

A locally deployed multimodal literature model system for lithium battery research

Yi-Feng Wang¹, Yu Zhang¹, Yang-Zhong Li¹, Qin-Rui Zheng¹, Jian-Ting Liu¹, Rui Lin¹, Qi-Men Xu^{1,*}, Di-Xing Ni^{1,*}

¹High Performance Computing Department, National Supercomputing Center in Shenzhen, Shenzhen 518055, Guangdong, China.

Correspondence to: Dr. Di-Xing Ni, Dr. Qi-Men Xu, High Performance Computing Department, National Supercomputing Center in Shenzhen, Shenzhen 518055, Guangdong, China. E-mail: 12031137@mail.sustech.edu.cn; xuqm@nscsz.cn

Received: 20 September 2026 | Approved: 30 September 2026 | Online: 30 September 2026

Abstract

The analysis of scientific literature is usually constrained by the fact that textual conclusions are separated from the figures and tables which support them. General-purpose cloud models also rely on external computing services, which may increase latency, cost, and data privacy risks. The present study introduces a locally deployed multimodal literature model system and links textual results with their original graphical evidence while keeping the references to the source papers. The system can operate on existing consumer-grade hardware without additional cloud computing quotas or dedicated high-performance infrastructure. When tested using literature on lithium batteries, the system is found to reduce response latency while improving evidence localization, the quality of the answers remaining comparable to that of general-purpose models. The system is therefore offered as a practical solution for verifiable literature analysis in local research settings where data privacy is a concern.



© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Keywords: Materials informatics, lithium batteries, multimodal retrieval, graphical evidence, scientific literature analysis, local deployment

INTRODUCTION

The transition toward low-carbon energy systems has increased the importance of lithium-ion batteries in electric vehicles and stationary energy storage^[1,2]. Research on battery materials and performance has consequently produced a growing body of scientific literature in which experimental conditions, performance data, and the mechanistic conclusions are spread out among the text, the figures, and the tables. This dispersion makes it more difficult to compare studies and to verify the sources.

Materials informatics increasingly uses large language models and automated literature mining methods to organize scientific knowledge^[3,4]. General-purpose language models can produce domain specific errors or unsupported statements^[5], while retrieval augmented generation (RAG) helps ground generation in external evidence^[6]. Materials literature studies have converted unstructured publications into structured materials records^[7-9]. However, these methods still mainly concentrate on textual or structured outputs, with the original figures and tables not always being kept as retrievable evidence. Long document multimodal benchmarks show that scientific questions may depend on information distributed across text and graphical evidence^[10]. Document layout benchmarks indicate that complex scientific pages remain difficult to process consistently^[11,12]. Document understanding has progressed toward models that jointly use text, layout, and visual information^[13-15], and scientific PDF recognition can recover machine readable content from page images^[16]. However, large general-purpose models can require substantial memory and computation during adaptation and inference^[17,18]. These advances therefore do not by themselves provide a resource efficient local literature retrieval workflow that keeps answers linked to their original graphical evidence while running on existing consumer-grade hardware and avoiding recurring cloud computing costs.

To address these limitations, the present study develops a locally deployed multimodal literature model system, with lithium battery research as its initial application domain. The methodological contribution revolves around three closely linked components. First, the system uses complementary table, image object, and caption channels to identify

graphical areas from the heterogeneous PDF structures and at the same time consolidates duplicate candidates while keeping the persistent identifiers. Second, a source-grounded progressive description generation selects the most likely original caption and generates a contextual description only when the source information is insufficient. Third, a fusion retrieval mechanism connects literature passages with the corresponding original figures and tables, allowing generated answers to be verified against graphical evidence and source papers. Local domain adaptation and inference optimization enable the system to function on consumer-grade hardware.

METHODS

This section first presents the system architecture and literature dataset preparation. It then introduces compact text retrieval and its text-only limitation, followed by graphical evidence detection, description generation, and fusion retrieval of text and graphical evidence. Local model adaptation and inference optimization are presented after the retrieval workflow, and the final subsection defines the evaluation protocol.

System architecture

Figure 1 shows a three-layer architecture that connects document processing, model inference, and user interaction within a local environment. Although textual content and graphical evidence are kept in separate indexes, they are still connected to their source papers by means of persistent identifiers. This design enables multimodal question answering as well as the ability to browse figures and tables without having to transfer the documents or the inference tasks to external services.



Figure 1. Architecture of the locally deployed multimodal literature model system. Q&A: question answering; LoRA: low-rank adaptation.

Literature dataset preparation

A domain dataset was assembled from open scientific platforms and institutionally authorized academic sites for model adaptation, retrieval, and knowledge base construction. The papers covered different publication periods and areas of lithium battery research. The PDF files were parsed with pdfplumber and PyMuPDF; headers, footers, and layout noise were removed while text, captions, figures, and tables were retained. The cleaned fields were aligned and stored in JSON for knowledge base construction [Figure 2].

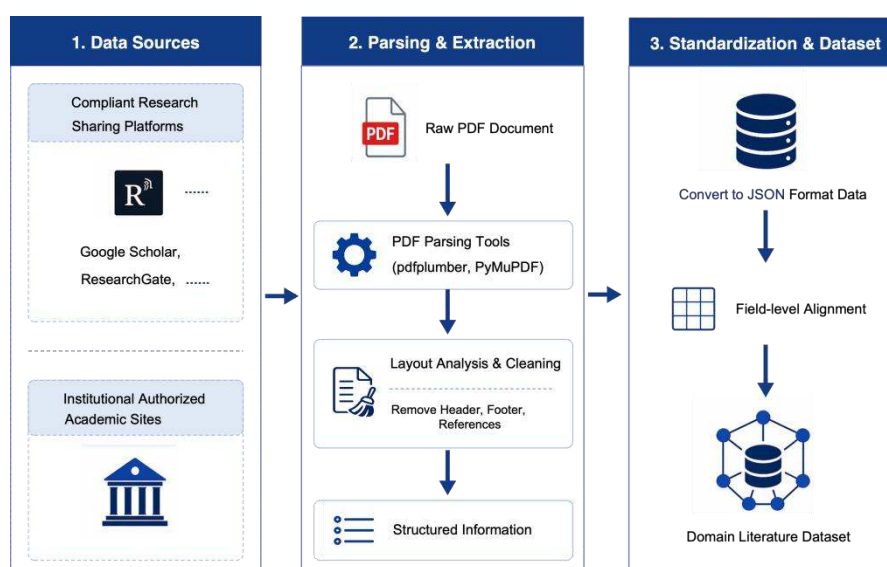


Figure 2. Literature dataset preparation process. PDF: portable document format; JSON: JavaScript Object Notation.

