Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2026, **5**, 1401

Chat-RFB: a flow battery chat system leveraging knowledge graphs and large language models

Hao-Tian Wang,[ab] Xuefeng Bai,[ab] Zhiling Zheng,[c] Xin Zhang,[ab] Ruipeng Jin,[ab] Hao-Tian An,[ab] Zheng-He Xie,[abd] Xiu-Liang Lv *[ab] and Jian-Rong Li *[ab]

Received 10th November 2025
Accepted 12th February 2026

DOI: 10.1039/d5dd00494b

rsc.li/digitaldiscovery

The interdisciplinary nature of redox flow batteries (RFBs), spanning chemistry, materials science, and engineering, has led to a vast and fragmented body of research, hindering the efficient synthesis of knowledge. An intelligent question-answering system is therefore essential to organize this dispersed knowledge, enhance information retrieval, and lower the barrier to comprehensive understanding. In this study, we leveraged the natural language processing capabilities of large language models (LLMs) and the structured nature of knowledge graphs (KGs) to establish a chat model in the field of RFBs, named Chat-RFB. By analyzing 5353 articles related to flow batteries and deconstructing the text content, we learned contextual relationships and generated nearly 164 232 nodes, constructing 853 939 relationships among nodes. This process enhances the professional domain knowledge question-answering ability of LLMs. Given the limited research on the responsiveness of evaluation models in the flow battery field, we conducted model performance evaluations using both choice and non-choice questions. The results indicate that by incorporating a professional knowledge base, Chat-RFB enhanced the level of professional domain knowledge. Choice question accuracy was: Chat-RFB 94.9%, DeepSeek-v3 90.9%, GPT-4o 90.7%, Qwen-Max 90.4%, and Gemini-2.5-Flash 91.1%. Non-choice question accuracy was: Chat-RFB 93.3%, DeepSeek-v3 73.3%, GPT-4o 68.9%, Qwen-Max 75.6%, and Gemini-2.5-Flash 86.7%.

Introduction

With the rapid advancement of renewable energy technologies, including uninterruptible power supplies, emergency backup systems, and smart grid applications,[1] the demand for large-scale battery energy storage systems has been steadily increasing.[2,3] Among the various grid-connected rechargeable energy storage solutions, redox flow batteries (RFBs) have emerged as particularly promising due to their distinctive advantages. RFBs represent a critical and inherently interdisciplinary area within energy storage research, requiring a breadth of knowledge from chemistry,[4–7] materials science,[8–10] and systems engineering.[11,12] They are characterized by a flexible, modular design that allows for the independent adjustment of storage capacity and power output, along with a long cycle life and high safety profile.[13–15] These features have attracted significant research interest and exploration in the field of RFB technology.[16–20] The burgeoning research on flow batteries has yielded a substantial body of valuable and reliable data in the scientific literature. However, manually sifting through large volumes of publications to extract fundamental information about the materials presents a formidable challenge, potentially hindering researchers' efficiency and increasing the time investment required for the research process.

Large Language Models (LLMs) exhibit strong language comprehension abilities, having been pre-trained on extensive corpora, which enable them to automatically extract,[21–23] semantically analyze,[24] and logically reason[25] literature content in response to users' natural language queries. Owing to their generalization, multitasking capabilities, and contextual understanding, LLMs hold promise for integration into agentic systems,[26] where they can act as intelligent agents to assist with scientific research.[27,28] However, it is important to note that LLMs face critical limitations for scientific applications. Their training on general knowledge datasets limits their expertise in specialized domains.[29,30] Furthermore, the finite context window of LLMs restricts the amount of information they can process at once, which can lead to significant information loss or "contextual forgetting" during the analysis of long documents. Conventional search processes are similarly limited, often retrieving only superficial information from article abstracts. To address these challenges, the Retrieval-Augmented Generation (RAG)[31] technology of artificial intelligence frameworks has emerged in response to the times. RAG enables the system to precisely retrieve the most relevant information fragments from the knowledge base based on user input. Subsequently, LLMs utilize

[a]State Key Laboratory of Materials Low-Carbon Recycling, Beijing University of Technology, Beijing 100124, China. E-mail: jrli@bjut.edu.cn; lvxiuliang@bjut.edu.cn
[b]Department of Chemical Engineering, College of Materials Science and Engineering, Beijing University of Technology, Beijing 100124, China
[c]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
[d]Beijing Energy Holding Co., Ltd, Beijing 100022, PR China

the retrieved information as context to generate more accurate, fact-based, and enriched responses. For example, ChemReactSeek[32] is an artificial intelligence platform for heterogeneous hydrogenation reactions built using a text vectorization-based RAG method. RAG frameworks that incorporate Knowledge Graphs (KGs) have emerged as a promising solution.[33–39] While any form of structured data can enhance information density for LLMs, KGs are uniquely suited due to their ability to model the complex, multi-relational nature of scientific knowledge—a capability not fully matched by simpler structured formats. KGs utilize graphical models to delineate entities, concepts, and their interrelationships, providing a clear logical structure that is essential for the deep reasoning and relationship traversal required in scientific question-answering, thus facilitating the understanding of complex concepts.[37,40–42] In addition, by highly summarizing the key information extracted from the entire text, KGs can provide a clear logical structure, which helps to obtain important node information in a limited space and accelerate scientific discoveries. In particular, the integration of LLMs with KGs has been explored in fields such as medicine,[43] materials science,[35,44] and chemistry.[33,39] However, constructing an end-to-end framework that spans automated knowledge extraction, domain-specific knowledge graph construction, and RAG-integrated question answering—all supported by systematic evaluation—remains a cutting-edge challenge. Our work is dedicated to precisely this endeavor. By implementing and validating such a comprehensive framework within the field of redox flow batteries, we demonstrate its immense potential as a next-generation assistant for scientific research.

To address this gap, we developed Chat-RFB. This domain-specific intelligent assistant integrates an LLM with a structured knowledge base, enabling accurate and efficient retrieval of knowledge related to flow batteries. In this study, we employed an LLM to parse text and extract keywords from an extensive corpus of over 5000 relevant studies of literature in a high-throughput manner. The procedure yielded a KG comprising 164 232 nodes and 853 939 relationships, with each article's Digital Object Identifier (DOI) serving as a unique identifier. We developed an automated test set to evaluate our system's capacity for expert-level analysis and summarization within the flow battery domain. Testing revealed that the performance of our system, Chat-RFB, surpasses that of the native model. And during the testing process, by establishing a knowledge graph of node information throughout the text,

