Perspective | Open Access

AI Agent

Chen *et al. AI Agent* **2026**, *2*, 12 **DOI:**10.20517/aiagent.2026.06

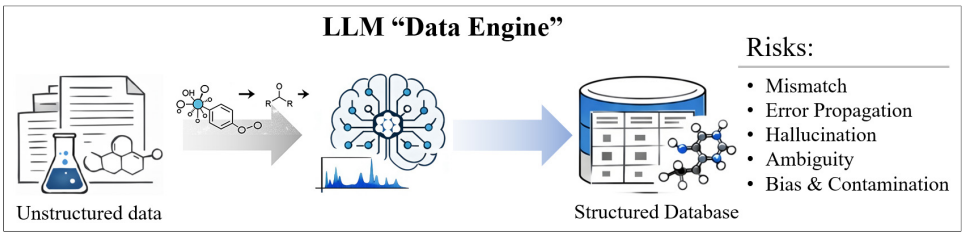
From unstructured to structured: opportunities, risks and engineering practices for LLM-based data extraction in chemistry and materials

Weijian Chen[1], Haoxiang Xu[1,2], Daojian Cheng[1,2]

Citation: Chen, W.; Xu, H.; Cheng, D. From unstructured to structured: opportunities, risks and engineering practices for LLM-based data extraction in chemistry and materials. *AI Agent* **2026**, *2*, 12. https://dx.doi.org/10.20517/aiagent.2026.06

Received: 2 Mar 2026
First Decision: 10 Mar 2026
Revised: 24 Mar 2026
Accepted: 22 May 2026
Published: 3 Jun 2026

Academic Editor: Hao Li
Copy Editor: Xing-Yue Zhang
Production Editor: Xing-Yue Zhang



The vast majority of experimental knowledge in chemistry and materials science remains locked in unstructured natural language texts, creating a fundamental bottleneck for data-driven discovery. Large language models (LLMs), with their remarkable in-context learning and instruction-following capabilities, offer a transformative solution by enabling rapid, scalable conversion of scientific literature into structured databases. This perspective critically examines the end-to-end workflow of LLM-based data extraction, from preprocessing and LLM interaction to postprocessing and validation, while highlighting the unique advantages of chemical domain knowledge for constraining and verifying outputs. We systematically identify hidden risks in large-scale extraction, including condition-value mismatches, error accumulation, "ingratiation" bias and instruction ambiguity, drawing on practical observations. To move toward reliable and robust extraction, we propose four engineering principles, including task decomposition, multilayer validation, clear division of labor between LLMs and deterministic tools, and standardized data reporting in publications. Finally, we outline future research frontiers, namely multimodal agents, crossdocument integration and lowresource benchmarks, that will shape the next generation of autonomous scientific data assistants.

INTRODUCTION

The paradigm of chemical materials discovery is shifting toward datadriven approaches, where highthroughput screening, machine learning and artificial intelligence rely on large volumes of highquality structured data[1-3]. Yet the vast

[1]Beijing Key Laboratory of Intelligent Design and Manufacturing for Hydrogen Energy Materials, Beijing University of Chemical Technology, Beijing 100029, China.
[2]Deep Intelligence Experiment Technology (Beijing) Co., Ltd., Beijing 100029, China.

Correspondence to: Prof. Haoxiang Xu, Prof. Daojian Cheng, Beijing Key Laboratory of Intelligent Design and Manufacturing for Hydrogen Energy Materials, Beijing University of Chemical Technology, Beijing 100029, China. E-mail: xuhx@mail.buct.edu.cn; chengdj@mail.buct.edu.cn