Text and graphical evidence fusion retrieval

Text retrieval

Text retrieval uses a BERT family semantic encoder and dense vector search to match user questions with literature passages^[19]. A compact domain term expansion step is applied before encoding to reduce vocabulary mismatch in materials science expressions. Related studies support prompt based extraction and knowledge grounded agents^[20,21], while dense passage retrieval and sentence embedding enable effective matching between questions and candidate passages^[22,23]. For text retrieval, the current system uses domain term expansion with passage level retrieval and sentence level ranking. This helps find text that closely matches the user's question.

The system first expands the user question with a compact set of domain terms before encoding. Cosine similarity is then used to compare the encoded query with the literature passages, and the highest ranked passages are retained. Sentence ranking is applied within these passages to select the most relevant sentences before answer generation. Figure 3 shows the text retrieval process used in the system.

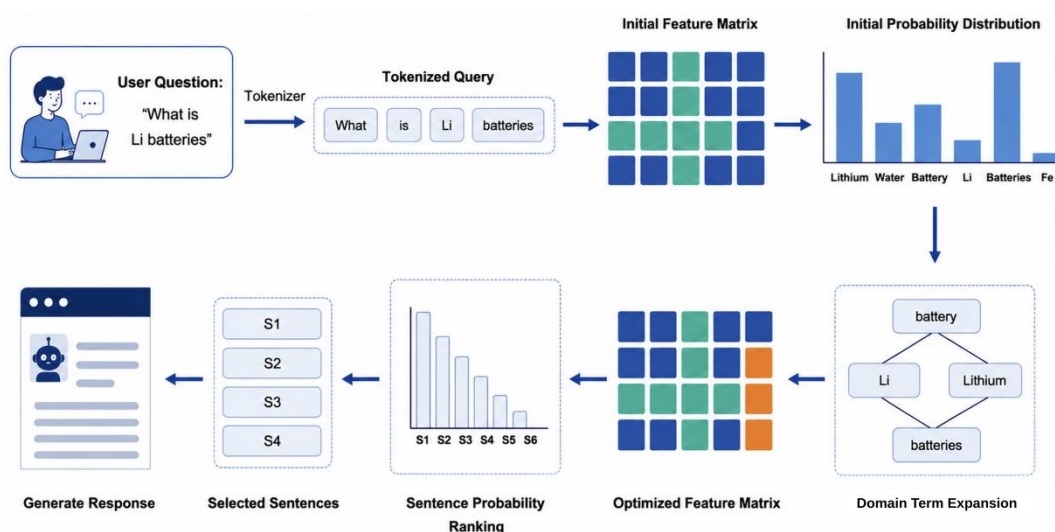


Figure 3. Text retrieval process.

While text retrieval is capable of identifying relevant passages, a great many retrieval systems are still mainly focused on text, which means that the original figures and tables necessary to verify a scientific conclusion are not retrieved along with the answer. An answer may therefore contain the relevant textual conclusion without locating the

original graphical evidence needed for verification. This limitation is particularly important in lithium battery research, where X-ray diffraction patterns, microscopy images, impedance spectra, performance curves, and parameter tables usually contain essential experimental evidence. The proposed method therefore extends text retrieval by detecting and retrieving graphical evidence and links the two evidence types to their sources through persistent associations.

Graphical evidence detection

Figures and tables can appear in the form of table structures, as PDF image objects, or as regions identified by their captions. As there is no one extraction method that can reliably deal with all these different forms, the system makes use of three complementary channels - the table channel, the image object channel, and the caption channel - to identify possible graphical evidence. Graphical evidence may be picked up by more than one of these channels, and the duplicate candidates thus obtained are then consolidated before being stored in the graphical evidence knowledge base.

The candidate regions are consolidated into a unified set and filtered by size, page position, and duplication. For two candidates B_i and B_j , the consolidation decision is based on intersection over union (IoU), as defined in Equation (1).

$$M(B_i, B_j) = \frac{|B_i \cap B_j|}{|B_i \cup B_j|} \quad (1)$$

In Equation (1), $I(\cdot)$ denotes the indicator function. Candidates that satisfy the condition are treated as duplicate detections and consolidated. The IoU threshold was set to 0.7 in the current implementation.

Caption keywords and object features are then used to classify each candidate as a table, scanning electron microscopy image, X-ray diffraction pattern, electrochemical impedance spectrum, performance curve, or another supported category. Before being

The diagram illustrates a document processing pipeline. It starts with a **PDF** document containing various elements. These elements are processed by **Region Detection**, which identifies and categorizes them into three types: **Table**, **Image**, and **Caption**. These detected regions are then passed to a **Consolidate** step. The consolidated data is then processed by a **Filter** step, which shows a 'Before' state with multiple items and an 'After' state with a single item, indicating a filtering process. Finally, the filtered data is stored in a **Store**, which contains a list of items with their **ID** and a **Caption**.