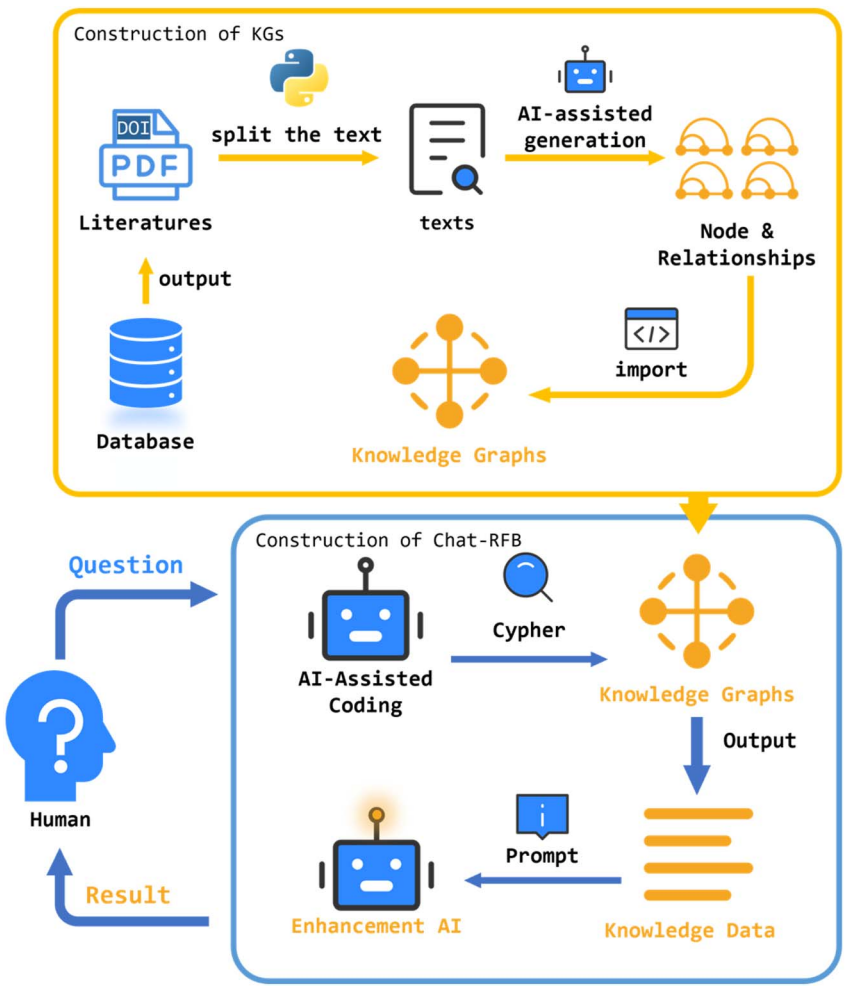

Fig. 1 KG enhanced LLM system in the field of flow batteries (Chat-RFB) model construction and application workflow. The Chat-RFB workflow follows two stages: the construction of KGs is shown in the yellow square and the construction of Chat-RFB is shown in the blue square.

some important information in the text that cannot be summarized can be queried, which effectively broadens the system's usage scenarios. The workflow of Chat-RFB is illustrated in Fig. 1. Technical implementation and applications are detailed in the subsequent sections.

Methods

Collecting information from relevant literature

To identify relevant journal articles on flow batteries, we conducted a comprehensive search of the Web of Science database, collecting all publications available up to May 2025. In order to make the search cover relevant literature in the field as widely as possible, we used "flow battery" as the keyword search: "TI = (flow battery)".

From the Web of Science database, we exported the DOI information for all retrieved papers. Leveraging these DOIs, the full text of the literature was subsequently downloaded locally using high-throughput methods. Specifically, we employed a customized Python script utilizing requests and langchain_text_splitters libraries to automate the retrieval of PDF files from publisher websites. Given the model's context token input limitations, long texts were segmented into manageable chunks. To achieve this, we utilized a character-based text splitter (CharacterTextSplitter), configuring it to divide the text based on space characters. Each chunk was set to a maximum size of 20 000 characters, with an overlap of 500 characters between adjacent chunks. This approach ensures that semantic continuity is maintained across segment boundaries while respecting the model's input limits (Fig. S2 and S3). DeepSeek-v3 was then employed for high-throughput extraction of node and relationship information from these segmented text paragraphs. For the convenience of subsequent retrieval, we use the prompt engineering method to constrain LLMs to add label information on node outputs, which facilitates retrieval and increases the information content of nodes. The task and output format are defined in the prompt words (Fig. S1), and a formatted JSON file is generated (Fig. S4 and S5). To quantitatively assess the quality of the information extraction, we established a rigorous manual evaluation process conducted by domain experts. We framed the task as a decision-making process: for each potential piece of knowledge in the source text, the model must decide whether to extract it as an accurate and relevant relationship. This framing allows for the application of a standard evaluation framework based on the concepts of true positives, false positives, and false negatives.

The evaluation was performed by two researchers with expertise in the RFB field, who independently assessed a random sample of 500 extracted node-relationship triplets against their original source literature. Any disagreements were resolved through discussion to reach a consensus. The evaluation criteria were defined as follows:

• True positive (TP): an extracted relationship is considered a TP if it is both factually accurate according to the source text and represents a meaningful, key piece of information relevant to the RFB domain.

• False positive (FP): an extracted relationship is considered an FP if it is factually incorrect, a misinterpretation of the source text, or a hallucination not present in the literature.

• False negative (FN): a FN occurs when a key, explicit piece of information clearly stated in the source text was not extracted by the model.

This framework allows us to calculate standard metrics such as precision and recall to robustly evaluate the performance of our knowledge extraction pipeline.

KG construction and usage

To build a structured and reliable dataset for subsequent analysis, we first integrate all JSON files into raw comma separated value (CSV) files, and then extract target list information from the CSV files. This data extraction process was conducted using standard parsing techniques to ensure accurate retrieval of all relevant entries. To facilitate precise tracking and referencing of individual records, each entry was then assigned a unique identifier (UUID), ensuring global uniqueness across the dataset. Duplicate entries were identified and removed through a systematic comparison of UUIDs across all records, resulting in a clean, non-redundant dataset. The deduplication process was critical in eliminating redundancy-induced biases and enhancing the overall integrity and analytical validity of the dataset. The KGs were built using Neo4j software. We used the import command of Neo4j admin to batch process CSV files and create a database. Then we called the graphics database through Neo4j and visualized and analyzed the data. The Neo4j version we used in this study is 5.25.1.

Integration of LLM with the KGs

The RAG process using LLMs divides the question-answering process into four steps, as illustrated in Fig. 2:

(1) Generate Cypher queries: first of all, the LLM extracts keywords from the user's questions to identify inference nodes or relational information. It constructs Cypher query statements to access the Neo4j database and retrieve relevant information.

(2) Execute queries and retrieve data: these Cypher queries are executed on the Neo4j database, retrieving data pertinent to the user's queries from the KGs.

(3) Generate answers: the LLM uses the retrieved KG data as a prompt to comprehend and analyze user questions, developing precise and professional responses.

(4) Generate new conversation content: due to the LLM's context length limitation, it is not possible to save an unlimited number of data query results from previous interactions as part of the conversation history. When starting new conversation content, the system deletes previous search data while retaining earlier conversation content. Upon receiving a new question, steps 1–3 are repeated.

LLMs are accessed *via* the OpenAI extension library API, with specific version details provided in Table S1. Apart from the context length limitation, all models share the same parameters.