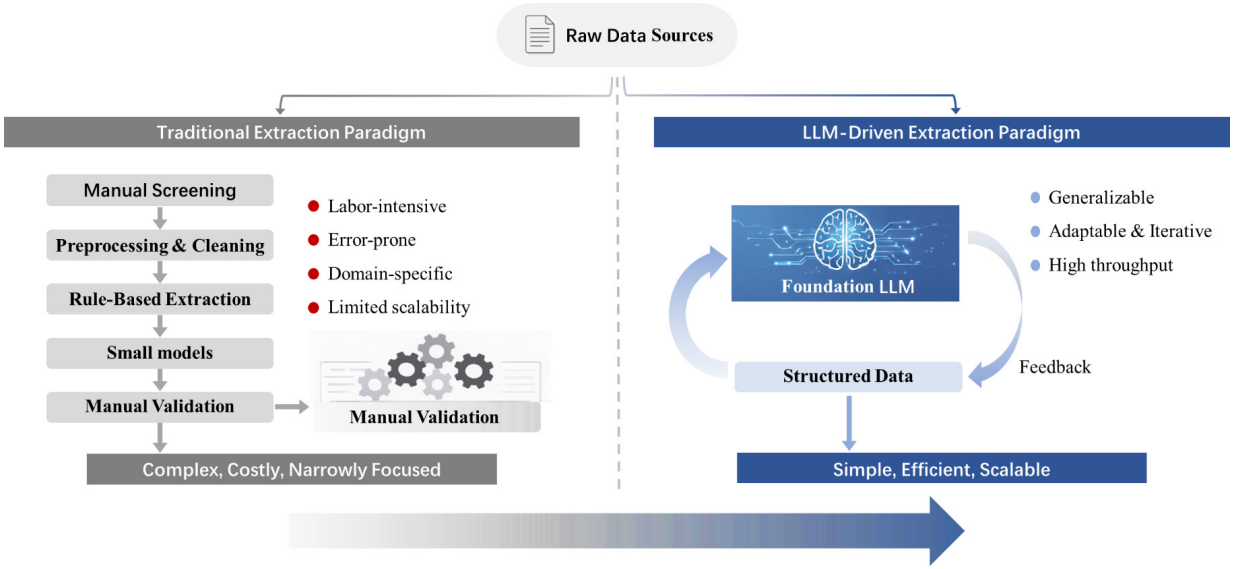


Figure 1. Comparison of data extraction paradigms.

majority of accumulated experimental knowledge, such as synthesis conditions, characterization results, property measurements, resides in unstructured natural language texts, including millions of research articles, patents and supporting information files[4, 5]. This "textual data warehouse" stands in stark contrast to the needs of datadriven chemical materials discovery workflows. Manual curation is prohibitively slow, and traditional natural language processing tools, often rulebased or trained on narrow corpora, suffer from poor generalization and require weeks of reengineering for each new task. In contrast to traditional methods that rely on predefined rules or task-specific training data, LLM-based approaches offer greater flexibility and generalization through in-context learning. However, they also introduce new challenges related to reliability, interpretability, and error control [Figure 1].

Large language models (LLMs) have emerged as a gamechanger. Their ability to perform incontext learning from a few examples[6,7] and to follow complex instructions enables researchers to prototype data extraction pipelines in days rather than months[8,9]. By treating LLMs as versatile "data engines," we can now convert unstructured text into structured formats, such as tables, databases, knowledge graphs and so on, at an unprecedented scale[10-13]. This transformation promises to accelerate the construction of comprehensive experimental databases that fuel downstream applications, from catalyst screening to property prediction.

Several recent reviews have surveyed the emerging role of LLMs and AI-driven frameworks in chemical and materials science. For instance, Zhang *et al.*[14] summarize the use of LLMs and related AI tools for knowledge extraction and data-driven research, while Zhang *et al.*[15] present the concept of a "digital materials ecosystem," highlighting integrated workflows that combine databases, AI agents, and automated experimentation. These works collectively emphasize the growing importance of structured data extraction and AI-enabled discovery pipelines.

Despite these advances, these studies primarily focus on model capabilities, system architectures, or high-level workflows, often under idealized settings, with comparatively less attention paid to the practical challenges encountered in real-world deployment. Consequently, the systematic risks inherent to LLM-based data extraction remain insufficiently examined. These risks manifest in multiple forms. LLMs may