Graphical evidence descriptions

For graphical evidence item g_i , candidate captions C_i are collected from caption-like text blocks on the same page and adjacent pages, and the selected source caption is determined by identifier agreement, geometric proximity, semantic consistency, and caption-format confidence as defined in Equation (2).

$$c_i^* = \arg \max_x \{ \alpha R(c, i) + \beta P(c, i) + \gamma S(c, g_i) + \delta F(c) : c \in C_i \}$$

In Equation (2), R represents agreement with the figure or table identifier, P represents spatial proximity, S measures semantic consistency between the candidate caption and the graphical evidence item together with its referencing context, and F represents conformity with scientific caption patterns. The nonnegative weights α , β , γ , and δ sum to 1. The system accepts the selected source caption only if its score is above the confidence threshold τ_c .

If the selected caption does not meet this threshold, or if no source caption is found, the system moves to context-based generation. The local model combines knowledge learned during domain adaptation with the paragraphs that cite the graphical evidence, neighboring text, and structured metadata to generate a contextual description. The output of each stage is checked for identifier consistency, source-term coverage, and

unsupported content. If the context-based generation fails this check, the available metadata are used to fill a template for the relevant graphical evidence type. Figure 5 shows the progressive description generation process.

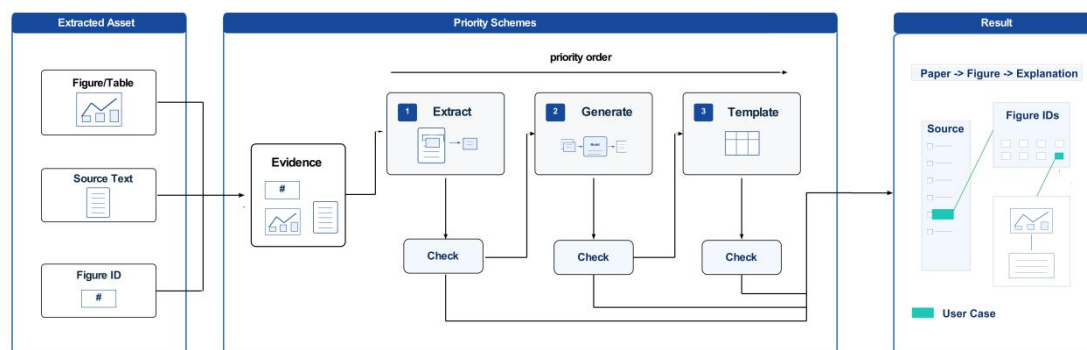


Figure 5. Source-grounded progressive description generation.

Once it has been processed, each knowledge unit includes the source paper, the original graphical evidence, a source-grounded description of it, and a persistent identifier which connects these components. Prioritizing source captions supports factual consistency, while context-based generation and template fallback provide coverage when the original caption is incomplete.

Fusion retrieval

Literature passages and graphical evidence are stored in separate indexes and are linked to their source papers through persistent identifiers. When answering questions, both indexes are searched in parallel; after each has been ranked independently, the graphical evidence items that are associated with the most highly ranked passages are promoted, while strong direct matches are retained.

The final evidence set includes the relevant passages and their corresponding original figures or tables. As shown in Figure 6, the interface presents the answer with its supporting graphical evidence so that users are able to check the result and trace it back to the source paper.

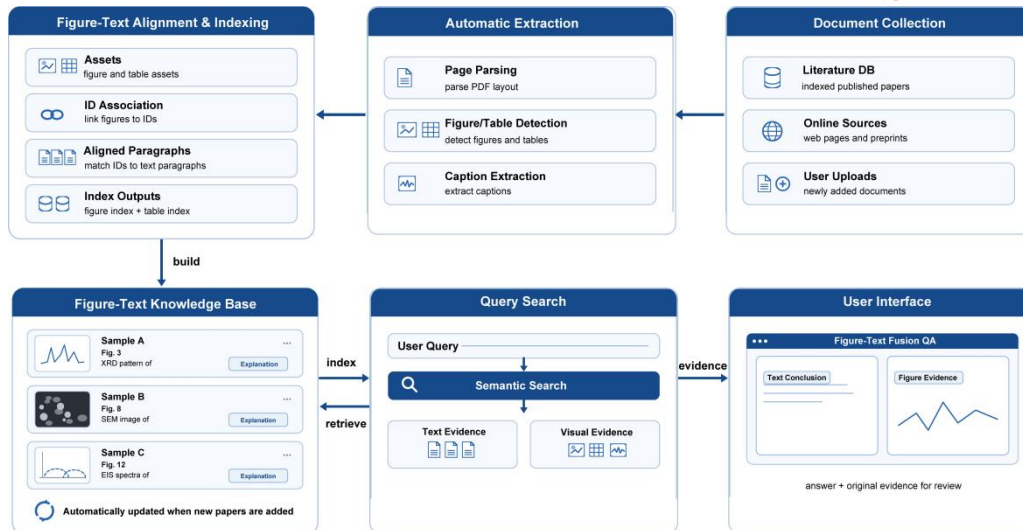


Figure 6. Fusion retrieval of text and graphical evidence.

Local model adaptation and inference optimization

Hybrid LoRA-Adapter fine-tuning

Parameter-efficient tuning is widely used to adapt large pretrained models without updating the full backbone, and quantized tuning can further reduce memory requirements^[18,26]. The present implementation combines two lightweight adaptation paths because local deployment places a stronger constraint on trainable parameters and runtime memory.

For domain adaptation under local resource constraints, the backbone model is frozen while low-rank adaptation (LoRA)^[27] and adapter modules^[28] are trained. For a frozen weight matrix W_0 , the scaled LoRA update is given in Equation (3).

$$W' = I \quad (3)$$

In Equation (3), A and B are trainable low-rank matrices, W_0 remains frozen, r is the low-rank dimension, and α/r is the LoRA scaling factor. The adapter modules provide a separate residual correction pathway, as given in Equation (4).

$$h' = h \quad (4)$$

In Equation (4), h denotes the input hidden representation, the down-projection and up-projection matrices are trainable, and σ represents the nonlinear activation function. This simplified bottleneck relation highlights the residual adapter pathway. Combining the two lightweight transformations limits the number of trainable parameters and allows their weights to be loaded independently during local inference. The evaluation compares full-parameter fine-tuning, LoRA-only fine-tuning, and hybrid LoRA-Adapter fine-tuning.

