



LISA: A Lithium-Ion Solid-State Assistant using large language models for knowledge defragmentation in battery science and beyond

Yinghan Zhao^{a,1}, Anna-Lena Hansen^{b,c,1}, Anna Dahlhaus^c, Nico Brandt^a, Michael Selzer^{a,d}, Arnd Koeppel^{a,*}, Britta Nestler^{a,d}, Michael Knapp^b, Helmut Ehrenberg^b

^a Institute for Applied Materials-Microstructure Modelling and Simulation, Karlsruhe Institute of Technology, Strasse am Forum 7, Karlsruhe D-76131, Germany

^b Institute for Applied Materials-Energy Storage Systems, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Karlsruhe D-76344, Germany

^c opencampus.sh // Campus Business Box e.V., Fraunhoferstrasse 13, Kiel D-24118, Germany

^d Institute of Digital Materials Science, Karlsruhe University of Applied Sciences, Moltkestrasse 30, Karlsruhe D-76133, Germany

ARTICLE INFO

Keywords:

Large language model
Retrieval augmented generation
Solid-state battery
Research assistant

ABSTRACT

This work presents the development and implementation of a research assistant tool, Lithium-Ion Solid-State Assistant (LISA), based on the Retrieval-Augmented Generation (RAG) architecture. This assistant has been specifically tailored to enhance the retrieval and extraction of information from the domain of solid-state battery research. The system employs sophisticated retrieval techniques to efficiently identify the most pertinent document segments in response to researcher queries. The segments above are subsequently collated into prompts for a Large Language Model (LLM), which generates accurate, contextually enhanced responses to queries about solid-state battery-related subjects. This approach has the potential to markedly improve the accessibility and usability of a range of documentation, from project reports to complex scientific literature. The system provides researchers with a powerful tool to bridge disciplinary gaps, facilitate cross-disciplinary communication, accelerate knowledge discovery, and drive innovation in the field of solid-state batteries. A comprehensive evaluation was conducted to assess the system's performance, with results indicating its potential to transform scientific research workflows. The system offers a robust open-source framework for future advancements in automated knowledge retrieval, understanding, and management, particularly in supporting the development of new materials.

1. Introduction

The field of solid-state battery research is an emerging, dynamic, and interdisciplinary area of study, as evidenced by the recently rising number of publications (Fig. 1) and the wide range of scientific fields covered (Fig. 2), including chemistry, materials science, physics, electrochemistry and more. With such a diverse set of disciplines involved, researchers are addressing many facets of solid-state battery development, from fundamental material properties to practical applications [1]. This interdisciplinary approach facilitates a comprehensive understanding of the challenges and opportunities of advancing solid-state battery technology. Nevertheless, due to this multidisciplinary approach and highly specialized research, knowledge fragmentation is inevitable and poses a significant challenge to efficiently addressing this issue.

The phenomenon of knowledge fragmentation is defined as the splitting of a subject into smaller components due to increasing specialization. As the pieces become smaller, the challenge of reassembling the puzzle intensifies, as described by Karlqvist [3]: “*There is a price to be paid for fragmentation and specialization [...] The parts can no longer be put together easily [...] As science moves closer to applications, decision- and policy making, problems occur that cannot be confined to narrow disciplines or kept within the borders of specific departments*”.

The challenge of knowledge fragmentation extends beyond merely putting the pieces back together. The emergence of intellectual silos and rivalry, in addition to the difficulty of appreciating common ground due to insufficient communication, divergent terminology, or differing priorities, also constitute significant obstacles. The broader the scope of a research project, the more likely it is that the scientists involved are highly specialized with diverse backgrounds, thereby complicating

* Corresponding author.

E-mail address: arnd.koeppel@kit.edu (A. Koeppel).

¹ These authors contributed equally.

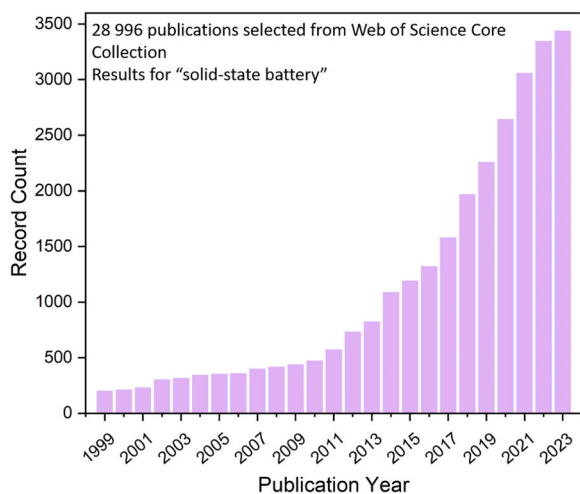


Fig. 1. Search results from Web of science (visited at 19.01.2024) [2] for the number of publications related to the topic “solid-state battery”.

establishing a shared language.

It is not uncommon for researchers to fail to identify shared ontologies or to employ vague notations specific to their respective fields. For example, chemical compounds can be described in many ways, from stoichiometry, IUPAC nomenclature [4], trivial names, mineral names, and structure types to cryptic abbreviations or industry codes. As a consequence, misunderstandings can arise, obstructing fruitful exchanges of ideas and potentially hindering rather than facilitating collaboration. The breakdown of intellectual silos necessitates deliberate efforts to promote openness, inclusivity, and connectivity across disciplines. The importance of shared knowledge and active collaboration cannot be overstated. Regardless of the scale of the project, ranging from modest intra-institutional endeavors to extensive international ventures spanning millions of euros, shared knowledge remains the cornerstone.

An additional obstacle is posed by the dilution of knowledge due to an increasing volume of publications, making it time-consuming and

challenging to find relevant information for specific scientific questions. All the aforementioned phenomena - fragmentation of knowledge, intellectual silos, and dilution of knowledge - are adversaries for efficient research projects.

In this work, a solution has been proposed for defragmenting knowledge and ensuring more effective progress in research projects, large or small, through the centralization of knowledge (e.g., publications) and the use of Large Language Models (LLMs) to handle them [5, 6]. LLMs are advanced computational systems designed to process and generate human-like text by leveraging vast datasets and sophisticated algorithms based on neural network architectures. These models demonstrate the capacity to comprehend textual data, generate coherent responses, and perform a range of natural language processing tasks, offering promising prospects by enabling seamless integration and transfer of knowledge among diverse disciplines. Notable advantages of employing LLMs in this regard include [6]:

- Cross-disciplinary contextualization: LLMs are proficient at deciphering and extracting semantic content from extensive textual data, thereby enabling the identification of interconnections among disparate academic literature.
- Unified terminology: LLMs can facilitate the acquisition and application of consistent terminology across disciplines, thereby reducing the communication barriers that may arise from inconsistent or ambiguous terminology.
- Accessible expertise: LLMs democratize access to specialized knowledge, offering expert guidance irrespective of geographical location or institutional affiliation. Individuals previously confined to their respective silos can now expand their horizons, broadening their respective perspectives.
- Relevant information: LLMs can quickly find relevant publications within a vast repository based on the researcher’s query, enabling targeted exploration of relevant content without overwhelming the user with the sheer quantity of available information.

To illustrate the potential of LLMs as a solution to the problem of knowledge fragmentation, the FestBatt initiative in Germany is presented as a case study, a nationwide cluster of competence focusing on

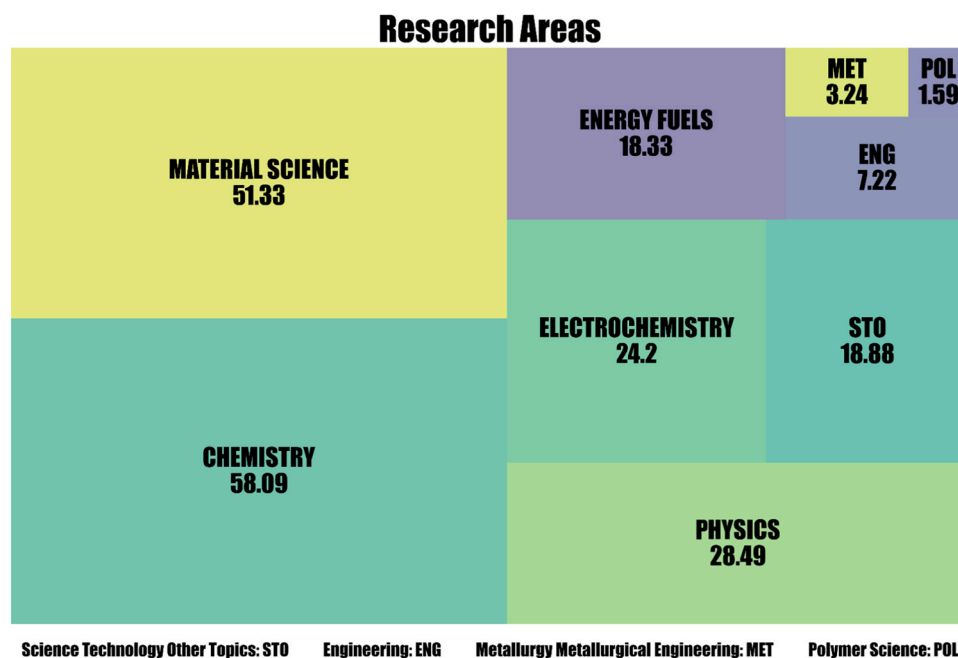


Fig. 2. Research areas working on “solid-state batteries” based on results from Web of Science (visited at 19.01.2024) [2]. Numbers are represented in percentage, with respect to 28.996 publications in total. Only areas are displayed that have a share > 1 %.