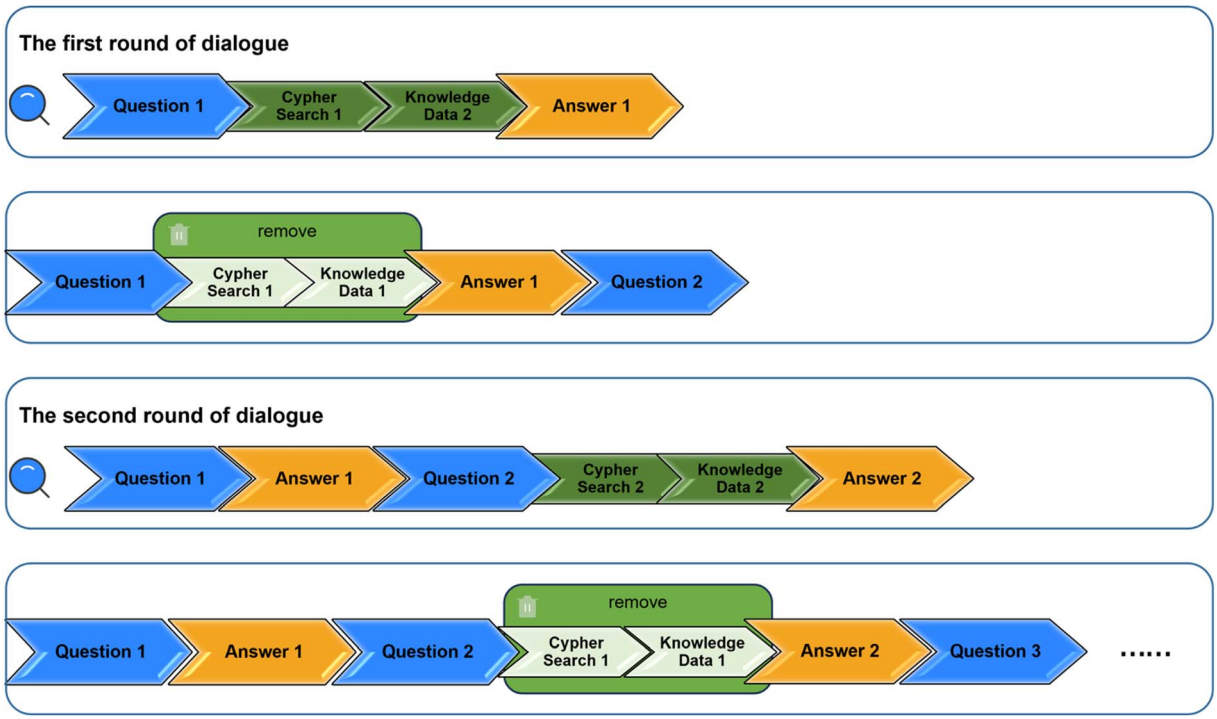

Fig. 2 Multi round dialogue context stitching method. In order to overcome the contextual limitations of the model, the data information used in the previous text was removed.

Evaluation of the LLM-KGs system

In order to comprehensively evaluate the knowledge, intuition, and reasoning abilities of the developed Large Language Model (LLM) in the field of chemistry, ChemBench[45] serves as a benchmark dataset for evaluating chemical abilities, which can effectively assess an LLM's application to a wide range of general science topics such as molecular properties, chemical reactions, materials science, and modeling design. However, although ChemBench covers a wide range of chemical topics, its design is not specifically targeted at the specific and highly interdisciplinary field of flow batteries. To evaluate the performance of KG enhanced LLM in the field of flow batteries, we designed a test set consisting of flow-battery-related problems. The test set includes 450 choice questions (203 true/false and 247 choice questions) and 45 non-choice questions. Randomly batch read literature through a large language model, and generate all questions based on the content of the literature. The questions cover a comprehensive study of flow batteries from basic materials science (molecular design, electrolyte chemistry, and electrode materials) to electrochemical performance (redox, kinetics, and decay mechanisms), to systems engineering (flow field design, energy density enhancement, and thermal management), as well as application level (economy, environmental friendliness, safety, and extreme condition adaptability), and emphasize the important role of characterization analysis techniques in revealing mechanisms and optimizing performance. These questions and their reference answers can be accessed in SI2. To ensure fairness and rigor of the test set, the prepared questions are rigorous, and all answers to the questions are sourced from the literature. While systematically evaluating choice questions, we further examined the performance of the model using 45 non-choice questions in order to be close to everyday conversational language. We evaluate LLM responses with scoring criteria. During evaluation, in order to facilitate the analysis of LLM usability, if the model clearly outputs incorrect information, it is considered that the output answer is unqualified. The evaluation criteria are as follows:

(1) Excellent (completely correct): the answer fully meets the question's requirements, is accurate, logically clear, and perfectly resolves the questioner's doubt.

(2) Good (partly correct): the answer is somewhat correct but may only cover some of the question's key points or have slight inaccuracies.

(3) Normal (no obvious errors, but the answer is too broad and lacks professionalism): the answer has no obvious errors but is too general and lacks the necessary depth and focus on the core of the question.

(4) Poor (obvious errors exist): the answer contains obvious misinformation that may mislead the questioner and fail to solve the problem.

For the convenience of evaluating the model's capability, we consider answers that do not contain serious errors (with a level higher than poor) to be qualified.

Results and discussion

Construction of Chat-RFB

In the data preprocessing stage, we retrieved and collected 5353 highly relevant academic papers on flow batteries that have

been published. Through text conversion, the content of the literature was segmented and converted, and stored in .txt files for easy retrieval by the LLM, without affecting the semantic structure of the article.[46] After that, we used DeepSeek-v3 to perform deep parsing on these files, identifying key information and organizing it into nodes with labels and relationships. Finally, we converted text content into relational data format, resulting in 164 232 nodes and 853 939 relationship links.

As shown in Fig. S2–S5, the LLM can accurately identify the core elements of an article, including research content, research methods, and theoretical concepts. Through customized prompt engineering, this information was successfully converted into a structured JSON format.[47] This step is crucial for the construction of a KG, as it directly affects the quality and accuracy of nodes in the graph.

To further ensure quality, we manually reviewed 500 nodes and their related relationships, as extracted by the LLM.[48] We evaluated the effectiveness of the LLM in extracting text entities during the process (Fig. S6). The model achieved a true positive rate (TP) of 97.8% in accurate and comprehensive information extraction, with an accuracy rate of 2% in false negative (FN) examples where key information is missing, and a false positive (FP) example rate of 0.2% in inaccurate information extraction. After comprehensive evaluation, the F1 score was as high as 0.9889. The LLM performed well in the task of extracting text information from liquid flow battery articles given prompt engineering, which were subsequently used in large-scale automated extraction processes. Subsequently, we annotated the information in bulk and organized it into the Neo4j database. Ultimately, we can obtain an LLM intelligent system that integrates KGs. For more detailed information, refer to the source code.