Semantic chunking of long documents

Long scientific papers often exceed the effective context window of a local language model. Token free byte level models provide one alternative, but generally require a different pretrained architecture rather than a direct modification of an existing token based local model^[29]. Long context processing can still degrade when lengthy sequences are analyzed directly^[30], and a larger context window does not necessarily guarantee reliable access to information across the full input sequence^[31]. At the same time, fixed length segmentation may divide coherent experimental descriptions. These limitations motivate a chunking strategy that considers semantic continuity rather than relying only on the nominal context limit.

The system determines where to place chunk boundaries by considering both semantic continuity and the effective context length; for adjacent text units u_i and u_{i+1} , a new boundary is introduced when their semantic similarity falls below τ_s or when appending u_{i+1} to the current chunk C_j would exceed $L_{\max}$, as defined in Equation (5).

$$b_i = \begin{cases} 1 & \text{if } s(e_i, e_{i+1}) < \tau_s \text{ or } L_{C_j} + |u_{i+1}| > L_{\max} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

In Equation (5), e_i and e_{i+1} are the semantic embeddings of adjacent text units, $s(\cdot)$ is the semantic similarity score, and b_i is the binary boundary variable. A value of $b_i = 1$ starts

a new chunk when either condition is satisfied, whereas $b_i = 0$ retains the adjacent units in the same chunk.

This rule divides long or structurally complex text into coherent chunks, keeping each section within the effective context window while preserving continuous experimental logic and table fields. The resulting chunks are encoded independently and can be processed in parallel to reduce redundant computation over long contexts. Figure 7 shows the semantic chunking process.

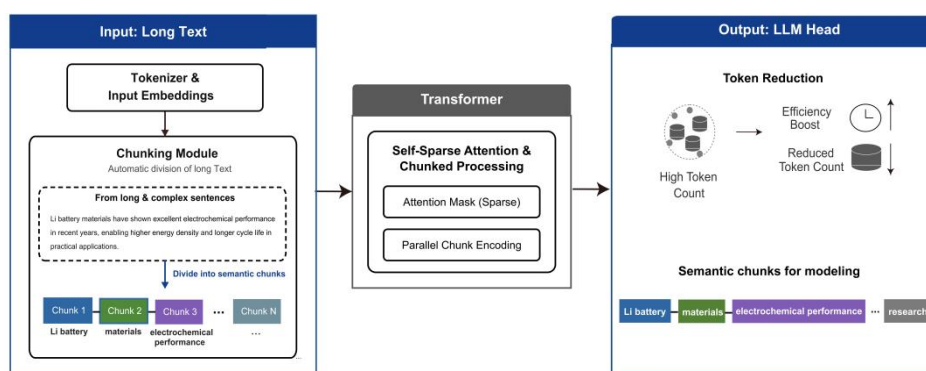


Figure 7. Semantic chunking of long scientific papers. LLM: large language model.

Parallel resource scheduling

Local inference can be delayed when central processing unit (CPU) preparation and graphics processing unit (GPU) computation are executed serially. The CPU performs data preparation, context construction, prompt organization, cache management, and adapter preparation, while the GPU performs model loading, prefill, and decoding. Memory management and adapter switching can further increase waiting time in large language model serving^[32,33].

During inference round r , the GPU executes the current round while the CPU prepares the context, prompt, and adapter for round $r + 1$. Over R successive rounds, the total pipeline time is given by Equation (6).

$$T_{\text{pipeline}}^{(6)}$$

In Equation (6), each overlapped stage is determined by the slower of the CPU preparation and the GPU inference, while the initial CPU preparation and the final GPU inference are considered as boundary terms. Once the first input has been prepared, the pipeline reaches a steady state at which the prepared inputs are passed on to the GPU as soon as the current inference round is completed.

The CPU then starts preparing the next input, allowing successive rounds to proceed without an additional serial waiting period. This staggered transfer across the various rounds is shown in Figure 8. The latency reduction therefore arises from overlapping preparation and inference rather than omitting either stage of the original computation.

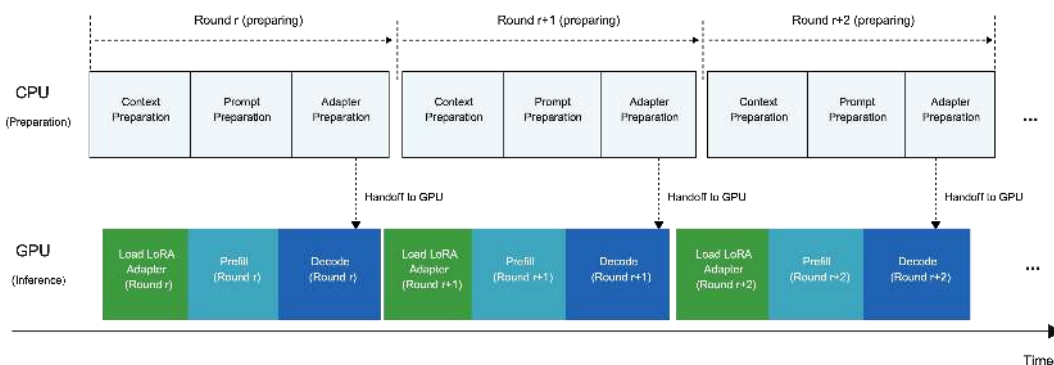


Figure 8. Parallel resource scheduling across successive inference rounds. CPU: central processing unit; GPU: graphics processing unit; LoRA: low-rank adaptation.

Evaluation protocol

Fine-tuning performance was assessed under the same training conditions, and the three fine-tuning strategies were compared in terms of training cost, stability, and task performance.

At the system level, the proposed system, GPT-5.6, and Gemini-3.5 were tested on the same lithium battery scientific question answering set. Response latency was summarized using the mean, median (P50), and 95th percentile (P95).