solid-state battery development [7]. The project brings together over 22 institutes and a team of over 180 scientists. The diversity of academic backgrounds among these specialists encompasses chemistry/electrochemistry, materials science, computer science, theoretical physics, and engineering. While these specialists collectively strive toward the common objective of advancing solid-state battery technology, the development and maintenance of a common language are ongoing dynamic efforts. Each scientific community possesses unique (explicit or implicit) ontologies, frequently undisclosed methodologies, and occasionally prejudices against alternative approaches to tackling challenges.

The fragmentation of knowledge in battery research is more than just a consequence of the involvement of numerous participants. It is also a consequence of the increasing sophistication of experimental techniques. It is now possible to analyze a battery at every length scale, from the atomic to the macroscopic. Ideally, these experiments are conducted in a standardized multi-technique experiment at large-scale facilities. However, this is only sometimes possible in practice [8]. Hence, smaller fragments of the whole picture are published. Additionally, review articles condense and summarise recent articles to present information that allows the reader to understand the current state of knowledge in that field. These articles are typically written by experts in the field and are of high value, although their references may be limited, and their scope may not answer very specific questions.

In summary, knowledge fragmentation and dilution are commonly associated with:

- Specialized knowledge staying in isolated groups
- Too much information to process
- Different systems that do not work well together
- Varying methods or approaches
- Organizational obstacles

and can lead, especially in interdisciplinary teams, to research inefficiencies, duplicated efforts, communication barriers between specialists, missed opportunities for innovation at disciplinary intersections, delayed scientific breakthroughs, and ultimately slower progress in critical fields like solid-state battery development where interdisciplinary collaboration is essential for advancement.

To address knowledge fragmentation, we propose the Lithium Ion Solid-state Assistant (LISA), an LLM-based virtual research assistant specifically designed to defragment knowledge in battery science, using publications from Germany's FestBatt initiative as a case study. LISA works by centralizing diverse publications—from scientific papers to dissertations and technical reports - in a shared virtual research environment (Kadi4Mat [9]) and employs Retrieval-Augmented Generation (RAG [10]) technology to intelligently access, connect, and contextualize information across disciplinary boundaries. By providing cross-disciplinary contextualization, unified terminology, accessible expertise, and targeted retrieval of relevant information, LISA helps researchers overcome the barriers created by knowledge fragmentation and accelerates collaborative scientific progress in this critical field. Thus, the deployment of virtual assistants powered by LLMs may facilitate the efficient coordination of multidisciplinary teams, streamline workflows, and harmonize objectives.

The following sections of this work are organized as follows: first, a comprehensive review of related literature is presented, which provides a foundation for the current study. Secondly, the underlying principles of the system are presented, along with a detailed explanation of the corresponding User Interface (UI). Thirdly, an in-depth evaluation of the system's performance is carried out, followed by a discussion of the results and their implications.

2. LLM applications in materials research

Materials science, in general, is an interdisciplinary field that studies

the properties, performance, and applications of materials to develop new and specifically optimized materials. It combines principles from physics, chemistry, and engineering to understand how a material's structure at the atomic or molecular level affects its macroscopic properties. In recent years, Machine Learning (ML) has emerged as a powerful tool in materials science, revolutionizing how researchers approach traditional challenges. Generally, ML algorithms can analyze vast amounts of data generated from experiments, characterization, and simulations, revealing patterns and relationships that may be missed by traditional methods [11]. The advent of large language models has provided researchers with access to sophisticated ML methods that can be further applied in various aspects of the development of materials [12,13]. Scientists are inspired by the vast majority of possible applications and try different implementations in their respective fields [14]. For example, Jablonka et al. [15] explored exemplary prototypes in the field of materials science and chemistry developed during a hackathon, from predicting the properties of materials at different scales to molecule discovery and extracting knowledge from unstructured documents, demonstrating the versatility of the LLM in materials science and its significant potential to build bridges between the diverse scientific disciplines.

Among possible applications, one of the most discussed topics is using ML to predict the various types of properties of materials [16]. This capability is particularly valuable in the early stages of materials development, where experimental testing can be time-consuming and costly. Especially, using LLMs to predict properties is usually achieved by utilizing the so-called In-Context Learning (ICL) ability of the ML model, where a few example input/output pairs (e.g., materials compositions and interested properties) are provided to the model in the prompt for helping the model to figure out how the task should be completed. Successful examples include the prediction of physical and electronic properties of crystalline solids [17], electrical performance in solar cells [18], and various molecule properties [19].

LLMs can also play a critical role in helping researchers discover and design materials [20]. The traditional trial-and-error approach is often inefficient for exploring the vast space of possible material combinations. In contrast, powerful generative models, such as LLMs, show great promise in efficiently extracting and capturing complex knowledge from training data and suggesting potential correlations, thus exploring this space and suggesting new materials with desired properties [21,22]. The LLMs can be further empowered by various tools [23] and integrated into the research workflow [24], whereby they can act as intelligent agents, thereby providing significant value in autonomous experimental design and execution [25].

Another area of growing interest is using LLM to revolutionize data extraction in materials science, which is achieved by automating the conversion of unstructured text from research papers, patents, and reports into structured data formats [26]. These models can parse complex narratives to identify and extract key information such as material properties, synthesis methods, and experimental conditions [27–31]. More recently, LLMs have also been extended to handle visual inputs (namely multimodal LLMs) and have shown considerable potential to transform materials science by integrating and analyzing textual and visual data [6]. This approach can potentially address the challenges mentioned above more effectively, and they can process written content from research papers and patents alongside visual data such as graphs, spectra, and microscopy images [32]. This enables them to enhance their ability to assist in analyzing the complicated data generated from modern, advanced characterization techniques. By correlating information across these modalities, multimodal LLMs can accurately extract and synthesize comprehensive datasets, uncovering intricate relationships between material properties. This automation significantly accelerates data analysis, enabling researchers to quickly compile comprehensive datasets and find trends, facilitating meta-analyses and informed decision-making.

This work examines the use of LLMs to address knowledge

fragmentation, and an RAG research assistant system has been developed to address the previously outlined issues. The results suggest that this system possesses the potential to transform scientific research workflows. It provides a robust framework for future advancements in automated knowledge retrieval while accelerating the process of developing materials.

3. Retrieval augmented generation

LLMs have demonstrated considerable achievements in a variety of domains. Nevertheless, they still face numerous constraints, particularly in domain-specific or knowledge-intensive tasks. One significant challenge is the generation of “hallucinations” when these models are tasked with queries beyond their training data, particularly in cases where updated or real-time information is required. LLMs have a relatively poor ability to incorporate long-tail knowledge, which may also contribute to generating “hallucinations”, leading to incorrect results [33]. To address these challenges, RAG is a promising technique that can augment LLMs by allowing the referencing of external knowledge information, such as personal document collection. This is typically achieved by retrieving pertinent document chunks from an external knowledge source through semantic similarity calculations. This approach can effectively address the issue of generating factually incorrect content and provides a cost-effective solution to the problem of insufficiently up-to-date knowledge of the model itself. The integration of RAG into LLMs has been widely adopted, establishing it as a pivotal technology for advancing chatbots and enhancing the practicality of LLMs in real-world applications. In contrast to specifically trained LLMs, the knowledge base of RAG architectures can readily be changed, making it easy to apply to different fields of study.