Applications of Chat-RFB

The system not only helps users efficiently query the KGs using natural language, but also addresses the key challenges of LLMs, such as inaccurate facts and limited domain specificity, by providing feedback on the KGs output data. Especially, after passing the query results of KGs as contextual prompt information to LLMs, the flexibly written knowledge information can be adjusted according to user needs to ensure the high relevance of model knowledge acquisition, which can significantly improve the inference quality and interpretability of the model, simultaneously reducing model illusions, thus providing accurate guidance for research scientists.

Soft–hard zwitterionic trappers (SH-ZITs)[20] are a kind of novel additive material for developing efficient and high-performance battery technology due to their high solubility, stability, and electrochemical performance. Because of effective prompt engineering guide (Fig. S10), the LLM identified the problem content and used a simple Cypher statement to query information directly related to SH-ZITs in KGs. Fig. 3 is a schematic diagram of the relationship nodes centered around SH-ZITs, which expresses the highly summarized content of the

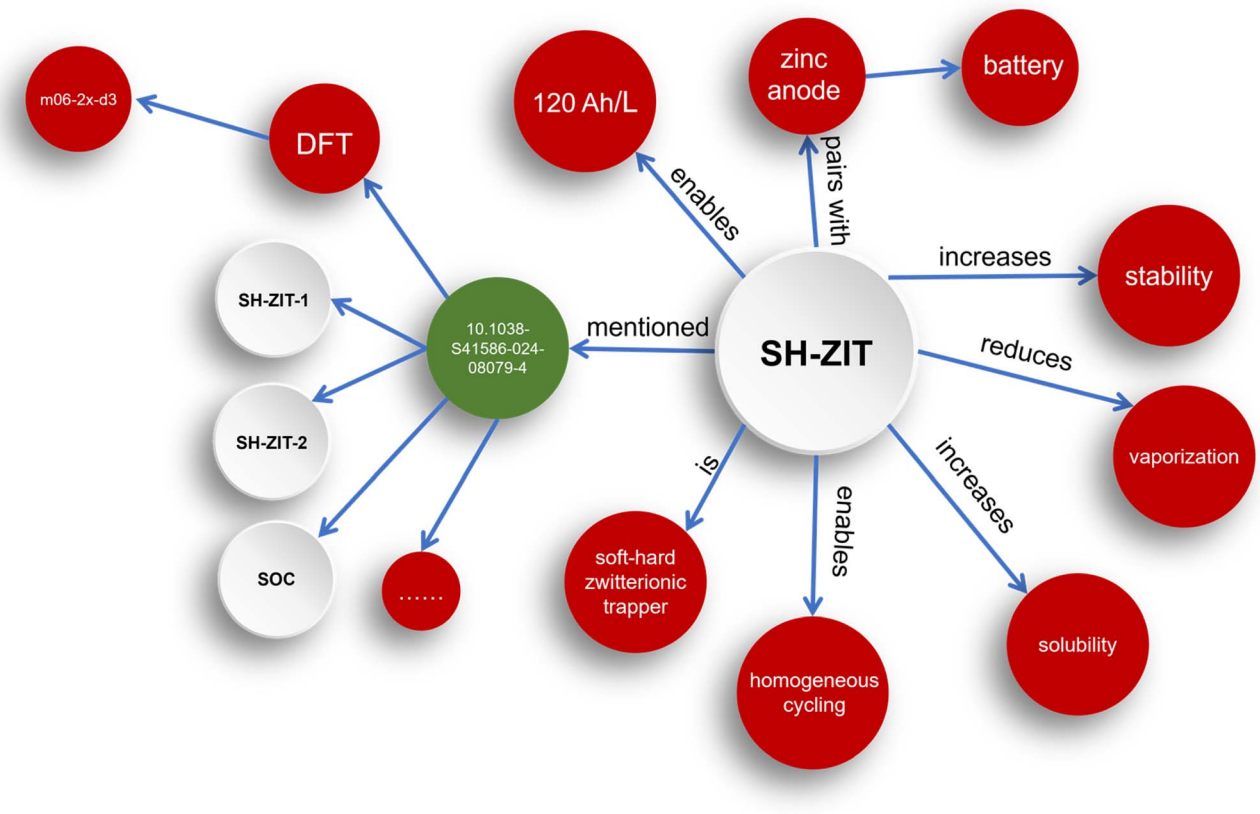

Fig. 3 Schematic diagram of the relationship nodes.

literature. The KGs not only provide researchers with a comprehensive and multidimensional material information database, but also offer strong data support for the research and development of materials science. Researchers can access the chat system through natural language dialogue to obtain cutting-edge information in the field of flow batteries. This structured data representation enables researchers to intuitively grasp the relationship between materials and their potential value in various application scenarios.

The integration of KGs and LLMs to improve their responsiveness has been implemented in various fields.[33,49,50] We used DeepSeek-v3 as the base model and the RAG method to construct KGs based on literature in the field of flow batteries. We generated Chat-RFB, which improved the model's information acquisition ability in the field of flow batteries and enhanced its ability to answer professional questions. As shown in the comparison between the left and right sides of Fig. 4, we compared the DeepSeek-v3 model enhanced by KGs with the native LLM and the commonly used LLMs: GPT-4o, Qwen-Max and Gemini 2.5 Flash. For specialized queries such as "What is SH-ZIT in flow battery?", general-purpose LLMs often produce ambiguous or incorrect responses due to knowledge limitations or hallucinations. In contrast, Chat-RFB, which leverages Cypher queries on structured KGs, retrieves accurate definitions and source references directly from the literature. This structured approach significantly reduces the risk of hallucinations, ensuring that responses are factually grounded and traceable to verified sources.

It is worth noting that compared to the network retrieval enhanced models DeepSeek-v3 (Web Search) and GPT-4o (Web Search), Chat-RFB can demonstrate higher information accuracy. The specific Q&A process is shown in Fig. S13–S19. When asked about the experimental details of SH-ZIT, such as the Gaussian simulation calculation method used in the literature, detailed experimental data cannot be obtained due to the one-sidedness of network search. In contrast, Chat-RFB can obtain real-time parameter data for DFT calculations and provide accurate and effective answers due to the rich content of real-time database queries. This is very meaningful for researchers to conduct experimental comparisons under the same conditions. This example highlights how KG-enhanced LLMs improve scientific accuracy, making them more reliable for domain-specific research applications. This process showcases how the KG's explicit modeling of relationships—from a material ('SH-ZIT') to its source literature ('DOI'), and from that literature to its specific 'calculation methods'—enables the system to answer complex, multi-step queries. Such queries are often intractable for models relying solely on unstructured text