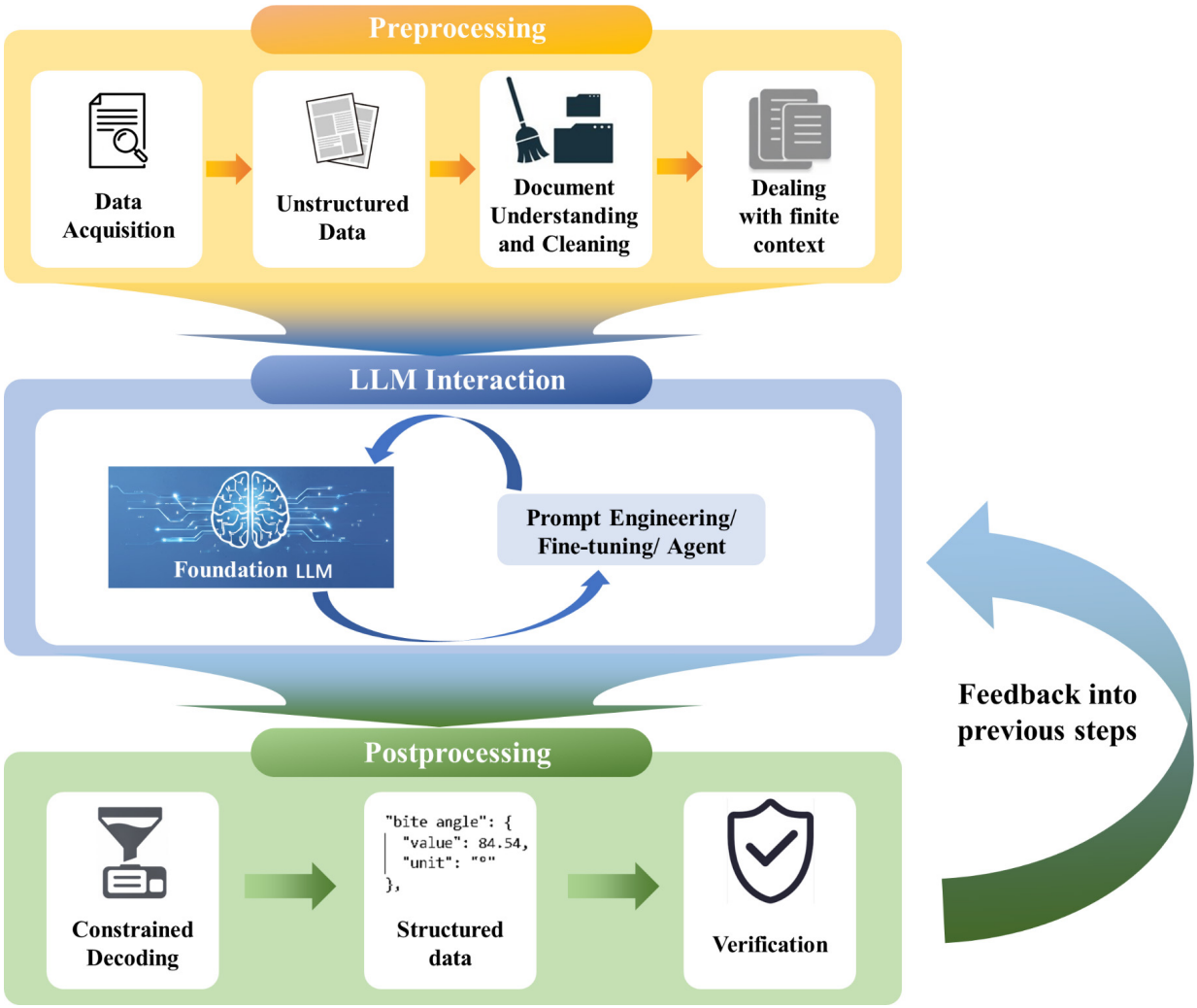


Figure 2. Three-stage workflow for LLM-driven data extraction. LLM: Large language model.

hallucinate facts, make misassociations between conditions and values, propagate errors in multi-step workflows, or quietly turn uncertainties into certainties. Recognizing these risks and developing engineering practices to mitigate them is essential if we are to build reliable, reusable experimental databases. This perspective addresses this gap by examining LLM-based data extraction from an engineering practice perspective, with a focus on identifying systematic risks and developing practical mitigation strategies.

In this perspective, we first dissect the typical LLMbased extraction workflow, including preprocessing, model interaction and postprocessing. We then expose the hidden traps encountered during largescale extraction, informed by practical experience. Finally, we propose a set of guiding principles and future directions to steer the field toward trustworthy, scalable and impactful data extraction.

LLM AS A DATA ENGINE: FROM UNSTRUCTURED TEXT TO STRUCTURED DATA

A successful extraction pipeline consists of three interconnected stages, namely preprocessing, LLM interaction and postprocessing [Figure 2]. Each stage requires careful choices that balance accuracy, cost and robustness.

Preprocessing: making documents digestible

The raw material, like PDFs, HTML or XML, must be converted into a machinereadable format that LLMs can process. Document parsing tools such as OCR engines (Tesseract) or endtoend models (Nougat, Marker) extract text and layout information, producing Markdown or plain text while preserving section structure. Cleaning removes irrelevant parts (references, acknowledgments, page numbers) to reduce noise and token cost. Without such cleaning, irrelevant content can dilute the model's attention, increase the likelihood of extracting spurious or misattributed information (e.g., confusing references with experimental data), and exacerbate context length limitations, ultimately degrading extraction accuracy. For chemistry, retaining only the experimental sections often suffices.

LLMs have finite context windows, so long documents must be chunked[16]. Simple fixed-size chunking risks breaking semantic units, while overlapping chunks or semantic chunking based on section boundaries preserve context[17-19]. For informationsparse corpora, retrievalaugmented generation (RAG) can be employed, where chunks are embedded and stored in a vector database, and only those relevant to the extraction task are retrieved and fed to the LLM, drastically cutting cost while maintaining performance.

LLM interaction: prompting, fine-tuning and beyond

The core decision is how to query the LLM. Prompt engineering is the first resort. A system prompt defines the model's persona ("You are a chemistry expert"), while the user prompt contains the task and the text[20]. Zeroshot prompts are simple but may lack domain knowledge. Fewshot prompts provide a handful of inputoutput examples, enabling incontext learning[21]. The selection of examples, including both positive and negative cases, strongly influences performance. For complex tasks, chainofthought prompting decomposes the extraction into reasoning steps, improving accuracy[22].