The workflow of a typical RAG system consists of several key steps, as shown in Fig. 3. The process begins with indexing, where raw documents in various formats are standardized into plain text, and these texts are then broken down into chunks, encoded into vectors, and stored in a vector database. In the retrieval step, user queries are transformed into vector representations and similarity scores are computed between these query vectors and the stored chunk vectors to retrieve the top k most relevant chunks. Finally, in the generation step, the user query and the retrieved chunks are combined to create a prompt for the large language model. The model can generate a response, incorporating the conversation history for multi-turn interactions. The main extension of LLM by RAGs, namely information retrieval, is a popular area of research. This can be traced back to the 1970s, which marked a pivotal development within the field of computer science [34]. In contrast to generation, retrieval involves identifying pertinent existing objects within a vast repository of resources. The most prevalent retrieval application can be observed in web search engines, which primarily focus on document retrieval. In the contemporary era, effective information retrieval systems can handle document collections containing billions of items.

Moreover, built upon the typical foundational RAG process,

numerous enhancements have also been proposed to improve the overall quality. These enhancements encompass methods for specific components, optimizations targeting the entire pipeline, and subsequently, RAG technology has experienced rapid development in recent years. The evolution of RAG in the context of LLMs can exhibit several distinct stages with the advances in the machine learning field. Its popularity was boosted further after the arrival of Generative Pretrained Transformer (GPT) models like ChatGPT by OpenAI, which marked a pivotal moment [35], with LLMs demonstrating powerful ICL capabilities [5]. Currently, advanced RAG usually introduces targeted improvements to address the limitations of basic RAG. It focuses on enhancing retrieval quality through both pre-retrieval and post-retrieval strategies. Pre-retrieval strategies optimize the indexing structure and the original query, while post-retrieval strategies select essential information, emphasize critical sections, and shorten the context to be processed. Recently, RAG research shifted towards providing better information for LLMs to answer more complex and knowledge-intensive tasks during the inference stage, leading to rapid development in RAG studies. As research progresses, enhancing RAG can offer enhanced adaptability and versatility. This includes various strategies for improving its components, including adding a search module for similarity searches, the refinement of the retriever through fine-tuning [36], and considering hierarchy-structured information such as knowledge graphs [37]. Although RAG architectures are more commonly used in the text domain, they have rapidly expanded to multimodal domains such as image, audio, and video data [6]. RAG architecture has also been applied to assist researchers in developing novel materials in different areas, such as building a chatbot system for Metal–Organic Framework (MOF) material [38] and enhancing the construction of knowledge graphs [39].

4. Architecture of LISA

In this work, the LISA research assistant employs the RAG architecture, which represents a cutting-edge approach that significantly enhances the performance of language models by integrating external knowledge into the generation process. An illustration of this system is shown in Fig. 4. The system’s workflow commences with document parsing and metadata loading, a critical step where raw documents are analyzed to extract relevant information and metadata such as author information or titles. This metadata enriches the document’s context, improves the search’s precision, and is the source for the LLM to track the citation of the answer to be generated. Next, the documents are segmented into smaller, more manageable chunks. This segmentation ensures that the content is sufficiently granular to be processed efficiently while retaining contextual integrity. The chunk size is an important hyper-parameter that must be carefully configured, as it strikes a balance between shorter chunks that offer more precise retrieval and larger chunks that provide richer contextual information. Each chunk is then converted into vector embeddings using advanced

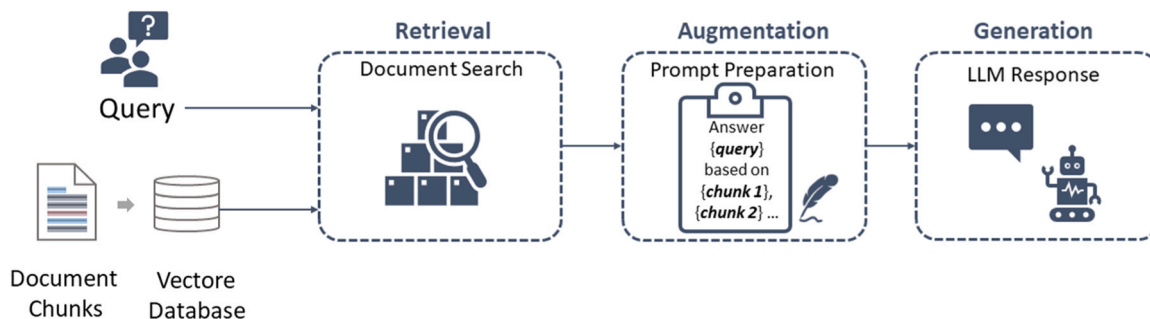


Fig. 3. The workflow of a basic RAG system typically involves several key steps: retrieval, augmentation, and generation. First, user queries search a vector database of document chunks. Relevant chunks are then combined with the query to form a prompt. Finally, a language model generates a response to this prompt.

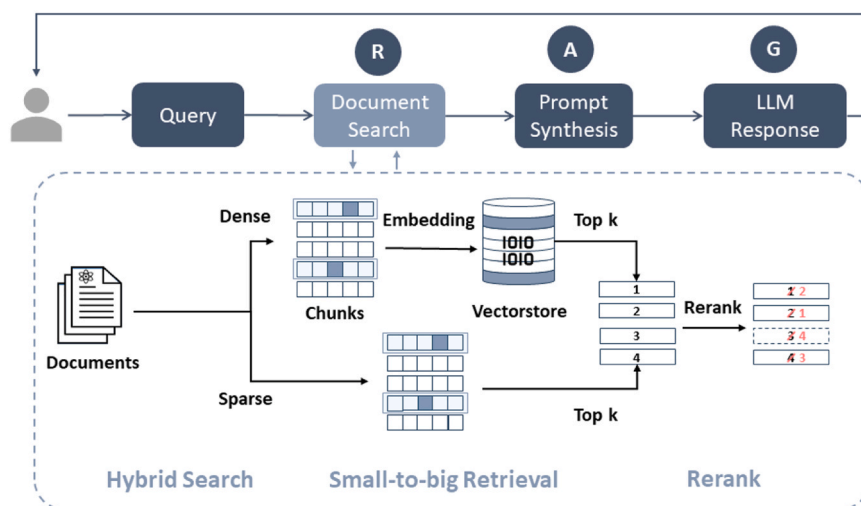


Fig. 4. The LISA architecture involves segmenting texts into chunks and embedding these fragments into vectors. During the search phase, a hybrid method is adopted (dense search with embedded vectors and sparse search with keywords), and the pre-filtered chunks are reranked for better order. For each chunk, the text in its accompanying context window is also returned for richer information (the so-called small-to-big retrieval [40]). Together with the query, a prompt is crafted as the input for LLM, and this prompt instructs the model to generate responses based on the user’s query within the context gathered during the search phase.

natural language processing techniques, where these embeddings encapsulate the semantic meaning of the text. The choice of the embedding model plays a crucial role, as a well-selected model can significantly improve system performance. By enhancing the quality of the embeddings, the model boosts the system’s ability to accurately retrieve the most relevant information from the entire database, leading to faster and more precise similarity searches.

Upon receipt of a user query, the RAG system employs a hybrid search mechanism [41]. This approach combines dense retrieval, which

leverages the vector embeddings to identify semantically similar chunks, with sparse retrieval, which utilizes traditional keyword-based methods to identify relevant documents. This dual strategy ensures that both conceptual relevance and specific keyword matches are considered, leading to more accurate and comprehensive retrieval results. Once the relevant chunks have been identified, they are reranked to prioritize the most pertinent information [42,43]. The reranking process considers several factors, including the relevance score assigned during the retrieval phase, the quality of the chunks, and their contextual