Evidence localization was assessed by comparing the set of the top three returned evidence items $R_i^{(3)}$ with the gold-standard evidence set G_i , as defined in Equation (7).

$$LocAcc = \frac{1}{N} \quad (7)$$

In Equation (7), $LocAcc$ refers to the fraction of the questions for which at least one gold-standard item appears in the top three results, and $LocCount$ is the average number of gold-standard items retrieved within the same range.

Answer quality was summarized by the study-specific answer quality index (AQI), as defined in Equation (8).

$$AQI = 0.3 S_{length} + 0.4 S_{keyword} + 0.3 S_{feasibility} \quad (8)$$

In Equation (8), S_{length} , $S_{keyword}$, and $S_{feasibility}$ represent the response-length, keyword-coverage, and feasibility component scores, respectively. Each of the components is scaled to a range from 0 to 100, and the weights of 30%, 40%, and 30% ensure that the content coverage receives the greatest contribution while still including the two complementary components.

RESULTS AND DISCUSSION

Fine-tuning performance

Figure 9 compares the resource use and task performance of the three fine-tuning strategies. Full-parameter fine-tuning achieved strong task performance but required the greatest training time and resources, whereas LoRA-only fine-tuning minimized training cost at lower task performance. Hybrid LoRA-Adapter added limited overhead over LoRA-only while approaching full-parameter performance, making it more suitable for repeated local adaptation.

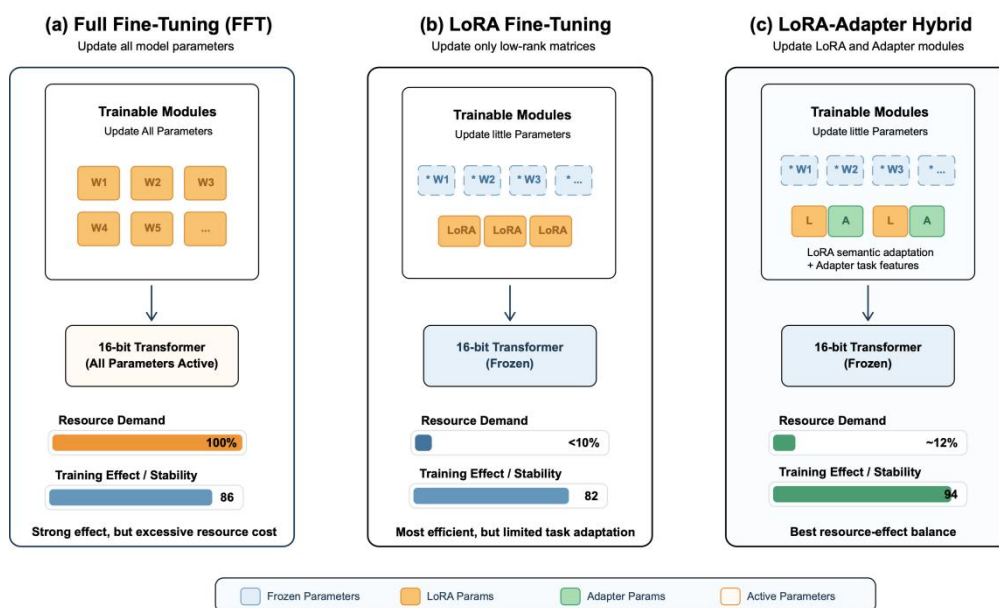


Figure 9. Resource use and task performance of the three fine-tuning strategies.

The training curves in Figure 10 show that the hybrid model remained above the base and LoRA-only models for most of training and achieved the highest final performance. This indicates that the adapter pathway improved domain adaptation.

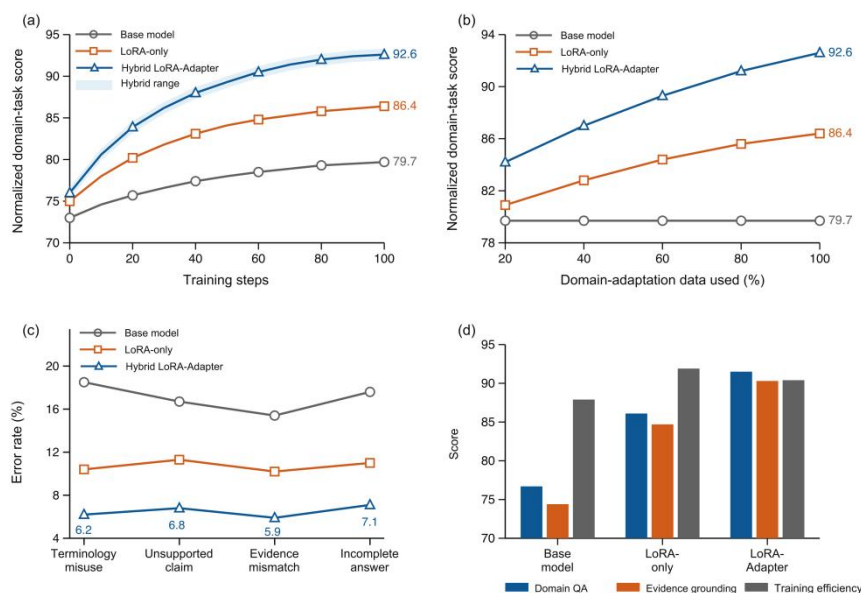


Figure 10. Training performance of the base, LoRA-only, hybrid LoRA Adapter models.

Comparison with general-purpose language models

Same scientific dataset was used to assess the proposed system, GPT-5.6, and Gemini-3.5. According to the metrics set out in the Evaluation protocol, Table 1 shows a comparison of their response latency, evidence localization, and answer quality.

Table 1. Performance comparison on lithium battery scientific question answering.

