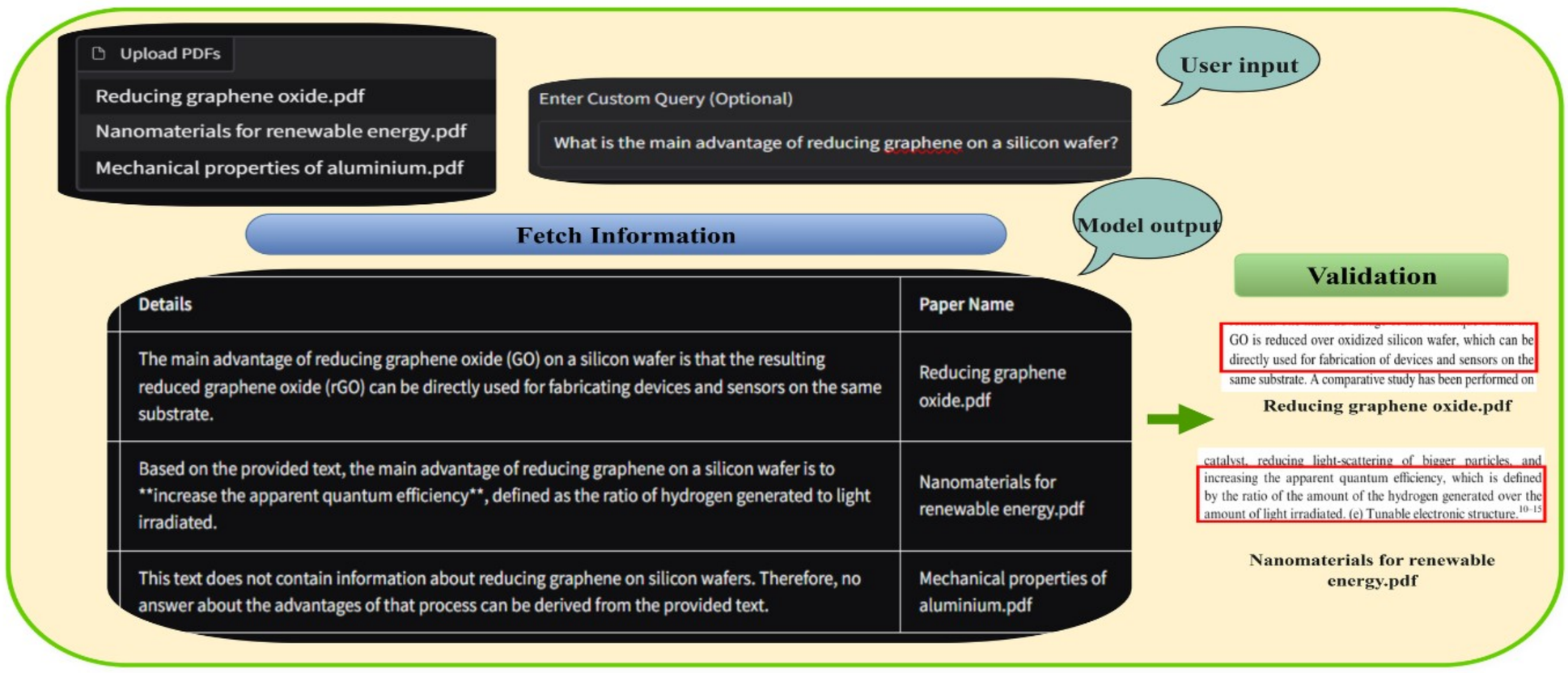

Fig. 4. The user-defined query was executed across three PDF sources, with relevant content retrieved exclusively from the first document.

C. Human Cross Evaluation

To assess the factual correctness and relevance of the model-generated answers in the material science domain, a structured human evaluation was conducted. This involved manually reviewing outputs from 15 research papers, covering 155 queries evenly distributed across four categories: properties, synthesis methods, applications, and custom questions. For each query, the model's response was manually cross-checked against the corresponding source PDF to verify whether the answer was present and correctly aligned with the user's intent. This ensured that the evaluation focused on both the accuracy and verified presence of each generated answer [10].