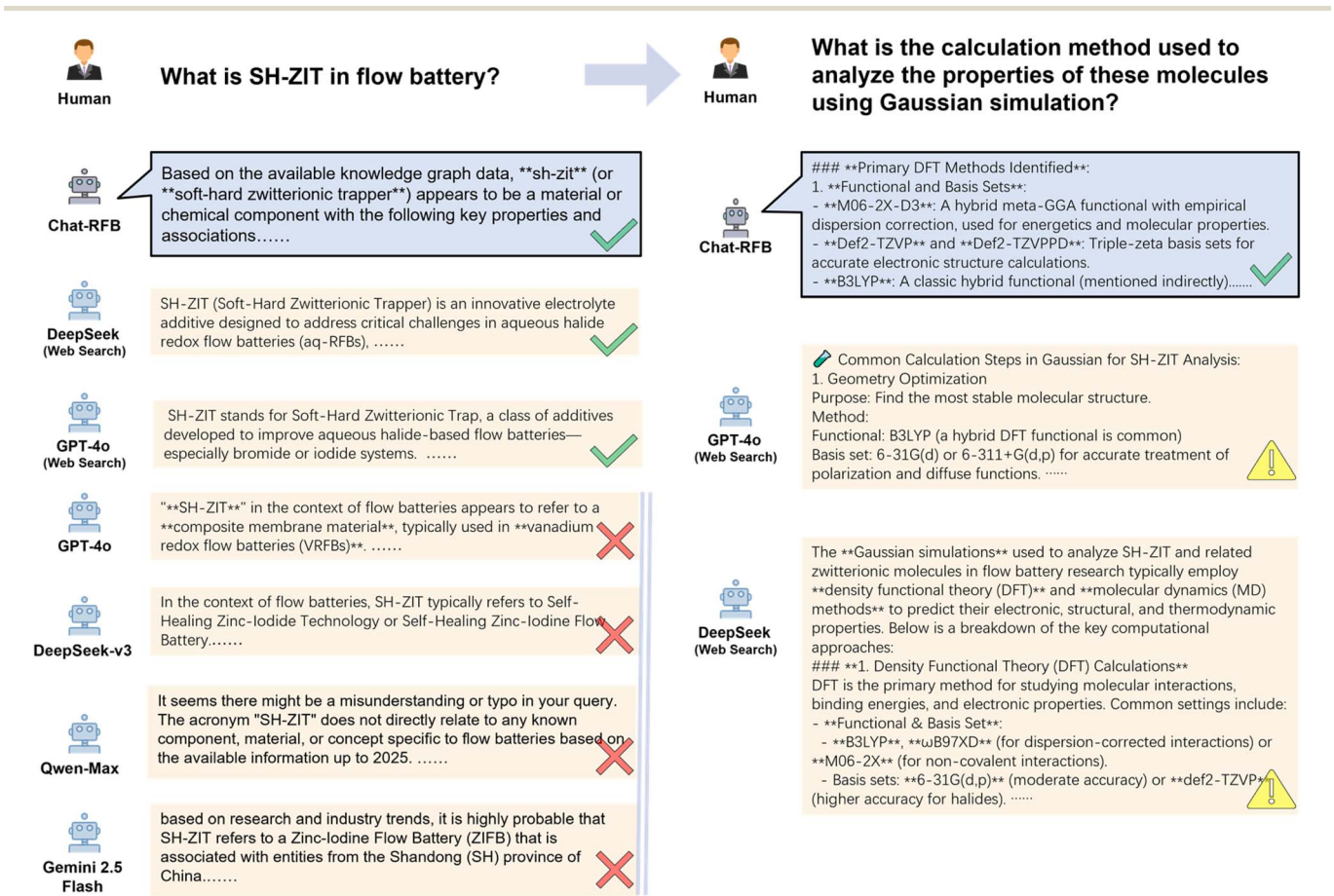

Fig. 4 Comparison of Chat-RFB with multi-turn question-and-answer format of five native LLMs: "What is SH-ZIT in flow battery?" and "What is the calculation method used to analyze the properties of these molecules using Gaussian simulation?".

or less relational data structures. Chat-RFB can now provide accurate information sources, effectively addressing the limitations of LLMs in citation and source tracing. As users inquire about the model, it can further understand the contextual content, conduct deep searches, and provide users with highly relevant literature indexes. Chat-RFB can now provide accurate information sources, effectively addressing the limitations of LLMs in citation and source tracing. The nearly 4% improvement in choice questions suggests that Chat-RFB helped refine factual accuracy, but its impact was limited since many answers were already within the LLMs' pretrained knowledge. In contrast, the 16–22% gain in non-choice questions highlights the key role of KGs in reducing hallucinations and improving response completeness. Unlike choice tasks, non-choice questions require retrieval, synthesis, and articulation of specialized knowledge, whereas purely pretrained models often struggle. By leveraging structured retrieval, Chat-RFB ensures that responses are grounded in verified literature, enhancing accuracy. This indicates that KG-enhanced LLMs excel in complex reasoning tasks, while their impact on fact-based recall is more modest.

Subsequently, we used 450 choice questions to analyze the performance improvement effect of the Chat-RFB model quantitatively. As shown in Fig. 5a, the current general model performs similarly, and with the assistance of KG data, Chat-RFB achieves an accuracy of 94.9% (Table S3). Moreover, LLMs are commonly used in natural language dialogue, so we not only prepared choice questions with fixed answers, but also created a test set consisting of 45 non-choice questions, reviewed by domain experts. As shown in Fig. 5b, the performance comparison results of general LLMs and Chat-RFB optimized using KGs on non-choice questions are presented. Fortunately, with the guidance of KG data, the hallucination of LLMs was effectively suppressed. The model qualification rate reached 93.3%. Beyond overall qualification rates, a closer look at fully correct ("excellent") responses further highlights Chat-RFB's advantage. Among the 45 non-choice questions, it achieved 27 fully correct answers (60.0%, Table S2), significantly outperforming the baseline models, which had less than 30% in this category. This suggests that Chat-RFB not only improves general accuracy but also enhances the precision and completeness of responses. The substantial gap indicates that KG integration helps the model generate more reliable, domain-specific answers, reinforcing its effectiveness in handling complex scientific inquiries.

System modularity and long-term sustainability

Any scientific tool relying on LLMs for rapid iteration must consider its long-term sustainability. We recognize that tying Chat-RFB's future exclusively to any single commercial API is unsustainable. To address this, we adopted a modular architecture to ensure robustness, flexibility, and long-term viability.

Chat-RFB is cleanly decoupled into two core modules: a persistent knowledge base and a replaceable language model interaction layer. This design separation makes swapping underlying LLMs a standardized engineering task rather than a disruptive system overhaul. To address the risk of specific API versions being deprecated or becoming inaccessible in the future, we have planned a clear migration path:

Migrating to other commercial LLM APIs: we handle interactions with LLMs through a dynamic API invocation module. This module essentially acts as a wrapper layer calling different Python libraries (*e.g.*, OpenAI, anthropic, google.generativeai). We designed a Priority and Fallback mechanism: the system first attempts to invoke the preferred model (currently DeepSeek-v3). If the API call fails (*e.g.*, due to network errors, API deprecation, or access permission changes), the module automatically catches the exception and seamlessly switches to fallback models (such as GPT-4, Gemini Pro, *etc.*) in a predefined priority order, ensuring service continuity.

Furthermore, it is worth emphasizing that during the knowledge extraction phase, LLMs are primarily used for one-time, offline batch processing. The outcome—the constructed knowledge graph—is an independent and persistent asset.