Metric	Proposed system	GPT-5.6	Gemini-3.5
Response latency (s)			
Mean	2.9	6.2	3.8
Median (P50)	2.6	5.6	3.4
95th percentile (P95)	5.0	12.0	7.5
Evidence localization			
Accuracy (LocAcc)	0.910	0.850	0.880
Count in top three (LocCount)	2.300	2.050	2.200
Answer quality			
AQI (0-100)	88	90	87

Mean, P50, and P95 refer to the mean, median, and 95th percentile response latency respectively. LocAcc refers to evidence localization accuracy, and LocCount is the average number of gold-standard items among those in the top three results. AQI stands for the answer quality index.

The proposed system was faster than both of the reference models and obtained the highest scores for LocAcc and LocCount. Its AQI stayed near that of GPT-5.6 and Gemini-3.5. These findings show that the reduction in response latency and improvements in evidence localization did not result in a substantial decrease in answer quality.

Fusion retrieval ablation

In order to find out whether combining literature passages with graphical evidence gives an advantage over using either type of evidence alone, a representative question-level ablation study was carried out with the generation settings left unchanged. The text-only retrieval, the graphical-only retrieval, and the fusion retrieval of both text and graphical

evidence were compared using the same evaluation criteria. The resulting AQI values are summarized in Table 2.

Table 2. Representative question-level ablation of text-only, graphical-only, and fusion retrieval.

Question Sample	Text-only AQI	Graphical-only AQI	Fusion AQI
Q01	84	80	90
Q02	82	—	89
Q03	—	83	91

AQI is reported on a 0-100 scale; — indicates that no corresponding evidence was located.

Local inference efficiency and operational stability

Figure 11 evaluates the final local model from two perspectives: latency decomposition identifies the main contributors to response latency, while batch level evaluation examines the stability of successful and evidence-grounded responses.

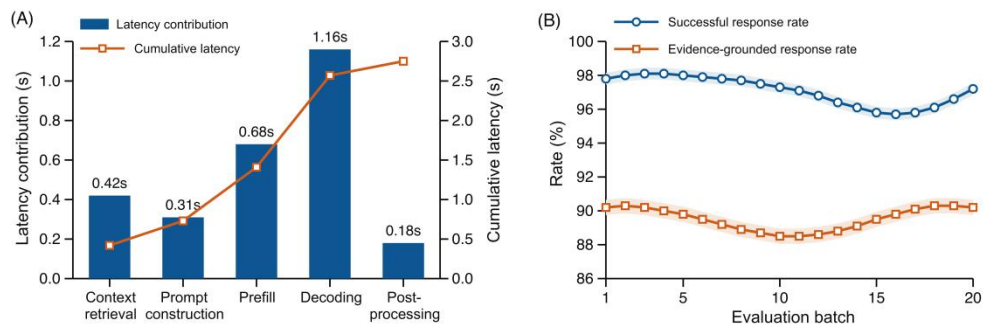


Figure 11. Local inference efficiency and stability of the hybrid LoRA-Adapter model.

Figure 11A shows that all stages were completed within a short time, with the displayed stages totaling 2.75 s, close to the 2.9 s mean response latency in Table 1. In Figure 11B, the successful response rate remained at 96% to 98%, showing that the system returned an answer for most questions, while the evidence-grounded response rate, which additionally requires the answer to be associated with supporting evidence, remained at 88% to 90% across 20 evaluation batches. The consistently high rates and limited variation indicate stable answer generation and evidence association during repeated use.

Case study and limitations

To examine the system in an integrated application scenario, a case study was conducted using 96 lithium battery cathode papers collected from open scientific platforms and institutionally authorized academic sites. The papers were processed through the proposed graphical evidence detection and consolidation workflow, resulting in 356 indexed graphical evidence items, including figures and tables, which were stored in the local knowledge base along with the relevant source papers and the corresponding literature passages. The case study aimed to determine whether the full workflow could enable multimodal question answering and browsing of the graphical evidence on a paper-by-paper basis when deployed locally. Figure 12 shows the multimodal question answering and graphical evidence catalog interfaces.