D. Examples

Figures 3 and 4 show how the system processes queries across multiple PDFs while checking results against the source documents. Figure 3 highlights how the model extracts applications of graphene and cross-checks each answer with the uploaded PDFs, keeping responses present in the source. Figure 4 illustrates a custom query about reduced graphene oxide. The system retrieves the explanation from *Reducing graphene oxide.pdf* and validates it against the text segments, preventing unsupported claims and keeping the process transparent.

In cases where the query lacked sufficient context or the information was absent from the PDF, the system transparently returned "Not found!!!", which reflects its ability to avoid hallucination and maintain response integrity.

E. Why is it better than models like ChatGPT

While models like ChatGPT are powerful LLMs, it has clear limitations for domain-specific tasks like material science information retrieval. A key drawback is that such models cannot process multiple PDFs at once. For example, uploading 15 research papers into ChatGPT simultaneously is not possible.

ChatGPT relies on its pre-trained knowledge, which may be incomplete, outdated, or irrelevant to the specific documents

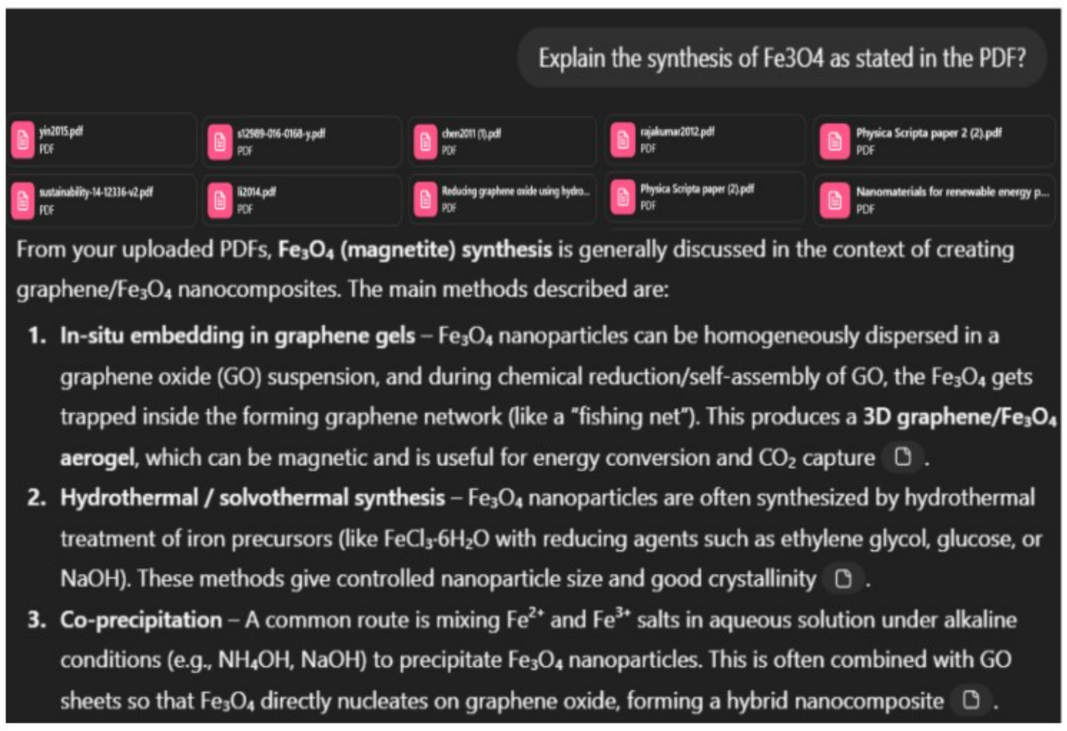

Fig. 5. Response to the query on synthesis in ChatGPT. The 1st method (in-situ embedding) is partially extracted and widely generated from PDF, while the 2nd (hydrothermal/Solvothermal synthesis) and 3rd (co-precipitation) methods are incorrectly pulled, showing hallucination.

at hand. When queries are provided without attaching PDFs, ChatGPT often generates generic or incorrect answers because it attempts to predict a likely response rather than retrieve facts from the actual sources. Even when PDFs are supplied, ChatGPT may fabricate references or attribute information to a document where it does not actually exist. This behavior reduces reliability, especially for scientific research that demands precision and verifiable presence in source material, as shown in Figure 5.