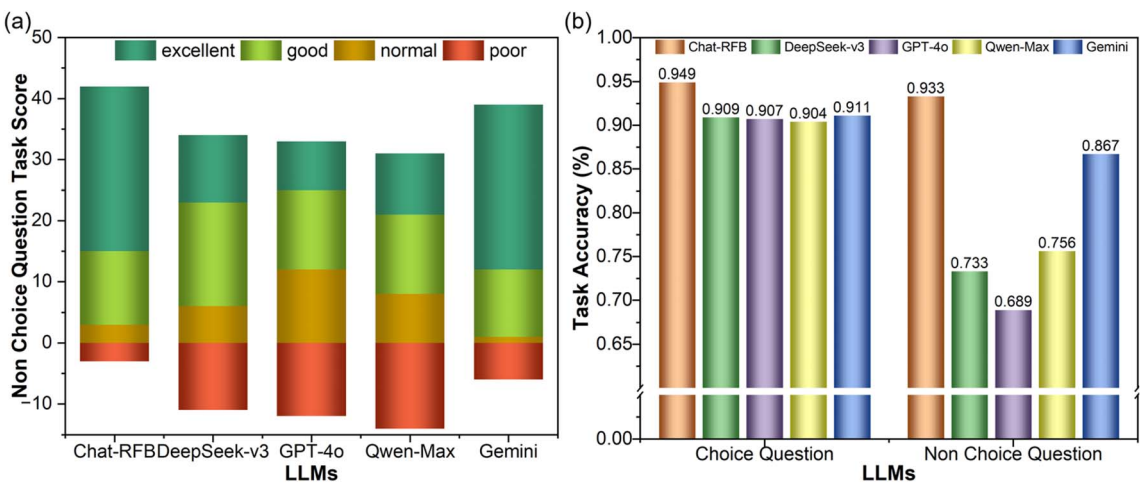

Fig. 5 The answering ability of Chat-RFB compared with that of DeepSeek-v3, GPT-4o, Qwen-Max and Gemini: (a) comparison of the task accuracy of non-choice answer situations and (b) bar chart comparing the accuracy of choice and non-choice questions.

Real-time API dependency is primarily manifested in the question-answering generation phase, which is precisely the core issue addressed by our modular and dynamic invocation mechanism.

However, due to the fact that the construction of the KGs only includes the relationships in the field of flow batteries to effectively obtain key content information of the model, and has not yet introduced knowledge from other fields, the model currently only remains at the level of summarizing known knowledge, lacking innovative thinking on related topics.

Conclusion

In this study, we developed Chat-RFB, the first LLM-KGs-integrated question-answering system for RFBs. By utilizing the natural language processing capabilities of LLMs to analyze and structure knowledge from over 5000 scientific articles, we constructed comprehensive, domain-specific KGs comprising over 164 232 nodes and 853 939 relationships. By leveraging KGs, Chat-RFB demonstrates significantly improved performance, achieving 94.9% accuracy in specialized question-answering tasks while reducing hallucinations compared to general LLMs. To quantitatively validate these results, we also designed a novel evaluation method combining choice and non-choice questions to assess the model's understanding of complex scientific problems. Functionally, Chat-RFB enhances literature retrieval, knowledge structuring, and automated reasoning, proving to be a valuable tool for energy storage research that, in comparison with online-search-augmented LLMs, shows a stronger ability to find details in professional fields. The combination of KGs and LLMs represents a crucial step forward for artificial intelligence in scientific research. The future of Chat-RFB involves expanding the knowledge base, integrating cross-domain expertise, ensuring real-time updates, and applying the system to other energy storage technologies. It is anticipated that such an integrated system will offer automated and intelligent support for scientific research, production, and customized applications, strengthening its role as an AI-driven research assistant and accelerating progress in sustainable energy storage.

Author contributions

Jian-Rong Li and Hao-Tian Wang conceived and designed the study. Hao-Tian Wang wrote the initial manuscript. Ruipeng Jin, Hao-Tian An and Zheng-He Xie participated in discussing the manuscripts. Jian-Rong Li, Xiu-Liang Lv, Xin Zhang and Zhiling Zheng performed the supervision, review and editing.

Conflicts of interest

The authors declare no competing interests.

Data availability

The code supporting the findings of this study is publicly available *via* Zenodo at https://doi.org/10.5281/zenodo.18476424, which archives the corresponding GitHub repository (https://github.com/WHTony1996/KG-RFB).

Supplementary information: supplementary material 1: the LLM information and prompt word information used in this study, as well as examples of KG invocation. Supplementary material 2: the test set table used in this study includes choice and non-choice questions, with each row containing questions, options (choice question), answers, model answers, and evaluation results. See DOI: https://doi.org/10.1039/d5dd00494b.

Acknowledgements

The authors acknowledge the financial support from the Beijing Outstanding Young Scientist Program (Project No. JWZQ20240102008). During the preparation of this work the author(s) used Gemini 2.5 pro in order to assist in the creation of illustrative figures and in correcting grammatical errors in the text. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

1 S. Ha and K. G. Gallagher, Estimating the System Price of Redox Flow Batteries for Grid Storage, *J. Power Sources*, 2015, **296**, 122–132, DOI: 10.1016/j.jpowsour.2015.07.004.
2 Z. Hou, X. Chen, J. Liu, Z. Huang, Y. Chen, M. Zhou, W. Liu and H. Zhou, Towards a High Efficiency and Low-Cost Aqueous Redox Flow Battery: A Short Review, *J. Power Sources*, 2024, **601**, 234242, DOI: 10.1016/j.jpowsour.2024.234242.
3 J. Luo, B. Hu, M. Hu, Y. Zhao and T. L. Liu, Status and Prospects of Organic Redox Flow Batteries toward Sustainable Energy Storage, *ACS Energy Lett.*, 2019, **4**(9), 2220–2240, DOI: 10.1021/acsenergylett.9b01332.
4 L. Zhang, R. Feng, W. Wang and G. Yu, Emerging Chemistries and Molecular Designs for Flow Batteries, *Nat. Rev. Chem.*, 2022, **6**(8), 524–543, DOI: 10.1038/s41570-022-00394-6.
5 P. Xiong, L. Zhang, Y. Chen, S. Peng and G. Yu, A Chemistry and Microstructure Perspective on Ion-Conducting Membranes for Redox Flow Batteries, *Angew. Chem., Int. Ed.*, 2021, **60**(47), 24770–24798, DOI: 10.1002/anie.202105619.
6 L. Zhi, C. Liao, P. Xu, F. Sun, F. Fan, G. Li, Z. Yuan and X. Li, New Alkalescent Electrolyte Chemistry for Zinc-Ferricyanide Flow Battery, *Angew. Chem., Int. Ed.*, 2024, **63**(28), e202403607, DOI: 10.1002/anie.202403607.
7 Z. Zhao, X. Liu, M. Zhang, L. Zhang, C. Zhang, X. Li and G. Yu, Development of Flow Battery Technologies Using the Principles of Sustainable Chemistry, *Chem. Soc. Rev.*, 2023, **52**, 6031–6074, DOI: 10.1039/d2cs00765g.
8 L. Zhang and G. Yu, Recent Developments in Materials and Chemistries for Redox Flow Batteries, *ACS Mater. Lett.*, 2023, **5**(11), 3007–3009, DOI: 10.1021/acsmaterialslett.3c01191.