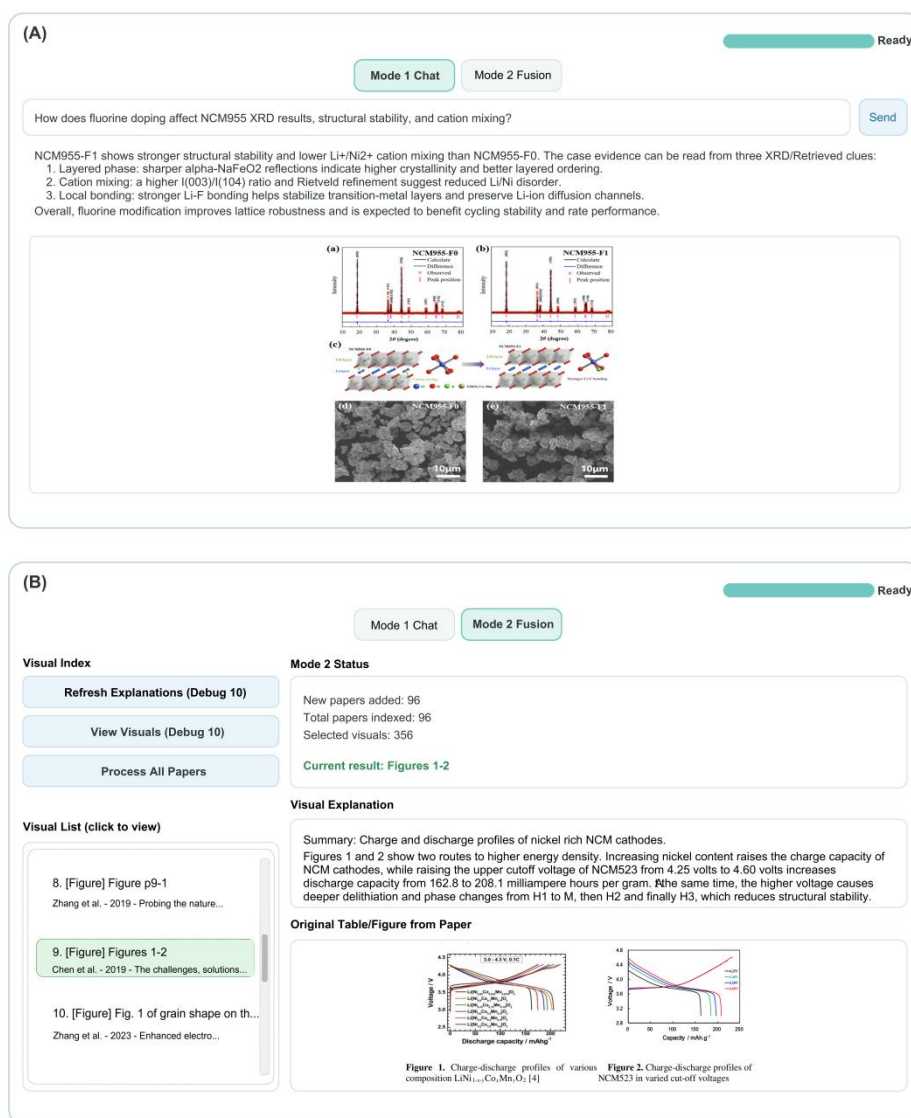


Figure 12. Interfaces of the locally deployed multimodal literature model system.

The question answering interface allows textual answers to be checked against the corresponding original graphical evidence and traced back to the source paper. The graphical evidence catalog interface supports paper-by-paper browsing of indexed graphical evidence items and comparison of their source-grounded descriptions with the corresponding figures or tables. Incremental caching updates the index as new papers are added, allowing the local knowledge base to expand without reprocessing existing content.

The current evaluation is limited to lithium battery literature and focuses primarily on evidence localization rather than the automatic validation of scientific hypotheses. Further evaluation across additional materials domains and more diverse scientific tasks is therefore needed to assess the generalizability of the system.

CONCLUSIONS

A locally deployed multimodal literature model system was developed for use in the study of lithium batteries. The system regards the original figures and tables as retrievable scientific evidence by means of multi-channel detection and consolidation, using source-grounded progressive description generation and fusion retrieval with literature passages. The system's local operation on consumer-grade hardware is supported by hybrid LoRA-Adapter fine-tuning, semantic chunking, and parallel resource scheduling.

On the common question answering test set, the system had a mean response latency of 2.9 seconds and an evidence localization accuracy of 0.910, while maintaining an AQI of 88. The case study also showed that multimodal question answering and browsing of a graphical evidence catalog were demonstrated within a local environment. Further research should focus on evaluating additional materials domains, improving the efficiency and coverage of fusion retrieval, and extending the system to support a broader range of modalities.

Looking ahead, domain specific model systems may increasingly complement general purpose models by adapting AI technologies to concrete scientific and professional workflows. The development of such systems may not only extend the technical boundaries of practical AI applications but also lower barriers to their use, enabling AI

to simplify work and improve efficiency while remaining focused on supporting human activities rather than displacing existing opportunities.

DECLARATIONS

Authors' contributions

Model development, performance optimization and writing - original draft: Wang Y;

Conceptualization, methodology and project supervision: Xu Q, Ni D;

Data cleaning and analysis: Zhang Y, Li Y;

Local deployment and validation: Zheng Q, Liu J;

Platform visualization: Lin R;

Writing - review and editing: Wang Y, Xu Q, Ni D.

Availability of data and materials

The program and source code developed in this study are available at:
<https://github.com/yifengwang123456/-Local-Multimodal-Literature-Model-System>.

AI and AI-assisted tools statement

Not applicable.

Financial support and sponsorship

This work was supported by the Guangdong Science and Technology Program under Grant No. 2024B0101040005, the National Natural Science Foundation of China under Grant No. 12404264, and the Greater Bay Area Institute of HPC-AI Co-Driven Innovation.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. International Energy Agency. Batteries and Secure Energy Transitions; *International Energy Agency: Paris*, 2024.
2. Liu, J.; Xiao, J.; Yang, J.; et al. The TWh challenge: next generation batteries for energy storage and electric vehicles. *Next Energy* 2023, 1, 100015.[DOI: 10.1016/j.nxener.2023.100015]
3. Jablonka, K. M.; Ai, Q.; Al-Feghali, A.; et al. 14 examples of how large language models can transform materials science and chemistry: a reflection on a large language model hackathon. *Digit. Discov.* 2023, 2, 1233-50.[PMID: 38013906 PMCID: PMC10561547 DOI: 10.1039/d3dd00113j]
4. Lei, G.; Docherty, R.; Cooper, S. J. Materials science in the era of large language models: a perspective. *Digit. Discov.* 2024, 3, 1257-72.[PMID: 42751567 PMCID: PMC13580199 DOI: 10.1039/d4dd00074a]
5. Vangala, B. P.; Mahmud, S.; Neupane, P.; et al. HalluMat: detecting hallucinations in LLM-generated materials science content through multi-stage verification. *arXiv* 2025, arXiv:2512.22396.[DOI: 10.48550/arXiv.2512.22396]
6. Lewis, P.; Perez, E.; Piktus, A.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *In Advances in Neural Information Processing Systems*; 2020; Vol. 33, pp 9459-74.[DOI: 10.48550/arXiv.2005.11401]
7. Tshitoyan, V.; Dagdelen, J.; Weston, L.; et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 2019, 571, 95-8.[DOI: 10.1038/s41586-019-1335-8]
8. Kononova, O.; Huo, H.; He, T.; et al. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* 2019, 6, 203.[DOI: 10.1038/s41597-019-0224-1]
9. Olivetti, E. A.; Cole, J. M.; Kim, E.; et al. Data-driven materials research enabled by natural language processing and information extraction. *Appl. Phys. Rev.* 2020, 7, 041317.[DOI: 10.1063/5.0021106]
10. Ma, Y.; Zang, Y.; Chen, L.; et al. MMLongBench-Doc: benchmarking long-context document understanding with visualizations. *Adv. Neural Inf. Process. Syst.* 2024, 37.[DOI: 10.52202/079017-3041]