In contrast, the proposed framework is explicitly designed to retrieve answers only from the uploaded PDFs. This reduces the risk of hallucinations and maintains transparency by indicating when relevant information is not present in the source documents.

In short, models like ChatGPT excels in open-ended reasoning, it cannot guarantee source-verified, multi-document

retrieval. The RAG-based framework directly addresses this gap by ensuring factual accuracy, domain relevance, and scalability for handling large sets of research papers.

F. Results

As shown in Table I, MatSciBERT combined with RAG consistently outperforms the all-MiniLM baseline across all evaluation metrics. Its domain-specific embeddings capture materials science terminology more effectively, leading to higher precision (0.898) and recall (0.844). In contrast, the general-purpose all-MiniLM model struggles with technical vocabulary, resulting in weaker retrieval and lower overall accuracy.

All evaluation metrics, including precision, recall, and accuracy, were derived through manual assessment of retrieved responses. The human-verified results yielded an overall accuracy of 0.784, demonstrating substantial consistency with verified information present in the documents and further validating the model's reliability.

Model	Accuracy	Precision	Recall	F1-Score
MatSciBERT + RAG	**0.784**	**0.898**	**0.844**	**0.858**
all-MiniLM-L6-v2 + RAG	0.670	0.843	0.727	0.775

TABLE I
CLASSIFICATION METRICS FOR GENERATED ANSWERS.

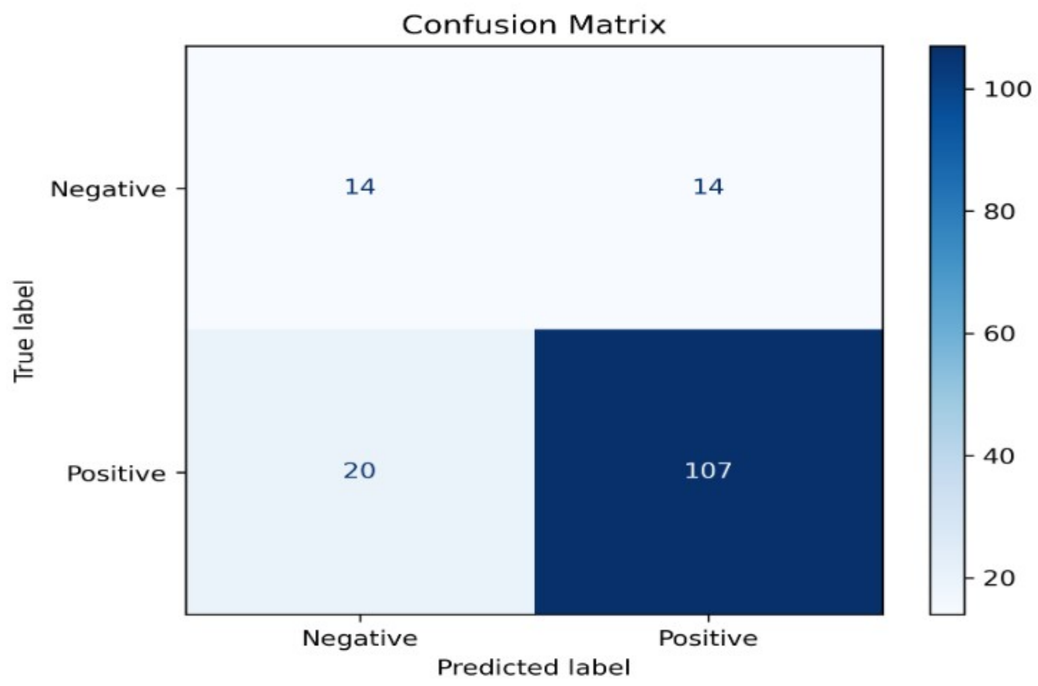

Fig. 6. Confusion matrix showing classification performance of the RAG framework using Matscibert model