9 T. Yuan, S. Qi, L. Ye, Y. Zhao, Y. Jiang, Z. Feng, J. Zhu, L. Dai, L. Wang and Z. He, Metal-Organic Frameworks-Based Materials: A Feasible Path for Redox Flow Battery, *Coord. Chem. Rev.*, 2025, **531**, 216503, DOI: 10.1016/j.ccr.2025.216503.

10 H. He, S. Tian, B. Tarroja, O. A. Ogunseitan, S. Samuelsen and J. M. Schoenung, Flow Battery Production: Materials Selection and Environmental Impact, *J. Clean. Prod.*, 2020, **269**, 121740, DOI: 10.1016/j.jclepro.2020.121740.

11 M. Shoaib, P. Vallayil, N. Jaiswal, P. Iyapazham Vaigunda Suba, S. Sankararaman, K. Ramanujam and V. Thangadurai, Advances in Redox Flow Batteries – a Comprehensive Review on Inorganic and Organic Electrolytes and Engineering Perspectives, *Adv. Energy Mater.*, 2024, **14**(32), 2400721, DOI: 10.1002/aenm.202400721.

12 Z. Li, X. Fang, L. Cheng, X. Wei and L. Zhang, Techno-Economic Analysis of Non-Aqueous Hybrid Redox Flow Batteries, *J. Power Sources*, 2022, **536**, 231493, DOI: 10.1016/j.jpowsour.2022.231493.

13 W. Wang, Q. Luo, B. Li, X. Wei, L. Li and Z. Yang, Recent Progress in Redox Flow Battery Research and Development, *Adv. Funct. Mater.*, 2012, **23**(8), 970–986, DOI: 10.1002/adfm.201200694.

14 J. Noack, N. Roznyatovskaya, T. Herr and P. Fischer, The Chemistry of Redox-Flow Batteries, *Angew. Chem., Int. Ed.*, 2015, **54**(34), 9776–9809, DOI: 10.1002/anie.201410823.

15 A. Z. Weber, M. M. Mench, J. P. Meyers, P. N. Ross, J. T. Gostick and Q. Liu, Redox Flow Batteries: A Review, *J. Appl. Electrochem.*, 2011, **41**(10), 1137–1164, DOI: 10.1007/s10800-011-0348-2.

16 M. O. Bamgbopa, Y. Shao-Horn and S. Almheiri, The Potential of Non-Aqueous Redox Flow Batteries as Fast-Charging Capable Energy Storage Solutions: Demonstration with an Iron–Chromium Acetylacetonate Chemistry, *J. Mater. Chem. A*, 2017, **5**(26), 13457–13468, DOI: 10.1039/C7TA02022H.

17 L. Yu and J. Xi, Durable and Efficient PTFE Sandwiched SPEEK Membrane for Vanadium Flow Batteries, *ACS Appl. Mater. Interfaces*, 2016, **8**(36), 23425–23430, DOI: 10.1021/acsami.6b07782.

18 M. Park, J. Ryu, W. Wang and J. Cho, Material Design and Engineering of Next-Generation Flow-Battery Technologies, *Nat. Rev. Mater.*, 2016, **2**, 16080, DOI: 10.1038/natrevmats.2016.80.

19 Z. Xu and M. Wu, Toward Dendrite-Free Deposition in Zinc-Based Flow Batteries: Status and Prospects, *Batteries*, 2022, **8**(9), 117, DOI: 10.3390/batteries8090117.

20 G. Choi, P. Sullivan, X.-L. Lv, W. Li, K. Lee, H. Kong, S. Gessler, J. R. Schmidt and D. Feng, Soft–Hard Zwitterionic Additives for Aqueous Halide Flow Batteries, *Nature*, 2024, **635**(8037), 89–95, DOI: 10.1038/s41586-024-08079-4.

21 D. Peng, Z. Yu, X. Tongle, H. Luling, L. Xiong, L. Minjie, X. Guangrui, L. Wei and M. Weibin, Vertical Model for Polyimide Design Assisted by Knowledge-Fused Large Language Models, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2025-73z9t.

22 H. Yin, A. V. Kononova, T. Bäck and N. V. Stein, Optimizing Photonic Structures with Large Language Model Driven Algorithm Discovery, *arXiv*, 2025, preprint, arXiv:2503.19742, DOI: 10.48550/arXiv.2503.19742.

23 Z. Qin, Q. Dong, X. Zhang, L. Dong, X. Huang, Z. Yang; M. Khademi, D. Zhang, H. H. Awadalla, Y. R. Fung, W. Chen, M. Cheng and F. Wei, Scaling Laws of Synthetic Data for Language Models, *arXiv*, 2025, preprint, arXiv:2503.19551, DOI: 10.48550/arXiv.2503.19551.

24 Y. Tian, W. Li, L. Hu, X. Chen, M. Brook, M. Brubaker, F. Zhang and A. K. Liljedahl, Advancing Large Language Models for Spatiotemporal and Semantic Association Mining of Similar Environmental Events, *arXiv*, 2024, preprint, arXiv:2411.12880, DOI: 10.48550/arXiv.2411.12880.

25 H. Liu, Z. Fu, M. Ding, R. Ning, C. Zhang, X. Liu and Y. Zhang, Logical Reasoning in Large Language Models: A Survey, *arXiv*, 2025, preprint, arXiv:2502.09100, DOI: 10.48550/arXiv.2502.09100.

26 J. Qiu, K. Lam, G. Li, A. Acharya, T. Y. Wong, A. Darzi, W. Yuan and E. J. Topol, LLM-Based Agentic Systems in Medicine and Healthcare, *Nat. Mach. Intell.*, 2024, **6**(12), 1418–1420, DOI: 10.1038/s42256-024-00944-1.

27 Z. Zheng, N. Rampal, T. J. Inizan, C. Borgs, J. T. Chayes and O. M. Yaghi, Large Language Models for Reticular Chemistry, *Nat. Rev. Mater.*, 2025, **10**, 369–381, DOI: 10.1038/s41578-025-00772-8.

28 Z. Zheng, O. Zhang, H. L. Nguyen, N. Rampal, A. H. Alawadhi, Z. Rong, T. Head-Gordon, C. Borgs, J. T. Chayes and O. M. Yaghi, ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs, *ACS Cent. Sci.*, 2023, **9**(11), 2161–2170, DOI: 10.1021/acscentsci.3c01087.

29 X. Bai, Y. Xie, X. Zhang, H. Han and J.-R. Li, Evaluation of Open-Source Large Language Models for Metal–Organic Frameworks Research, *J. Chem. Inf. Model.*, 2024, **64**(13), 4958–4965, DOI: 10.1021/acs.jcim.4c00065.