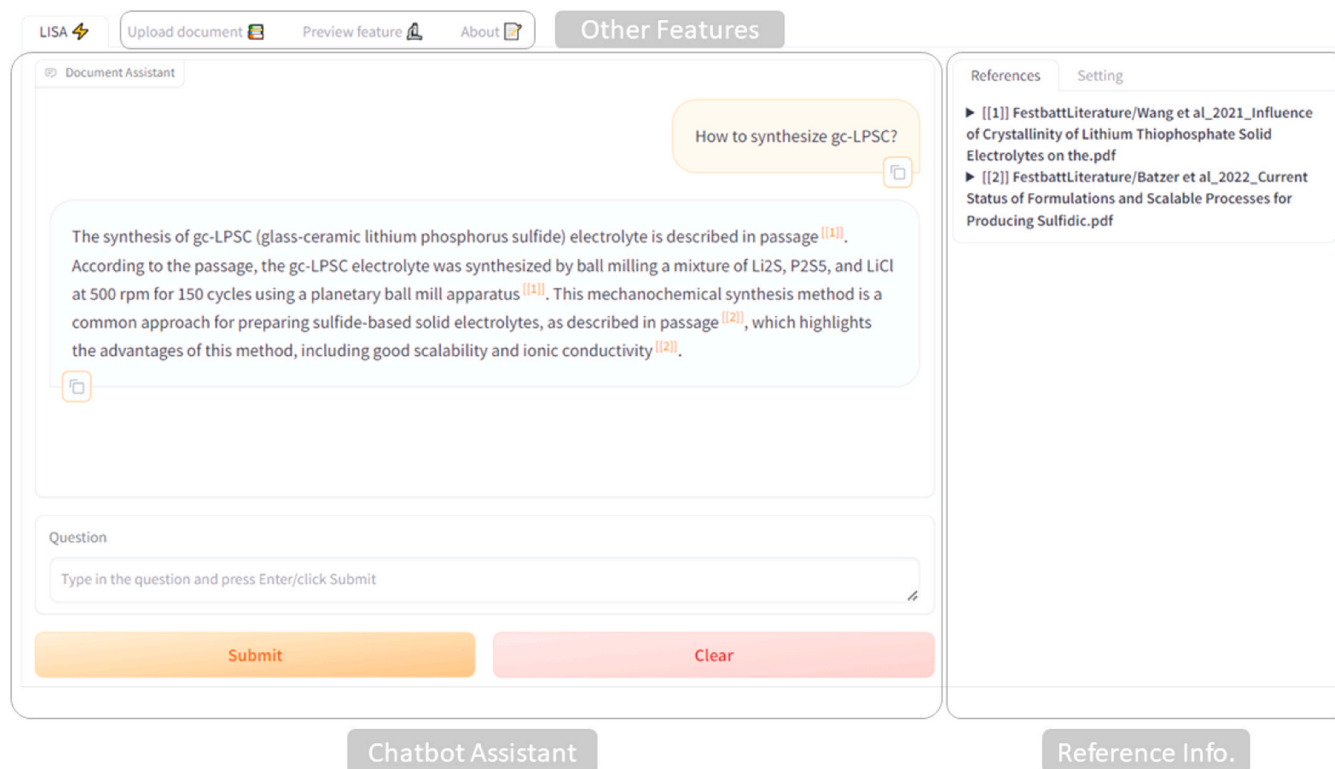


Fig. 5. UI of LISA. The interaction in this figure demonstrates the assistant’s ability to parse complex scientific questions and provide detailed, citation-backed answers. The user’s question about synthesizing a specific electrolyte in the solid-state battery is met with a precise procedural explanation, referencing specific scientific processes and studies, thereby illustrating the assistant’s utility in academic and research contexts.

alignment with the query [44]. This mechanism assists in reassessing and refining the retrieval results, thereby leading to additional performance gains.

Finally, the top-ranked chunks are aggregated into a cohesive prompt. This prompt, which now contains the most relevant and contextually appropriate information, is fed into the LLM. The LLM uses this enriched prompt to generate a response that is accurate, contextually rich, and informative. By integrating document retrieval and content generation, the RAG system effectively leverages large external knowledge bases and the advanced generative capabilities of modern LLMs, resulting in highly informed and accurate responses. This sophisticated interplay between retrieval and generation makes RAG a potent tool for various applications, such as this chatbot assistant, which can help researchers with scientific publications and knowledge. A detailed evaluation of the performance of this system is provided in the evaluation section.

Fig. 5 illustrates the chatbot assistant's User Interface (UI), designed to assist users by providing information based on uploaded documents and related references. The user interface is divided into three principal sections: the primary document assistance area, the references panel, and other feature panels. The central area of the user interface is the document assistant area, which features an interactive chatbot labeled "Document Assistant". In the displayed interaction, the user posed a query regarding the synthesis process for a specific electrolyte in solid-state batteries. In response, the assistant provides a comprehensive answer with detailed references to specific passages within the uploaded documents. To the right of the aforementioned "Document Assistant" area is a panel bearing the label "References". This panel lists the documents recorded in the database and used by the assistant to generate responses. These are of great importance, as they provide the source material for the information presented by the assistant, thus ensuring that the responses are grounded in credible and specific scientific literature. In addition to the primary functionality, the system offers a range of supplementary tools and options for user interaction with the assistant. To illustrate, the "Upload Document" function enables users to upload new documents for the assistant to reference. The "About" panel provides information about the development and functionalities of the system. The "Preview Feature" panel displays the latest experimental content or functionality for testing purposes and feedback collection, which facilitates the robust development of the system.

5. Results and evaluation

LISA generates responses based on selected and curated scientific documents while providing relevant references. Below are example prompts along with the corresponding answer, illustrating the qualitative capabilities of the system. Their corresponding references and further examples can be directly found in the online demo instance of LISA:

- General Knowledge Inquiry:** *What is a solid-state battery?*
- LISA's Response:** *"A solid-state battery (SSB) is a type of battery that uses a solid electrolyte instead of a conventional liquid electrolyte. This design enables SSBs to offer potentially higher energy storage capacities and improved performance, leading to longer driving ranges and shorter charging times, particularly in electromobility applications. The solid electrolyte facilitates ion transfer between electrodes and can be composed of various materials, including thiophosphate-based, oxide, sulfide, and polymer-based compounds."*
- Domain-Specific Inquiry:** *How to synthesize gc-LPSC (a type of solid-state electrolyte material)?*
- LISA's Response:** *"To synthesize gc-LPSC, a mixture of Li_2S , P_2S_5 , and LiCl is ball milled at 500 rpm for 150 cycles using a planetary ball mill apparatus. This mechanochemical synthesis method is widely used for sulfide-based solid electrolytes, as it offers a good balance between scalability and ionic conductivity. Compared to solid-state reaction methods,*

mechanochemical synthesis is more easily transferable to industrial-scale production. The specific synthesis parameters include the use of Li_2S (Alfa, 99.9 %), P_2S_5 (Sigma-Aldrich, 99 %), and LiCl (Alfa Aesar, 99 %) as starting materials, with ball milling conducted at 500 rpm for 150 cycles.

- Project Affairs Inquiry:** *Which department should I contact for support at Festbatt? LISA's Response:* *"For administrative support at FestBatt, you should contact ..., who is responsible for the technical-administrative coordination of FestBatt, at ... Additionally, you can reach out to the FB2-Koord group at Justus-Liebig Universitaet Giessen, Zentrum für Materialforschung (ZfM), where ... serve as contact persons. Further details on administrative support and contact information can also be found on the FestBatt website: [<https://festbatt.net>]."*

The quantitative evaluation of RAG models, such as LISA, is also a significant research focus. Its primary aim is to gain insight into and optimize RAG performance across various applications by examining key downstream tasks, datasets, and evaluation methods. Traditionally, RAG model assessments have targeted specific tasks, utilizing established metrics such as Exact Match (EM) and F_1 score for question answering [45], Accuracy for fact-checking, and BLEU and ROUGE for evaluating answer quality [46,47]. In the current era, the primary evaluation objectives for the RAG system include: (i) retrieval quality, which is assessed using metrics that are commonly used in search and recommendation systems [48]; and (ii) generation quality, which is evaluated based on faithfulness, relevance, non-harmfulness for unlabeled content, and accuracy for labeled content, using both manual and automatic methods. Several recent popular benchmark tests and tools, including ARES [49], RAGAS [50], and TruLens [51], offer quantitative metrics that assess RAG performance and deepen our understanding of its capabilities.

To assess the system's performance, RAGAS was adopted in this work as the evaluation framework, and our system database contains approximately 160 open-access publications and technical documents. The scoring metrics for this framework include five components, which are designed to evaluate the performance of the system from different perspectives:

- **Context precision:** a metric that gauges the accuracy of the context provided by the system. It indicates the degree to which the retrieved context aligns with the information required to answer a query.
- **Context recall:** assesses the system's capacity to retrieve comprehensive information, indicating the extent to which the relevant context is captured.
- **Faithfulness:** a metric that assesses the degree to which the generated answers adhere to the retrieved context, ensuring that the responses are grounded in accurate information.
- **Answer relevance:** assesses the degree to which the responses are pertinent and related to the queries posed, reflecting the system's ability to remain on-topic.
- **Answer correctness:** assesses the factual accuracy of the responses, ensuring that the answers provided are relevant and correct.

Together, these metrics provide a comprehensive framework for assessing the effectiveness and reliability of RAG systems in generating accurate, relevant, and contextually appropriate responses. For each document, approximately 10–20 question and answer pairs are generated with an LLM and manually checked by either the authors of the respective paper or the authors of this work for quality assurance, which serves as the test set to evaluate the system's performance.