When prompting falls short, finetuning on a small and curated dataset can adapt a general LLM to specialized chemical entities and relations[23]. Parameterefficient methods like LoRA (low-rank adaptation)[24] make fine-tuning feasible with modest computational resources. Recent work shows that a 7B-parameter model finetuned on chemistryspecific tasks can outperform much larger general models[25].

Beyond text, chemical data often reside in figures, spectra, and table images rather than plain text. Recent work has shown that combining image-text alignment with LLM-based reasoning enables the extraction of structured information directly from such multimodal sources[26] [Figure 3]. From a workflow perspective, this extends the three-stage paradigm discussed above: in preprocessing, figures and tables are parsed using vision models or OCR tools; in model interaction, cross-modal alignment associates visual elements with textual context; and in postprocessing, consistency checks integrate information across modalities. In practice, hybrid workflows, using vision-language models (VLMs) for initial interpretation and specialized tools (e.g., OSRA for structure diagrams) for verification, offer a pragmatic compromise.

Postprocessing: from model outputs to reliable data

The raw output of an LLM is a probability distribution over tokens. To enforce structure, constrained decoding techniques dynamically restrict the vocabulary to tokens that yield valid JSON, numbers or even chemically valid SMILES strings[27]. This eliminates syntactic errors at the source.

Data normalization is better performed after extraction using deterministic tools. For instance, units can be converted with libraries like pint, and chemical names can be canonicalized via PubChem or the Chemical Identifier Resolver. As a general rule of thumb, anything that can be handled deterministically should not be delegated to a probabilistic model.

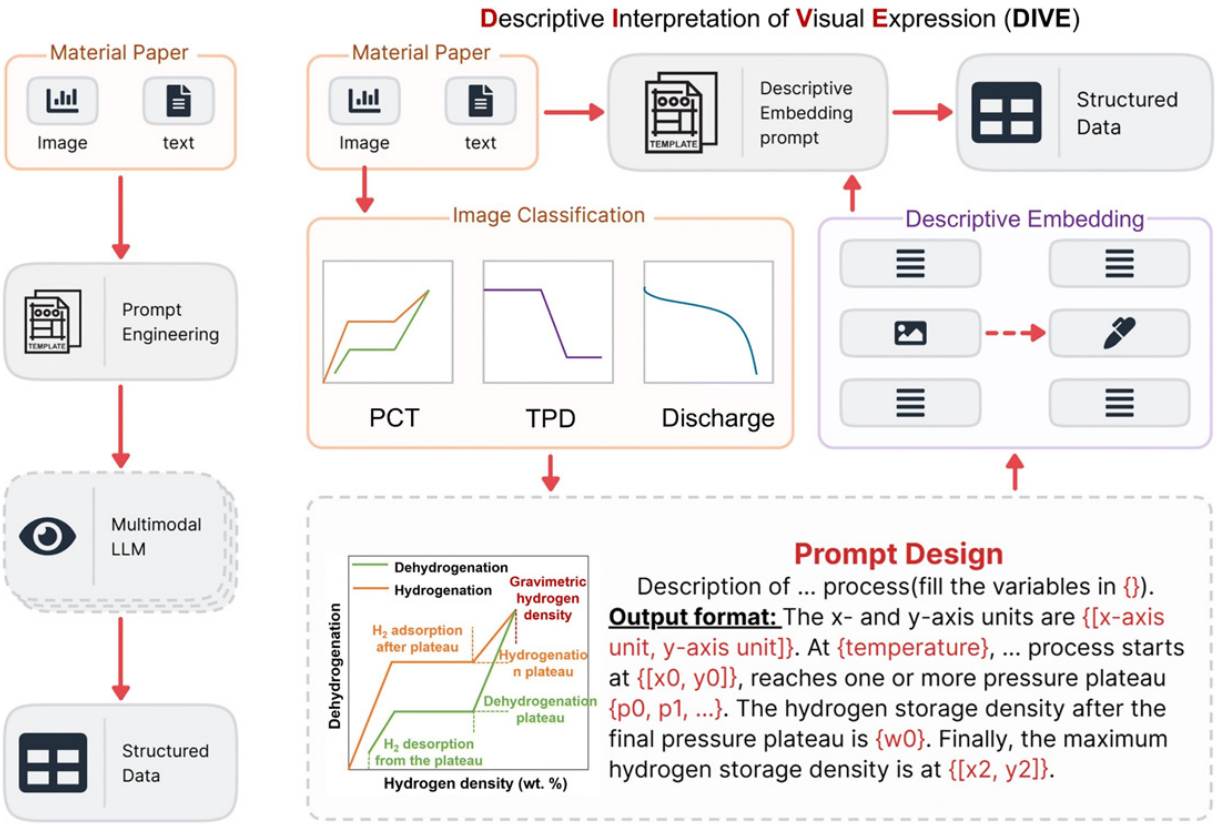


Figure 3. Schematic diagram and evaluation methods of the DIVE workflow[26].