30 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, Augmenting Large Language Models with Chemistry Tools, *Nat. Mach. Intell.*, 2024, **6**(5), 525–535, DOI: 10.1038/s42256-024-00832-8.

31 P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel and D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, *arXiv*, 2021, preprint, arXiv:2005.11401, DOI: 10.48550/arXiv.2005.11401.

32 Z. Gong, C. Zhang, D. Song, W. Xia, B. Shen, W. Su, H. Duan and A. Su, ChemReactSeek: An Artificial Intelligence-Guided Chemical Reaction Protocol Design Using Retrieval-Augmented Large Language Models, *Chem. Commun.*, 2025, **61**(70), 13137–13140, DOI: 10.1039/D5CC03155A.

33 X. Bai, S. He, Y. Li, Y. Xie, X. Zhang, W. Du and J.-R. Li, Construction of a Knowledge Graph for Framework Material Enabled by Large Language Models and Its Application, *npj Comput. Mater.*, 2025, **11**(1), 51, DOI: 10.1038/s41524-025-01540-6.

34 V. Venugopal and E. Olivetti, MatKG: An Autonomously Generated Knowledge Graph in Material Science, *Sci. Data*, 2024, **11**(1), 217, DOI: 10.1038/s41597-024-03039-z.

35 H. Fan, J. Huang, J. Xu, Y. Zhou, J. Y. H. Fuh, W. F. Lu and B. Li, AutoMEX: Streamlining Material Extrusion with AI Agents Powered by Large Language Models and Knowledge Graphs, *Mater. Des.*, 2025, **251**, 113644, DOI: 10.1016/j.matdes.2025.113644.

36 Q. Ma, Y. Zhou and J. Li, Automated Retrosynthesis Planning of Macromolecules Using Large Language Models and Knowledge Graphs, *Macromol. Rapid Commun.*, 2025, 2500065, DOI: 10.1002/marc.202500065.

37 Y. Fang, Q. Zhang, N. Zhang, Z. Chen, X. Zhuang, X. Shao, X. Fan and H. Chen, Knowledge Graph-Enhanced Molecular Contrastive Learning with Functional Prompt, *Nat. Mach. Intell.*, 2023, **5**(5), 542–553, DOI: 10.1038/s42256-023-00654-0.

38 Y. An, J. Greenberg, F. J. Uribe-Romo, D. A. Gómez-Gualdrón, K. Langlois, J. Furst, A. Kalinowski, X. Zhao, and X. Hu, Knowledge Graph Question Answering for Materials Science (KGQA4MAT), in *Metadata and semantic research*, ed. Garoufallou, E., and Sartori, F., Springer Nature Switzerland, Cham, 2024, pp. 18–29.

39 A. Kondinski, P. Rutkevych, L. Pascazio, D. N. Tran, F. Farazi, S. Ganguly and M. Kraft, Knowledge Graph Representation of Zeolitic Crystalline Materials, *Digital Discovery*, 2024, **3**(10), 2070–2084, DOI: 10.1039/D4DD00166D.

40 S. Ji, S. Pan, E. Cambria, P. Marttinen and P. S. Yu, A Survey on Knowledge Graphs: Representation, Acquisition, and Applications, *IEEE Transact. Neural Netw. Learn. Syst.*, 2022, **33**(2), 494–514, DOI: 10.1109/tnnls.2021.3070843.

41 Y. Gao, L. Wang, X. Chen, Y. Du and B. Wang, Revisiting Electrocatalyst Design by a Knowledge Graph of Cu-Based Catalysts for CO2 Reduction, *ACS Catal.*, 2023, **13**(13), 8525–8534, DOI: 10.1021/acscatal.3c00759.

42 A. Gaudry, M. Pagni, F. Mehl, S. Moretti, L.-M. Quiros-Guerrero, L. Cappelletti, A. Rutz, M. Kaiser, L. Marcourt, E. F. Queiroz, J.-R. Ioset, A. Grondin, B. David, J.-L. Wolfender and P.-M. Allard, A Sample-Centric and Knowledge-Driven Computational Framework for Natural Products Drug Discovery, *ACS Cent. Sci.*, 2024, **10**(3), 494–510, DOI: 10.1021/acscentsci.3c00800.

43 J. Wu, J. Zhu, Y. Qi, J. Chen, M. Xu, F. Menolascina and V. Grau, Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation, *arXiv*, 2024, preprint, arXiv:2408.04187, DOI: 10.48550/arXiv.2408.04187.

44 M. J. Statt, B. A. Rohr, D. Guevarra, J. Breeden, S. K. Suram and J. M. Gregoire, The Materials Experiment Knowledge Graph, *Digital Discovery*, 2023, **2**(4), 909–914, DOI: 10.1039/D3DD00067B.

45 A. Mirza, N. Alampara, S. Kunchapu, M. Ríos-García, B. Emoekabu, A. Krishnan, T. Gupta, M. Schilling-Wilhelmi, M. Okereke, A. Aneesh, A. M. Elahi, M. Asgari, J. Eberhardt, H. M. Elbeheiry, M. V. Gil, M. Greiner, C. T. Holick, C. Glaubitz, T. Hoffmann, A. Ibrahim, L. C. Klepsch, Y. Köster, F. A. Kreth, J. Meyer, S. Miret, J. M. Peschel, M. Ringleb, N. Roesner, J. Schreiber, U. S. Schubert, L. M. Stafast, D. Wonanke, M. Pieler, P. Schwaller and K. M. Jablonka, Are Large Language Models Superhuman Chemists?, *arXiv*, 2024, preprint, arXiv:2404.01475, DOI: 10.48550/arXiv.2404.01475.

46 A. Singh, N. Singh and S. Vatsal, Robustness of LLMs to Perturbations in Text, *arXiv*, 2024, preprint, arXiv:2407.08989, DOI: 10.48550/arXiv.2407.08989.

47 Y. Liu, D. Li, K. Wang, Z. Xiong, F. Shi, J. Wang, B. Li and B. Hang, Are LLMs Good at Structured Outputs? A Benchmark for Evaluating Structured Output Capabilities in LLMs, *Inf. Process. Manag.*, 2024, **61**(5), 103809, DOI: 10.1016/j.ipm.2024.103809.

48 M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez and K. M. Jablonka, From Text to Insight: Large Language Models for Chemical Data Extraction, *Chem. Soc. Rev.*, 2025, **54**(3), 1125–1150, DOI: 10.1039/D4CS00913D.

49 B. Zhou, X. Li, T. Liu, K. Xu, W. Liu and J. Bao, CausalKGPT: Industrial Structure Causal Knowledge-Enhanced Large Language Model for Cause Analysis of Quality Problems in Aerospace Product Manufacturing, *Adv. Eng. Inform.*, 2024, **59**, 102333, DOI: 10.1016/j.aei.2023.102333.

50 Y. Li and B. Starly, Building a Knowledge Graph to Enrich ChatGPT Responses in Manufacturing Service Discovery, *J. Ind. Inf. Integr.*, 2024, **40**, 100612, DOI: 10.1016/j.jii.2024.100612.