A systematic study of common architectures and settings in RAG systems is conducted to ascertain how these elements influence system performance. Beginning with baseline configuration (chunk size with 512 characters, all-mpnet-base-v2 as embedding model [52], Mixtral 8x7B as the main LLM [53]), a step-by-step investigation is meticulously executed to analyze the impact of various factors such as chunk

configurations (chunk size and the number of returned chunks), embedding models, context windows, hybrid search functions, and response reranking on the system's performance. The issue of parameter combinatorial explosion is avoided by isolating each factor. The time the system takes to respond to a query is recorded (averaged over 500 queries on a platform with a single GPU A100–80GB). These results (averaged score on a scale of 0.0–1.0, with 1.0 being best) are listed in Table 1, highlighting the impact on both the effectiveness, measured by score (where higher values indicate better performance), and efficiency, measured by time (where lower values are preferable). More detailed results are provided in Supplementary Information S1.

Table 1 comprehensively evaluates different settings and configurations of the RAG architectures, demonstrating the impact of various components on the system's performance. The evaluation begins with a baseline configuration with a score of 0.761 (and a processing time of 9.208 s). It incorporates optimized chunk configs, embedding models, expanded context windows (small-to-big retrieval), reranking mechanisms, and hybrid search strategies [34]. These modifications contribute to successive performance improvements, as evidenced by the scores. To facilitate comparison, the relationship between scores and settings is also depicted as a bar graph and presented in Fig. 4(a). (Fig. 6)

The baseline setting, devoid of any additional optimizations, serves as the control setting, providing a reference point against which subsequent settings are evaluated. The introduction of advanced settings is gradual, with modifications made in a stepwise manner, again, to avoid the problem of combinatorial explosion. The first of these is the adjustment of the *chunk configurations* (opt. 1), which boosts the score to 0.818, leading to a considerable improvement in performance. This indicates that the model benefits from larger, more contextually rich text segments during retrieval. Incorporating *embeddings optimization* (opt. 2) resulted in a notable enhancement of the model's ability to understand and represent text semantically, as evidenced by the increase in score to 0.864. Incorporating a small-to-big retrieval method (opt. 3) has resulted in an incremental elevation of the score to 0.870, allowing the model to consider a broader span of text for richer contextual information, further refining its retrieval capabilities and improving performance to a certain extent. Incorporating a *reranking mechanism* (opt. 4) increased performance to 0.877. This demonstrates the value of a secondary evaluation layer (in addition to the vector-based search with embedded text chunks) in enhancing retrieval quality. Finally, the incorporation of *hybrid search methodologies* (opt. 5), which integrate multiple retrieval strategies, yielded the highest performance score 0.881, with a

processing time of 13.076 s. These results illustrate that including multiple advanced retrieval methods enhances the RAG system's performance, albeit with increased processing time. The above evaluation demonstrates the efficacy of a diversified approach to information retrieval, which employs a range of methodologies to identify the most pertinent data. This comprehensive evaluation illuminates the collective impact of these enhancements, demonstrating the substantial influence of each setting on the overall efficacy of the RAG model. Furthermore, it highlights the importance of a multifaceted optimization strategy in advanced retrieval systems.

To validate the statistical significance of our results, we conducted paired t-tests between consecutive optimizations. The results confirm that the improvements from basic to opt.4 are statistically significant ($p < 0.05$), demonstrating that each stepwise modification meaningfully contributes to performance gains. Specifically, the improvements from basic to opt.1 ($t = -5.8338$, $p < 0.0001$), opt.1 to opt.2 ($t = -5.0046$, $p < 0.0001$), opt.2 to opt.3 ($t = -3.3331$, $p = 0.0009$), and opt.3 to opt.4 ($t = -2.1074$, $p = 0.0356$) are all statistically significant, reinforcing the effectiveness of each optimization step. The improvement from opt.4 to opt.5 shows a trend towards significance ($t = -1.8296$, $p = 0.0679$), suggesting that hybrid search likely contributes to performance gains. While this result does not meet the conventional $p < 0.05$ threshold, it falls within the more lenient exploratory range ($p < 0.1$), indicating a potential benefit. These statistical results confirm that the proposed optimizations consistently improve system performance, with the strongest effects observed in earlier modifications and a diminishing effect size in later stages.

Additionally, an ablation (abl.) study was conducted to identify the most influential factor impacting the system's performance in response to various architectural and setting adjustments. The results of this study are presented in the lower section of Table 1 and are plotted in Fig. 4(b). This study aimed to investigate the impact of removing one of the five previously mentioned optimization methods. Each setting was identified by a check mark (✓) or cross (✗) symbol, indicating the presence or absence of that optimization method, respectively. Notably, abl. 1, which omits the configuration of chunks, shows a slight decrease in score to 0.873 but a significant reduction in time to 9.990 s, indicating a trade-off between performance and speed. Excluding the embedding optimization (abl. 2) substantially impacts performance, reducing the score to 0.845. Notably, removing the small-to-big retrieval method (abl. 3) causes a marginal score to decrease to 0.879. It nearly matches the optimal setting time at 12.984 s, suggesting that its role may be less

Table 1

Evaluation metrics of different RAG architecture and settings. There is a checkmark (✓) or a cross (✗) to indicate the presence or absence of this optimization method.

Setting	Chunk Config	Embedding	Small-to-big Retrieval	Rerank	Hybrid search	Score ±Std	Time (s) ±Std
basic	✗	✗	✗	✗	✗	0.761 ± 0.183	9.208 ± 0.559
opt. 1	✓	✗	✗	✗	✗	0.818 ± 0.132	10.647 ± 0.953
opt. 2	✓	✓	✗	✗	✗	0.864 ± 0.105	11.601 ± 1.077
opt. 3	✓	✓	✓	✗	✗	0.870 ± 0.102	11.873 ± 0.780
opt. 4	✓	✓	✓	✓	✗	0.877 ± 0.101	12.694 ± 0.958
opt. 5	✓	✓	✓	✓	✓	0.881 ± 0.097	13.076 ± 1.154
abl. 1	✗	✓	✓	✓	✓	0.873 ± 0.111	9.990 ± 0.819
abl. 2	✓	✗	✓	✓	✓	0.845 ± 0.114	10.415 ± 0.978
abl. 3	✓	✓	✗	✓	✓	0.879 ± 0.093	12.984 ± 1.033
abl. 4	✓	✓	✓	✗	✓	0.863 ± 0.106	11.754 ± 0.645
abl. 5	✓	✓	✓	✓	✗	0.877 ± 0.101	12.694 ± 0.958

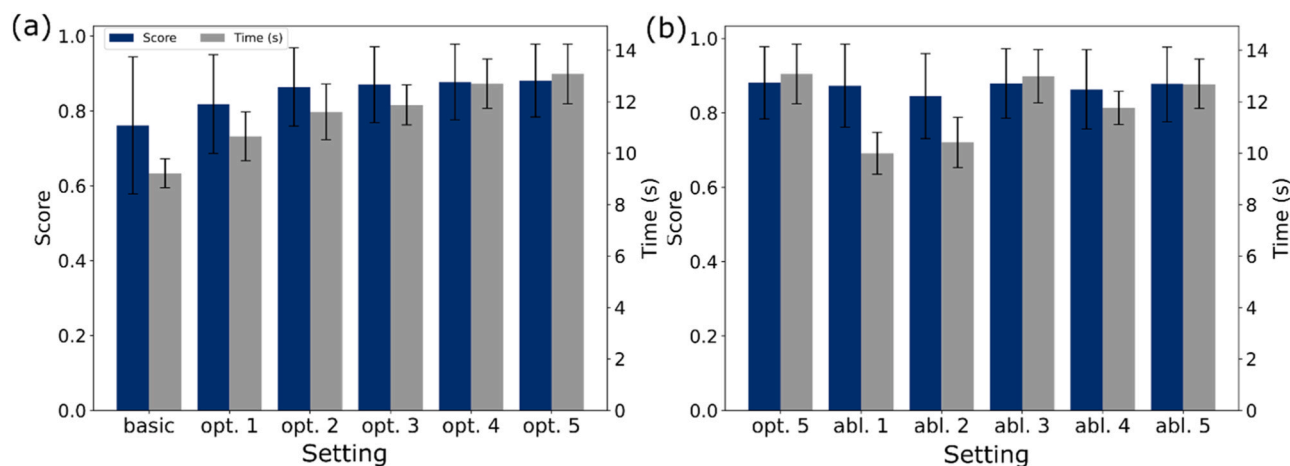


Fig. 6. (a) Performance measured in the score (dark blue bar) and response time (grey bar) of the system with different opt. (optimization) methods; (b) the influence of the absence of diverse optimization techniques (ablation study, short for abl.) on the performance score of the system.

critical in specific configurations. This implies that while it contributes positively to performance, its importance may be context-dependent and possibly compensated by other mechanisms within the system. The absence of reranking (abl. 4) also causes a more pronounced drop to 0.863 (with 11.754 s), highlighting the critical function of reranking in prioritizing the most relevant documents, thereby enhancing the overall quality of the generated responses. In contrast, abl. 5, which excludes hybrid search, demonstrates intermediate effects, with scores of 0.876 and times of 12.694 s. These findings underscore the importance of embeddings and reranking in achieving optimal performance while also showing the valuable contributions of chunk configs, context windows, and hybrid search in enhancing the RAG architecture, with specific configurations offering a balance between efficiency and effectiveness.