Evaluation of extracted data is nontrivial. Simple metrics such as precision and recall are commonly used for evaluation. For complex nested data, the "matching problem" must first be resolved, that is, determining which item in the ground-truth data corresponds to each extracted entry. Fuzzy matching algorithms such as edit distance can be used for optimal matching[28]. A complementary evaluation strategy is domain-knowledge-based verification, which represents a distinctive advantage in the field of chemistry. A "rationality check" of the extracted results can be performed using chemical rules, such as atomic balance across the reaction or consistency between spectral data and the proposed molecular formula.

However, a critical but often overlooked challenge lies in the construction of reliable ground-truth datasets for chemical information extraction. Unlike standard NLP tasks, chemical data are highly context-dependent, where correct interpretation may require domain expertise (e.g., associating experimental conditions with performance metrics or resolving implicit information in text and figures). As a result, annotations are time-consuming, potentially inconsistent across annotators, and often limited in scale. These limitations introduce uncertainty into evaluation and may obscure the true performance of extraction systems.

SYSTEMIC RISKS AND HIDDEN TRAPS IN LARGE-SCALE EXTRACTION

Despite the promise, deploying LLMs for bulk extraction reveals several insidious pitfalls that can undermine database quality [Table 1].

Risk 1: Condition-value mismatch - the most insidious error. LLMs often correctly extract a numeric value (e.g., "product selectivity") but fail to capture or misassign its experimental conditions (temperature, doping

Table 1. Common risks, manifestations, and mitigation strategies in LLM-driven data extraction

Risk categories	Specific manifestations	Core mitigation strategies	Evaluation signals
Condition mismatch	Extracted performance data but ignored or incorrectly associated the corresponding experimental conditions (e.g., temperature, pressure, doping concentration)	Design structured output schemas that explicitly require the joint extraction of "condition-performance" pairs; apply domain knowledge-based post-validation to verify consistency and correctness	Schema validation; cross-field consistency checks; automated detection of impossible condition-performance pairs
Error accumulation	Single-step accuracy is 99%, but the overall accuracy drops significantly after multiple sequential tasks	Decompose tasks by breaking complex tasks into simpler subtasks; insert validation steps at critical points	Multi-step validation; cumulative accuracy tracking; intermediate verification checkpoints
Alignment bias	Extracted "~400 °C" as "400 °C"; fabricated non-existent data based on context ("hallucination")	Explicitly prohibit fabrication (e.g., "if the source text is ambiguous, mark the field as 'uncertain'"); encourage the model to provide confidence scores	Confidence score thresholds; cross-evidence verification; detection of improbable or inconsistent values
Ambiguous instructions	When input is too long or instructions are unclear, the model engages in ineffective reasoning, degrading performance	Design instructions to be concise and clear; use system prompts to define explicit roles and task boundaries	Instruction clarity metrics; reasoning trace logs; evaluation on benchmark examples
Literature bias	Databases are biased toward positive results and often omit failed experimental data.	Acknowledge this limitation and explore the possibility of sourcing negative or failed data from alternative sources (e.g., laboratory notebooks)	Coverage statistics; ratio of positive/negative outcomes; completeness analysis

LLM: Large language model.

level, measurement method). This occurs because models excel at recognizing isolated facts but struggle with implicit contextual relationships. Downstream models trained on such mismatched data will learn spurious correlations.

Risk 2: Error accumulation in multi-step workflows. When processing thousands of documents at one time, even with a single-step accuracy rate of 99%, the final accuracy rate may drop sharply after the accumulation of multiple tasks (such as identifying compounds → extracting conditions → associating performance) [Figure 4A]. The autoregressive nature of LLMs means a mistake early on propagates through all subsequent generations.

Risk 3: "Ingratiation" bias - turning uncertainty into certainty. When the source text says "approximately 400 °C" or "a possible intermediate," LLMs tend to output "400 °C" or "intermediate: present," erasing ambiguity. Without explicit instructions to preserve uncertainty, the systematically over-confident predictions from LLMs create a false sense of precision in the database [Figure 4B].

Risk 4: Ambiguous instructions lead to ineffective reasoning. Overly long or vague prompts waste the LLM's context window on irrelevant details. In extreme cases, the model may overthink and hallucinate data to satisfy an unclear request. Concise, explicit instructions are essential.

Risk 5: Literature bias and data contamination. Scientific literature overrepresents positive results. Databases built solely from publications will lack failed experiment results, biasing any subsequent analysis. Moreover, LLMs may have been trained on a portion of the test corpus, leading to inflated benchmark scores that do not reflect true generalization.