11. Ouyang, L.; Qu, Y.; Zhou, H.; et al. OmniDocBench: benchmarking diverse PDF document parsing with comprehensive annotations. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2025; pp 24838-48.[DOI: 10.1109/CVPR52734.2025.02313]
12. Pfitzmann, B.; Auer, C.; Dolfi, M.; Nassar, A. S.; Staar, P. DocLayNet: a large human-annotated dataset for document-layout segmentation. *In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; 2022; pp 3743-51.[DOI: 10.48550/arXiv.2206.01062]
13. Zhong, X.; Tang, J.; Jimeno-Yepes, A. PubLayNet: largest dataset ever for document layout analysis. *In Proceedings of the 15th International Conference on Document Analysis and Recognition*; 2019; pp 1015-22.[DOI: 10.1109/ICDAR.2019.00166]
14. Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; Zhou, M. LayoutLM: pre-training of text and layout for document image understanding. *In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2020; pp 1192-200.[DOI: 10.1145/3394486.3403172]
15. Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; Wei, F. LayoutLMv3: pre-training for document AI with unified text and image masking. *In Proceedings of the 30th ACM International Conference on Multimedia*; 2022; pp 4083-91.[DOI: 10.1145/3503161.3548112]
16. Blecher, L.; Cucurull, G.; Scialom, T.; Stojnic, R. Nougat: neural optical understanding for academic documents. *In International Conference on Learning Representations*; 2024.[DOI: 10.48550/arXiv.2308.13418]
17. Sheng, Y.; Zheng, L.; Yuan, B.; et al. FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU. *In Proceedings of the 40th International Conference on Machine Learning; PMLR 202*; 2023; pp 31094-116.[DOI: 10.48550/arXiv.2303.06865]
18. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: efficient finetuning of quantized LLMs. *Adv. Neural Inf. Process. Syst.* 2023, 36.[DOI: 10.52202/075280-0441]
19. Chen, J.; Xiao, S.; Zhang, P.; Luo, K.; Lian, D.; Liu, Z. M3-Embedding: multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *In Findings of the Association for Computational Linguistics: ACL 2024*; 2024; pp 2318-35.[DOI: 10.18653/v1/2024.findings-acl.137]

20. Polak, M. P.; Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* 2024, 15, 1569.[DOI: 10.1038/s41467-024-45914-8]
21. Zhang, H.; Song, Y.; Hou, Z.; Miret, S.; Liu, B. HoneyComb: a flexible LLM-based agent system for materials science. *In Findings of the Association for Computational Linguistics: EMNLP 2024*; 2024; pp 3369-82.[DOI: 10.18653/v1/2024.findings-emnlp.192]
22. Reimers, N.; Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT networks. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*; 2019; pp 3982-92.[DOI: 10.18653/v1/D19-1410]
23. Karpukhin, V.; Oguz, B.; Min, S.; et al. Dense passage retrieval for open-domain question answering. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*; 2020; pp 6769-81.[DOI: 10.18653/v1/2020.emnlp-main.550]
24. Hsu, T.-Y.; Hsu, Y.-L.; Rohatgi, S.; et al. Do large multimodal models solve caption generation for scientific figures? Lessons learned from SciCap Challenge 2023. *Trans. Assoc. Comput. Linguist.* 2026, 14, 233-52.[DOI: 10.1162/tacl.a.653]
25. Kim, J.; Lee, J.; Choi, H.-J.; et al. Multi-LLM collaborative caption generation in scientific documents. *In AI for Research and Scalable, Efficient Systems; Communications in Computer and Information Science*, Vol. 2533; Springer: Singapore, 2025; pp 142-60.[DOI: 10.1007/978-981-96-8912-5_6]
26. Ding, N.; Qin, Y.; Yang, G.; et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mach. Intell.* 2023, 5, 220-35.[DOI: 10.1038/s42256-023-00626-4]
27. Hu, E. J.; Shen, Y.; Wallis, P.; et al. LoRA: low-rank adaptation of large language models. *In International Conference on Learning Representations*; 2022.[DOI: 10.48550/arXiv.2106.09685]
28. Houshy, N.; Giurghi, A.; Jastrzebski, S.; et al. Parameter-efficient transfer learning for NLP. *In Proceedings of the 36th International Conference on Machine Learning; PMLR* 97; 2019; pp 2790-9.[DOI: 10.48550/arXiv.1902.00751]
29. Xue, L.; Barua, A.; Constant, N.; et al. ByT5: towards a token-free future with pre-trained byte-to-byte models. *Trans. Assoc. Comput. Linguist.* 2022, 10, 291-306.[DOI: 10.1162/tacl_a_00461]

30. Hosseini, P.; Castro, I.; Ghinassi, I.; Purver, M. Efficient solutions for an intriguing failure of LLMs: long context window does not mean LLMs can analyze long sequences flawlessly. *In Proceedings of the 31st International Conference on Computational Linguistics*; 2025; pp 1880-91.[DOI: 10.48550/arXiv.2408.01866]
31. Liu, N. F.; Lin, K.; Hewitt, J.; et al. Lost in the middle: how language models use long contexts. *Trans. Assoc. Comput. Linguist.* 2024, 12, 157-73.[DOI: 10.1162/tac1_a_00638]
32. Kwon, W.; Li, Z.; Zhuang, S.; et al. Efficient memory management for large language model serving with PagedAttention. *In Proceedings of the 29th Symposium on Operating Systems Principles*; 2023; pp 611-26.[DOI: 10.1145/3600006.3613165]
33. Sheng, Y.; Cao, S.; Li, D.; et al. S-LoRA: serving thousands of concurrent LoRA adapters. *Proc. Mach. Learn. Syst.* 2024, 6.[DOI: 10.48550/arXiv.2311.03285]