To further analyze the contribution of each optimization component, we also conducted paired t-tests comparing the final optimized setting (opt.5) with ablation variants, where a specific component was removed. The results indicate that removing embedding optimization (abl.2) leads to the most significant performance drop ($t = 3.8030$, $p = 0.0001$), confirming its crucial role in improving retrieval quality. Additionally, removing chunk configuration (abl.1) ($t = 2.5171$, $p = 0.0121$), and reranking (abl.4) ($t = 2.4310$, $p = 0.0154$), all result in statistically significant performance degradation ($p < 0.05$), demonstrating their importance in the system's effectiveness. In contrast, removing small-to-big retrieval (abl.3) does not lead to a statistically significant difference ($t = 1.5216$, $p = 0.1282$), suggesting that this component contributes only marginally to overall performance in this setting. These findings validate that embedding optimization, chunk configuration, and reranking are the most influential factors, while small-to-big and hybrid-search retrieval provides only minor benefits in this particular setup in our experiment.

In conclusion, this detailed evaluation elucidates the multifaceted nature of RAG architectures, providing valuable insights into how individual components contribute to the overall performance. These findings are critical for guiding future developments and optimizations in RAG systems, ensuring that each element is effectively leveraged to maximize the system's retrieval and generation capabilities. By systematically analyzing the impact of each configuration, this study offers a robust framework for understanding and enhancing the performance of complex RAG systems.

6. Discussion and outlook

In this work, we present an implementation of the RAG system tailored explicitly for scientific publications on solid-state batteries. This system demonstrates notable advancements in information retrieval and

knowledge synthesis for complex scientific literature. By meticulously parsing documents and loading relevant metadata, segmenting content into manageable chunks, and transforming these chunks into vector embeddings, the system ensures the creation of a comprehensive database for the subsequent retrieval process. The hybrid search mechanism, which combines dense and sparse retrieval methods, significantly enhances the precision and breadth of search results, accommodating both nuanced and broad queries. The subsequent reranking of retrieved chunks further refines the output by prioritizing the most relevant information. This layered retrieval and ranking strategy enables the language model to generate highly accurate and contextually enriched responses, effectively supporting nuanced scientific inquiry and synthesis.

Developing a local RAG research system for scientific papers also presents significant advantages over commercial products. Comparing our open-source solution with several commercial products (an example is detailed in S3 in Supplementary Information), we highlight several distinct advantages and conclude that commercial platforms designed for a broad audience often use generic settings that can lead to inaccurate or irrelevant answers when handling domain-specific queries. These systems may fail to capture the nuanced terminology and context of scholarly documents, resulting in less reliable answers. In contrast, a local RAG system can be tailored to the specific corpus and knowledge base relevant to the research, thereby increasing the accuracy and relevance of its output. In addition, commercial products often restrict access to full document references and offer comprehensive citation functionality only as a paid add-on. This limitation can impede the research process, making verifying sources difficult and building on existing work. On the other hand, a local RAG chatbot system can be configured to provide full references and detailed citations as a standard feature, ensuring that researchers have access to all the information they need at no additional cost. As a result, the local RAG system ensures greater accuracy and completeness, essential for rigorous scientific research and advancement.

In addition to the above, other benefits of the local open-source system include enhanced privacy and security, as a local system ensures that sensitive research data remains within institutional boundaries, thereby reducing the risks associated with data breaches and third-party access. Moreover, utilizing open-source technologies facilitates transparency and reproducibility, enabling researchers to adapt and optimize the chatbot according to their specific requirements without the constraints imposed by proprietary software. This approach facilitates access to sophisticated AI tools and encourages collaborative innovation and collective problem-solving within the academic community. The present study demonstrates the efficacy of an open-source

RAG-based assistant, showcasing its robust performance and potential to streamline the research workflow significantly. By illustrating a practical implementation, this study highlights the feasibility and benefits of adopting open-source AI solutions in academia, ultimately contributing to more secure, transparent, and efficient scientific research practices.

While LLMs represent a significant advancement in artificial intelligence and natural language processing, it must be acknowledged that they have limitations and that human judgment and expertise play an essential role, particularly in specialized fields [54]. LLMs are particularly adept at processing and generating vast amounts of text, identifying patterns, and providing insights from extensive datasets. Nevertheless, they need to improve their capacity for nuanced comprehension, contextual awareness, and critical thinking, characteristic of human experts in their respective disciplines. It is therefore recommended that LLMs be employed as supplementary tools that enhance human capabilities, thus facilitating deeper comprehension and analysis.

Nevertheless, one has to be aware of the specific limitations of our approach and similar RAG systems:

- **Potential biases in LLM training data:** Large language models (LLMs) are trained on extensive and diverse corpora, which predominantly consist of general-domain text. As a result, they may not be specifically optimized for highly specialized fields such as solid-state battery research. This can lead to knowledge gaps, biases, or inaccuracies when generating responses related to niche scientific concepts. While RAG improves domain-specific accuracy by integrating external knowledge sources, technical jargon and evolving terminology in battery research still pose challenges. LLMs may struggle with disambiguating similar terms, correctly interpreting abbreviations, or adapting to emerging scientific concepts that were not well represented in their pretraining data.
- **Retrieval limitations:** Vector similarity search based on embeddings, which forms the basis of our RAG system, has inherent limitations when handling conceptually related but lexically dissimilar content. Technical concepts expressed in different terminology across disciplines may not be identified as semantically similar by embedding models, potentially missing valuable cross-disciplinary connections.
- **Knowledge boundaries:** LISA's knowledge is strictly limited to the documents in the corpus we provide. Unlike generative capabilities of base LLMs, our system cannot synthesize knowledge outside its retrieval database, making comprehensive document collection critical.
- **Specialized terminology challenges:** While RAG helps ground responses in domain-specific terminology from the corpus, embeddings may struggle with highly specialized or niche terminology, especially when terms have multiple context-dependent meanings across disciplines.
- **Evaluation constraints:** Our evaluation dataset, while diverse, cannot comprehensively cover all possible interdisciplinary queries that might arise in a real research environment. The performance metrics we report may not generalize equally across all subdomains of battery research.
- **Document segmentation trade-offs:** Our chunking approach inevitably creates artificial boundaries in the knowledge corpus. While we've optimized chunk size (as demonstrated), some contextual information may still be lost at chunk boundaries.

Based on all the discussed limitations, the objective of AI integration is to support, rather than replace, professionals' expert knowledge and analytical abilities. This ensures that AI integration enhances, rather than diminishes, the quality and accuracy of specialized work. In essence, one can interact directly with all documents, facilitating the rapid and convenient retrieval of specific information. It is important to note that this is not a replacement for personal discussions and meetings. However, it does provide an accessible option for knowledge

distribution and finding experts in the field to initiate new collaborations.

In addition to its core capabilities, LISA is equipped with advanced features such as web search integration and knowledge graph-enhanced RAG, enabling real-time searching from sources outside the internal database and improved structured information retrieval [37]. These features enhance LISA's ability to provide timely and comprehensive insight beyond its database and improve performance. However, due to considerations such as ensuring a fair comparison with the methods used in this work, avoiding excessive time and resource consumption in building a robust knowledge graph from the documents in its database, and maintaining focus on the primary defragmentation functionality, these additional features are detailed in the [Supplementary Information](#).