THE ROAD AHEAD: ENGINEERING PRACTICES AND RESEARCH FRONTIERS

To harness LLMs responsibly, we propose four guiding principles.

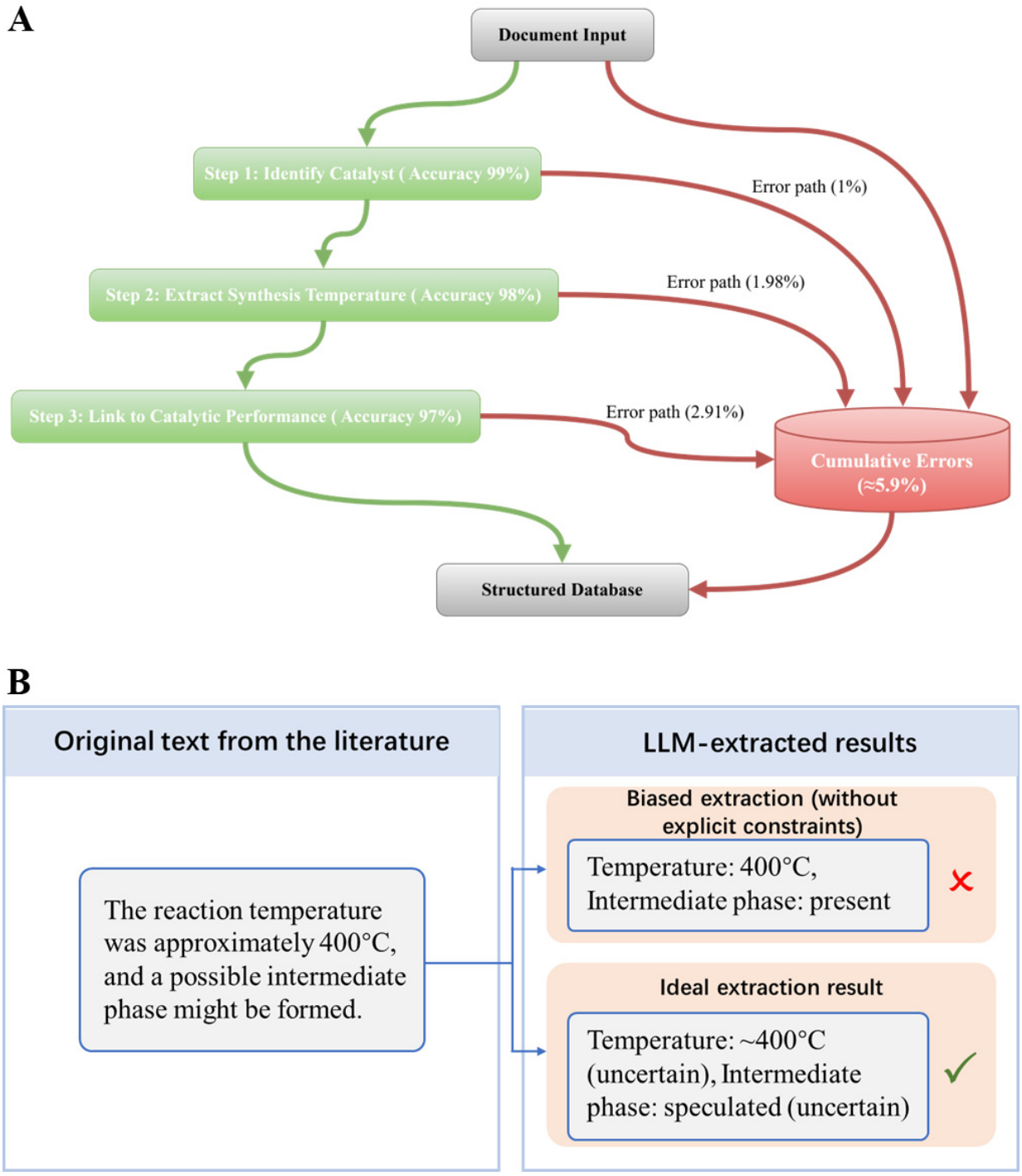


Figure 4. (A) Schematic illustration of error accumulation in large-scale extraction; (B) Schematic illustration of LLM "sycophancy" bias. LLM: Large language model.

Principle 1: Decompose tasks with divide and conquer. Complex extraction goals should be broken into a series of focused subtasks, each handled by a dedicated prompt or model call. This modularity reduces error rates and simplifies debugging.

Principle 2: Multi-layer validation. Build validation at four levels: (1) syntactic level- constrained decoding ensures format compliance; (2) semantic level - chemical rules (atom conservation, charge balance) flag implausible extractions; (3) consistency level- crossrecord checks (e.g., large property variations for the same

material) trigger human review; (4) Uncertainty quantification: where feasible, the model should output confidence scores for low-certainty fields.