The confusion matrix in Figure 6 shows that the model correctly classifies most positive instances (107) with relatively few false negatives (20). Negative instances are more balanced, with 14 true negatives and 14 false positives, indicating that errors are more likely when distinguishing negative cases. Overall, the model demonstrates strong performance on positive samples, which dominate the dataset.

Overall, the system offers fast, accurate, and scalable extraction and analysis of materials data, with potential for future enhancements in storage, UI, and analytics.

V. CONCLUSION

This work presents an efficient technique for extracting material science information from PDFs using RAG framework. By combining the abilities of KNN, MatSciBERT, and the Google Gemini Model, the system is able to promptly and accurately retrieve relevant document segments. This approach makes complex material science information more accessible, enabling users to efficiently extract specific material properties, applications, and synthesis processes from large datasets. While the system demonstrates high precision and scalability for processing technical texts, its performance can be affected by increased processing time when handling extremely large datasets.

Upcoming research will concentrate on enhancing computational performance by improving data source integration, refining the model, and applying more advanced embeddings to speed up retrieval and increase accuracy. Additional efforts will also be directed toward mapping the intricate relationships between material properties and their applications, particularly examining how specific features influence suitability for various purposes.

REFERENCES

[1] Ghanshyam Pilania. Machine learning in materials science: From explainable predictions to autonomous design. *Computational Materials Science*, 193:110360, 2021.

[2] Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. Data-driven materials science: status, challenges, and perspectives. *Advanced Science*, 6(21):1900808, 2019.

[3] Qintong Zhang, Bin Wang, Victor Shea-Jay Huang, Junyuan Zhang, Zhengren Wang, Hao Liang, Conghui He, and Wentao Zhang. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*, 2024.

[4] C Nidhisree, Ananya Paul, Anaswara Venunadh, and Rajat Subhra Bhowmick. Generative ai under scrutiny: Assessing the risks and challenges in diverse domains. In *2024 IEEE 6th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, pages 243–248. IEEE, 2024.

[5] Binglan Han, Teo Susnjak, and Anuradha Mathrani. Automating systematic literature reviews with retrieval-augmented generation: A comprehensive overview. *Applied Sciences*, 14(19):9103, 2024.

[6] Balduin Katzer, Steffen Klinder, and Katrin Schulz. Towards an automated workflow in materials science for combining multi-modal simulation and experimental information using data mining and large language models. *Materials Today Communications*, 45:112186, 2025.

[7] Achuth Chandrasekhar, Jonathan Chan, Francis Ogoke, Olabode Ajenifujah, and Amir Barati Farimani. Amgpt: a large language model for contextual querying in additive manufacturing. *Additive Manufacturing Letters*, 11:100232, 2024.

[8] Radeen Mostafa, Mirza Nihal Baig, Mashaekh Tausif Ehsan, and Jakir Hasan. G-rag: Knowledge expansion in material science. *arXiv preprint arXiv:2411.14592*, 2024.

[9] Nurshat Fateh Ali, Md Mahdi Mohtasim, Shakil Mosharrof, and T Gopi Krishna. Automated literature review using nlp techniques and llm-based retrieval-augmented generation. In *2024 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 1–6. IEEE, 2024.

[10] Juan José González Torres, Mihai Bogdan Bîndilă, Sebastiaan Hofstee, Daniel Szondy, Quang Hung Nguyen, Shenghui Wang, and Gwenn Englebienne. Automated question-answer generation for evaluating rag-based chatbots. In *1st Workshop on Patient-Oriented Language Processing, CL4Health 2024*, pages 204–214. European Language Resources Association (ELRA), 2024.