7. Conclusion

In conclusion, this work presents a robust Retrieval-Augmented Generation (RAG) system tailored for the retrieval and synthesis of information from scientific publications on solid-state batteries. This RAG system, integrated with advanced language model capabilities linked to external knowledge bases, demonstrates an advancement in retrieval technology, enabling streamlined access to complex scientific knowledge and supporting researchers in efficient information extraction. Our findings underscore the potential of RAG systems to enhance research workflows and advance knowledge dissemination within the scientific community, particularly in fast-evolving fields such as solid-state battery technology.

CRedit authorship contribution statement

Hansen Anna-Lena: Writing – review & editing, Writing – original draft, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Zhao Yinghan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Ehrenberg Helmut:** Writing – review & editing, Project administration. **Knapp Michael:** Writing – review & editing, Validation. **Selzer Michael:** Writing – review & editing, Funding acquisition. **Brandt Nico:** Writing – review & editing, Software. **Nestler Britta:** Writing – review & editing, Project administration. **Koeppe Arnd:** Writing – review & editing, Resources, Methodology, Conceptualization. **Dahlhaus Anna:** Writing – review & editing, Software.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The support from the following organizations and institutions is gratefully acknowledged: The "FESTBATT" consortium (grant number: 03XP0435D), funded by the German Federal Ministry of Education and Research (BMBF). This work also contributes to the research conducted at CELEST (Center for Electrochemical Energy Storage Ulm-Karlsruhe) and received partial funding from the German Research Foundation (DFG) under Germany's Excellence Strategy – EXC 2154 – Project number 390874152 (POLiS Cluster of Excellence). The authors also acknowledge funding from the Helmholtz Association through KNMFI and Joint Lab MDMC, programme no. 43.31.01.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.mtcomm.2025.112380](https://doi.org/10.1016/j.mtcomm.2025.112380).

Data availability

The running demo can be accessed via the following link: <https://huggingface.co/spaces/Kadi-IAM/LISA>. All datasets and custom code used in this study will be made available in the same repository upon acceptance of the manuscript for publication. For additional information or specific requests, please contact the corresponding author.

References

- [1] B. Wu, Addressing the battery talent shortage with interdisciplinarity, *Nat. Energy* 9 (2024) 1044–1045, <https://doi.org/10.1038/s41560-024-01576-w>.
- [2] Web of Science, Web of Science platform, 2024. (<https://www.webofscience.com>) (accessed January 19, 2024).
- [3] A. Karlqvist, Going beyond disciplines: the meanings of interdisciplinarity, *Policy Sci.* 32 (1999) 379–383.
- [4] Nomenclature of Inorganic Chemistry – IUPAC Recommendations 2005, *Chem. Int. Newsmag.* IUPAC 27 (2005), <https://doi.org/10.1515/ci.2005.27.6.25>.
- [5] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, *Adv. Neural Inf. Process. Syst.* (2020), <https://doi.org/10.48550/arXiv.2005.14165>.
- [6] W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A Surv. Large Lang. Models, Prepr. (2023), <https://doi.org/10.48550/arXiv.2303.18223>.
- [7] FestBatt, The BMBF competence cluster for solid-state batteries, 2024. (<http://festbatt.net/>) (accessed September 30, 2024).
- [8] D. Atkins, E. Capria, K. Edström, T. Famprikis, A. Grimaud, Q. Jacquet, M. Johnson, A. Matic, P. Norby, H. Reichert, J.-P. Rueff, K. C. Villeveuille, M. Wagemaker, S. Lyonnard, Accelerating battery characterization using neutron and synchrotron techniques: toward a multi-modal and multi-scale standardized experimental workflow, *Adv. Energy Mater.* 12 (2022) 2102694, <https://doi.org/10.1002/aenm.202102694>.
- [9] N. Brandt, L. Griem, C. Herrmann, E. Schoof, G. Tosato, Y. Zhao, P. Zschumme, M. Selzer, Kadi4Mat: a research data infrastructure for materials, *Sci., Data Sci. J.* 20 (2021), <https://doi.org/10.5334/dsj-2021-008>.
- [10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, Preprint (2021), <https://doi.org/10.48550/arXiv.2005.11401>.
- [11] X. Zhang, L. Wang, J. Helwig, Y. Luo, C. Fu, Y. Xie, M. Liu, Y. Lin, Z. Xu, K. Yan, K. Adams, M. Weiler, X. Li, T. Fu, Y. Wang, H. Yu, Y. Xie, X. Fu, A. Strasser, S. Xu, Y. Liu, Y. Du, A. Saxton, H. Ling, H. Lawrence, H. Stärk, S. Gui, C. Edwards, N. Gao, A. Ladera, T. Wu, E.F. Hofgard, A.M. Tehrani, R. Wang, A. Daigavane, M. Bohde, J. Kurtin, Q. Huang, T. Phung, M. Xu, C.K. Joshi, S.V. Mathis, K. Azizzadenesheli, A. Fang, A. Aspuru-Guzik, E. Bekkers, M. Bronstein, M. Zitnik, A. Anandkumar, S. Ermon, P. Liò, R. Yu, S. Günemann, J. Leskovec, H. Ji, J. Sun, R. Barzilay, T. Jaakkola, C.W. Coley, X. Qian, X. Qian, T. Smidt, S. Ji, Artificial Intelligence for Science in Quantum, Atomistic, and Continuum Systems, Preprint, 2023. <https://doi.org/10.48550/arXiv.2307.08423>.
- [12] G. Lei, R. Docherty, S.J. Cooper, Materials science in the era of large language models: a perspective, *Digit. Discov.* 3 (2024) 1257–1272, <https://doi.org/10.1039/D4DD00074A>.
- [13] M. Zaki, J. Mausam, N.M.A. Krishnan, MaScQA: investigating materials science knowledge of large language models, *Digit. Discov.* 3 (2024) 313–327, <https://doi.org/10.1039/D3DD00188A>.
- [14] Y. Liu, Z. Yang, Z. Yu, Z. Liu, D. Liu, H. Lin, M. Li, S. Ma, M. Avdeev, S. Shi, Generative artificial intelligence and its applications in materials science: current situation and future perspectives, *J. Mater.* 9 (2023) 798–816, <https://doi.org/10.1016/j.jmat.2023.05.001>.
- [15] K.M. Jablonka, Q. Ai, A. Al-Feghali, S. Badhwar, J.D. Bocarsly, A.M. Bran, S. Bringuier, L.C. Brinson, K. Choudhary, D. Circi, S. Cox, W.A. De Jong, M. L. Evans, N. Gastellu, J. Genzling, M.V. Gil, A.K. Gupta, Z. Hong, A. Imran, S. Kruschwitz, A. Labarre, J. Lala, T. Liu, S. Ma, S. Majumdar, G.W. Merz, N. Moitessier, E. Moubarak, B. Mourinho, B. Pelkie, M. Pieler, M.C. Ramos, B. Ranković, S.G. Rodrigues, J.N. Sanders, P. Schwaller, M. Schwarting, J. Shi, B. Smit, B.E. Smith, J. Van Herck, C. Völker, L. Ward, S. Warren, B. Weiser, S. Zhang, X. Zhang, G.A. Zia, A. Scourtas, K.J. Schmidt, I. Foster, A.D. White, B. Blaiszik, 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon, *Digit. Discov.* 2 (2023) 1233–1250, <https://doi.org/10.1039/D3DD00113J>.
- [16] K.M. Jablonka, P. Schwaller, A. Ortega-Guerrero, B. Smit, Leveraging large language models for predictive chemistry, *Nat. Mach. Intell.* 6 (2024) 161–169, <https://doi.org/10.1038/s42256-023-00788-1>.
- [17] A.N. Rubungo, C. Arnold, B.P. Rand, A.B. Dieng, LLM-Prop.: Predict. Phys. Electron. Prop. Cryst. Solids Their Text. Descr., Prepr. (2023), <https://doi.org/10.48550/arXiv.2310.14029>.
- [18] T. Xie, Y. Wan, W. Huang, Y. Zhou, Y. Liu, Q. Linghu, S. Wang, C. Kit, C. Grazian, W. Zhang, B. Hoex, Large Lang. Models Master Key: Unlocking Secrets Mater. Sci. GPT, Prepr. (2023), <https://doi.org/10.48550/arXiv.2304.02213>.
- [19] S. Balaji, R. Magar, Y. Jadhav, A.B. Farimani, GPT-MolBERTa: GPT Mol. Features Lang. Model Mol. Prop. Predict., Prepr. (2023), <https://doi.org/10.48550/arXiv.2310.03030>.
- [20] S. Miret, N.M.A. Krishnan, Are LLMs Ready Real. -World Mater. Discov., Prepr. (2024), <https://doi.org/10.48550/arXiv.2402.05200>.
- [21] J. Boyko, J. Cohen, N. Fox, M.H. Veiga, J.I.-H. Li, J. Liu, B. Modenesi, A.H. Rauch, K.N. Reid, S. Tribedi, A. Visheratina, X. Xie, Interdiscip. Outlook Large Lang. Models Sci. Res., Prepr. (2023), <https://doi.org/10.48550/arXiv.2311.04929>.
- [22] A. Mirza, N. Alampara, S. Kunchapu, B. Emoeakabu, A. Krishnan, M. Wilhelm, M. Okereke, J. Eberhardt, A.M. Elahi, M. Greiner, C.T. Holick, T. Gupta, M. Asgari, C. Glaubitz, L.C. Klepsch, Y. Köster, J. Meyer, S. Miret, T. Hoffmann, F.A. Kreth, M. Ringleb, N. Roesner, U.S. Schubert, L.M. Stafast, D. Wonanke, M. Pieler, P. Schwaller, K.M. Jablonka, Are Large Lang. Models Superhuman Chem., Prepr. (2024), <https://doi.org/10.48550/arXiv.2404.01475>.
- [23] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A.D. White, P. Schwaller, Augmenting large language models with chemistry tools, *Nat. Mach. Intell.* 6 (2024) 525–535, <https://doi.org/10.1038/s42256-024-00832-8>.
- [24] M. Jalali, Y. Luo, L. Caulfield, E. Sauter, A. Nefedov, C. Wöll, Large language models in electronic laboratory notebooks: transforming materials science research workflows, *Mater. Today Commun.* 40 (2024) 109801, <https://doi.org/10.1016/j.mtcomm.2024.109801>.
- [25] D.A. Boiko, R. MacKnight, B. Kline, G. Gomes, Autonomous chemical research with large language models, *Nature* 624 (2023) 570–578, <https://doi.org/10.1038/s41586-023-06792-0>.
- [26] E.A. Olivetti, J.M. Cole, E. Kim, O. Kononova, G. Ceder, T.Y.-J. Han, A. M. Hiszpanski, Data-driven materials research enabled by natural language processing and information extraction, *Appl. Phys. Rev.* 7 (2020) 041317, <https://doi.org/10.1063/5.0021106>.
- [27] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A.S. Rosen, G. Ceder, K.A. Persson, A. Jain, Structured information extraction from scientific text with large language models, *Nat. Commun.* 15 (2024) 1418, <https://doi.org/10.1038/s41467-024-45563-x>.
- [28] Q. Ai, F. Meng, J. Shi, B. Pelkie, C.W. Coley, Extracting structured data from organic synthesis procedures using a fine-tuned large language model, *Digit. Discov.* 3 (2024) 1822–1831, <https://doi.org/10.1039/D4DD00091A>.
- [29] J. Choi, B. Lee, Accelerating materials language processing with large language models, *Commun. Mater.* 5 (2024) 1–11, <https://doi.org/10.1038/s43246-024-00449-9>.
- [30] M.P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, *Nat. Commun.* 15 (2024) 1569, <https://doi.org/10.1038/s41467-024-45914-8>.
- [31] S.J. Yang, S. Li, S. Venugopalan, V. Tshitoyan, M. Aykol, A. Merchant, E.D. Cubuk, G. Cheon, Accurate Predict. Exp. Band. Gaps Large Lang. Model-Based Data Extr., Prepr. (2023), <https://doi.org/10.48550/arXiv.2311.13778>.
- [32] Z. Zheng, Z. He, O. Khattab, N. Rampal, M.A. Zaharia, C. Borgs, J.T. Chayes, O. M. Yaghi, Image and data mining in reticular chemistry powered by GPT-4V, *Digit. Discov.* 3 (2024) 491–501, <https://doi.org/10.1039/D3DD00239J>.
- [33] N. Kandpal, H. Deng, A. Roberts, E. Wallace, C. Raffel, Large Lang. Models Struggle Learn Long. -Tail Knowl., Prepr. (2023), <https://doi.org/10.48550/arXiv.2211.08411>.
- [34] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retr. -Augment. Gener. Large Lang. Model.: A Surv., Prepr. (2024), <https://doi.org/10.48550/arXiv.2312.10997>.
- [35] OpenAI, Improving Language Understanding by Generative Pre-Training, 2018. (<https://openai.com/index/language-unsupervised/>) (accessed December 12, 2024).
- [36] O. Ovadia, M. Brief, M. Mishaeli, O. Elisha, Fine-Tuning Or. Retr. ? Comp. Knowl. Inject. LLMs, Prepr. (2024), <https://doi.org/10.48550/arXiv.2312.05934>.
- [37] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, J. Larson, Local Glob.: A Graph RAG Approach Query-Focus. Summ., Prepr. (2024), <https://doi.org/10.48550/arXiv.2404.16130>.
- [38] Z. Zheng, O. Zhang, C. Borgs, J.T. Chayes, O.M. Yaghi, ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis, *J. Am. Chem. Soc.* 145 (2023) 18048–18062, <https://doi.org/10.1021/jacs.3c05819>.
- [39] M.J. Buehler, Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design, *ACS Eng. Au* 4 (2024) 241–277, <https://doi.org/10.1021/acengineeringau.3c00058>.
- [40] P. Finardi, L. Avila, R. Castaldoni, P. Gengo, C. Larcher, M. Piau, P. Costa, V. Caridà, Preprint, Chron. RAG Retrieval., *Chunk Gener.* (2024), <https://doi.org/10.48550/arXiv.2401.07883>.
- [41] X. Wang, Z. Wang, X. Gao, F. Zhang, Y. Wu, Z. Xu, T. Shi, Z. Wang, S. Li, Q. Qian, R. Yin, C. Lv, X. Zheng, X. Huang, Search. Best. Pract. Retr. Augment. Gener. Prepr. (2024), <https://doi.org/10.48550/arXiv.2407.01219>.
- [42] M. Glass, G. Rossiello, M.F.M. Chowdhury, A.R. Naik, P. Cai, A. Gliozzo, Re2G: Retrieve, Rerank, Gener., Prepr. (2022), <https://doi.org/10.48550/arXiv.2207.06300>.
- [43] S. Zhuang, B. Liu, B. Koopman, G. Zuccon, Open-Source Large Lang. Models are Strong Zero-shot Query Likelihood Models Doc. Rank., Prepr. (2023), <https://doi.org/10.48550/arXiv.2310.13243>.
- [44] N.F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost Middle.: How Lang. Models Use Long. Contexts, Prepr. (2023), <https://doi.org/10.48550/arXiv.2307.03172>.