Principle 3: Clear division of labor - LLMs for what they do best. It is important to delineate what LLMs are best suited for, such as entity recognition, relation extraction and information summarization, and what should be delegated to deterministic postprocessing tools, such as unit conversion, name standardization and numerical computation. Anything that can be handled deterministically with conventional code should not be delegated to probabilistic LLMs. A mechanism can be established that combines automatic extraction with periodic verification by experts. Recent work demonstrates that such role separation can be effectively implemented through modular pipelines combined with deterministic tool-based postprocessing, with embedding-based matching scores providing a practical way to quantitatively assess extraction quality[26]. LLMs serve as powerful "postdoctoral researchers," whereas human scientists remain the ultimate "principal investigators." This hybrid approach maximizes reliability and reduces cost.

Principle 4: Appeal for standardized reporting at the source. Encourage journals and authors to present key data (synthesis parameters, performance metrics) in machine-readable tables within the main text, rather than burying them in unstructured paragraphs. FAIR data principles should become standard practice. Standardized reporting drastically lowers the barrier for subsequent extraction.

CONCLUSION AND OUTLOOK

Large language models have unlocked the ability to convert the vast unstructured literature of chemistry and materials into structured, actionable data. By following a systematic workflow, including preprocessing, careful LLM interaction and rigorous postprocessing, we can accelerate the construction of highquality experimental databases. Yet this promise comes with profound risks, namely mismatched conditions, error cascades, overconfidence and hidden biases. Realizing the full potential of LLMs requires not only technical innovation but also a disciplined engineering mindset encompassing task decomposition, multilayer validation, strategic leveraging of each tool's strengths, and advocacy for improved data-reporting practices in scientific publishing. When these principles are embraced, LLMs will become trusted partners in scientific discovery, transforming text into insight and powering the next wave of datadriven materials innovation.

Looking forward, several research frontiers will shape the next generation of extraction systems:

Multimodal agents that can seamlessly navigate text, figures, and tables, invoking specialized tools (e.g., crystallographic databases) when needed.

Crossdocument integration to link methods described in one paper with results reported in another, enabling the reconstruction of complete experimental workflows.

Lowresource benchmarks that assess model performance on rare entities, ambiguous descriptions, and outofdistribution samples, providing a realistic measure of generalization.

DECLARATIONS

Authors' contributions

Contributed to the writing and preparation of the manuscript: Chen, W.; Xu, H.; Cheng, D.

Availability of data and materials

Not applicable.

AI and AI-assisted tools statement

Not applicable.

Financial support and sponsorship

This work was supported by the Major Coal Project (No. 2025ZD1701302), the National Natural Science Foundation of China (No. 22522802), and the Fundamental Research Funds for the Central Universities (No. JD2536).

Conflicts of interest

Cheng, D. is an Editorial Board Member of *AI Agent* and was not involved in any step of the editorial process for this manuscript, including reviewer selection, manuscript handling, or decision-making. Cheng, D. and Xu, H. are affiliated with Deep Intelligence Experiment Technology (Beijing) Co., Ltd. and have declared no conflicts of interest. The other author has declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. Zhang, J.; Li, J.; Zhao, G.; Wang, Q.; Guo, Y.; Yang, C. Mining solid-state electrolytes from metal-organic framework databases through large language models and representation clustering. *J. Am. Chem. Soc.* **2025**, *147*, 40496-506. DOI PubMed
2. Moses, I. A.; Barone, V.; Peralta, J. E. Accelerating the discovery of battery electrode materials through data mining and deep learning models. *J. Power. Sources.* **2022**, *546*, 231977. DOI
3. Fan, Q.; Min, G.; Liu, L.; Zhao, Y.; Yu, X.; Yun, S. Accelerate the design of new superhard carbon allotropes in Pca21 space group: high-throughput screening and machine learning strategies. *Diamond. Relat. Mater.* **2024**, *143*, 110928. DOI
4. Tshitoyan, V.; Dagdelen, J.; Weston, L.; et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95-8. DOI PubMed
5. M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 525-35. DOI PubMed PMC
6. Brown, T. B.; Mann, B.; Ryder, N.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165. Available online: https://doi.org/10.48550/arXiv.2005.14165 (accessed 28 May 2026).
7. Ouyang, L.; Wu, J.; Jiang, X.; et al. Training language models to follow instructions with human feedback. *arXiv* **2022**, arXiv:2203.02155. Available online: https://doi.org/10.48550/arXiv.2203.02155 (accessed 28 May 2026).
8. Schilling-wilhelmi, M.; Ríos-garcía, M.; Shabih, S.; et al. From text to insight: large language models for chemical data extraction. *Chem. Soc. Rev.* **2025**, *54*, 1125-50. DOI PubMed
9. Wang, X.; Huey, S. L.; Sheng, R.; Mehta, S.; Wang, F. SciDaSynth: interactive structured data extraction from scientific literature with large language model. *Campbell. Syst. Rev.* **2025**, *21*, e70073. DOI PubMed PMC
10. Chen, H.; Liu, H.; Tew, Y.; Ren, X.; Tang, X.; Wang, X. Distilling knowledge from catalysis literature with long-context large language model agents. *ACS. Catal.* **2025**, *15*, 18244-54. DOI
11. Fu, F.; Li, Q.; Wang, F.; et al. Synergizing a knowledge graph and large language model for relay catalysis pathway recommendation. *Natl. Sci. Rev.* **2025**, *12*, nwaf271. DOI PubMed PMC
12. Bai, X.; He, S.; Li, Y.; et al. Construction of a knowledge graph for framework material enabled by large language models and its application. *npj. Comput. Mater.* **2025**, *11*, 51. DOI
13. Ma, Q.; Zhou, Y.; Li, J. Automated retrosynthesis planning of macromolecules using large language models and knowledge graphs. *Macromol. Rapid. Commun.* **2025**, 2500065. DOI PubMed
14. Zhang, D.; Chen, Y.; Liu, C.; et al. Accelerating catalyst materials discovery with large artificial intelligence models. *Angew. Chem. Int. Ed.* **2026**, *65*, e26150. DOI PubMed PMC
15. Zhang, D.; Jia, X.; Wang, Y.; et al. Digital materials ecosystem: from databases to AI agents for autonomous discovery. *Chem. Sci.* **2026**, *17*, 5782-804. DOI PubMed PMC
16. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762. Available online: https://doi.org/1.48550/arXiv.1706.03762 (accessed 28 May 2026).

17. Bennani, S.; Moslonka, C. A systematic analysis of chunking strategies for reliable question answering. *arXiv* **2026**, arXiv:2601.14123. Available online: https://doi.org/10.48550/arXiv.2601.14123 (accessed 28 May 2026).
18. Allamraju, A.; Chitale, M. P.; Adibhatla, H. S.; Mishra, R.; Shrivastava, M. Breaking it down: domain-aware semantic segmentation for retrieval augmented generation. *arXiv* **2025**, arXiv:2512.00367. Available online: https://doi.org/10.48550/arXiv.2512.00367 (accessed 28 May 2026).
19. Narimissa, E.; Raithel, D. Exploring information retrieval landscapes: an investigation of a novel evaluation techniques and comparative document splitting methods. *arXiv* **2024**, arXiv:2409.08479. Available online: https://doi.org/10.48550/arXiv.2409.08479 (accessed 28 May 2026).
20. Jiang, X.; Wang, W.; Tian, S.; Wang, H.; Lookman, T.; Su, Y. Applications of natural language processing and large language models in materials discovery. *npj. Comput. Mater.* **2025**, *11*, 79. DOI
21. Yong, G.; Jeon, K.; Gil, D.; Lee, G. Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. *Comput. Aided. Civil. Infrastruct. Eng.* **2023**, *38*, 1536-54. DOI
22. Wei, J.; Wang, X.; Schuurmans, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv* **2022**, arXiv.2201.11903. Available online: https://doi.org/10.48550/arXiv.2201.11903 (accessed 28 May 2026).
23. Gupta, T.; Zaki, M.; Krishnan, N. M. A.; Mausam. MatSciBERT: a materials domain language model for text mining and information extraction. *npj Comput. Mater.* **2022**, *8*, 102. DOI
24. Hu, E. J.; Shen, Y.; Wallis, P.; et al. LoRA: low-rank adaptation of large language models. *arXiv* **2022**, arXiv.2106.09685. Available online: https://doi.org/10.48550/arXiv.2106.09685 (accessed 28 May 2026).
25. Li, S.; Wei, S.; Huang, C.; Zhang, Y.; Zhang, G.; Sun, S. Extracting and reconstructing knowledge in materials science literature using large language models. *Commun. Mater.* **2026**, *7*, 31. DOI
26. Zhang, D.; Jia, X.; Tran, H. B.; et al. "DIVE" into hydrogen storage materials discovery with AI agents. *Chem. Sci.* **2026**, *17*, 3031-42. DOI PubMed PMC
27. Gu, J.; Jiang, X.; Shi, Z.; et al. A survey on LLM-as-a-judge. *The. Innovation.* **2026**, 101253. DOI
28. Leaman, R.; Islamaj, R.; Adams, V.; et al. Chemical identification and indexing in full-text articles: an overview of the NLM-Chem track at BioCreative VII. *Database* **2023**, *2023*, baad005. DOI PubMed PMC

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.

© 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.