- [45] P. Christen, D.J. Hand, N. Kirielle, A review of the f-measure: its history, properties, criticism, and alternatives, *ACM Comput. Surv.* 56 (2023) 73:1–73:24, <https://doi.org/10.1145/3606367>.
- [46] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation. in: *Proc. 40th Annu. Meet. Assoc. Comput. Linguist. Association for Computational Linguistics, USA, 2002*, pp. 311–318, <https://doi.org/10.3115/1073083.1073135>.
- [47] C.-Y. Lin, ROUGE: A Package for Automatic Evaluation of Summaries. in: *Text Summ. Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004*, pp. 74–81. (<https://aclanthology.org/W04-1013>).
- [48] X. Li, J. Jin, Y. Zhou, Y. Zhang, P. Zhang, Y. Zhu, Z. Dou, Matching Gener. A Surv. *Gener. Inf. Retr. Prepr.* (2024), <https://doi.org/10.48550/arXiv.2404.14851>.
- [49] J. Saad-Falcon, O. Khattab, C. Potts, M. Zaharia, ARES: Autom. Eval. Framew. Retr. -Augment. Gener. Syst., *Prepr.* (2024), <https://doi.org/10.48550/arXiv.2311.09476>.
- [50] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, RAGAS Autom. Eval. Retr. Augment. Gener. Prepr. (2023), <https://doi.org/10.48550/arXiv.2309.15217>.
- [51] TruLens, Eval. Track LLM Appl. (2024). (<https://www.trulens.org>) (accessed June 11, 2024).
- [52] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embed. Using Siamese BERT-Netw. Prepr. (2019), <https://doi.org/10.48550/arXiv.1908.10084>.
- [53] A.Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E.B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L.R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T.L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W.E. Sayed, Mixtral Experts, *Prepr.* (2024), <https://doi.org/10.48550/arXiv.2401.04088>.
- [54] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, M. Abdelrazek, Seven Fail. Points When Eng. a Retr. Augment. Gener. Syst., *Prepr.* (2024), <https://doi.org/10.48550/arXiv.2401.05856>.