PCCP

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2025, **27**, 22885

Construction of an artificial-intelligence agent for the discovery of next-generation white-LED phosphors

Zichun Zhou,[ab] Han Zhang,[ab] Chi Song,[a] Chen Ming *[ab] and Yi-Yang Sun *[ab]

Received 17th September 2025,
Accepted 30th September 2025

DOI: 10.1039/d5cp03582a

rsc.li/pccp

Large language models have been extensively employed for scientific research from different aspects, yet their performance is often limited by gaps in highly specialized knowledge. To bridge this divide, in this perspective we take phosphor materials for white LED applications as a model system and construct a domain-specific knowledge base that couples retrieval-augmented generation with a numerical-querying model context protocol. By automatically extracting and structuring data from more than 5400 publications—including chemical compositions, crystallographic parameters, excitation-emission wavelengths, and synthesis conditions—we construct an artificial-intelligence agent that delivers both broad semantic search and exact parameter lookup, each answer accompanied by verifiable references. This hybrid approach mitigates hallucinations, and improves recall and precision in expert-level question-answering. Finally, we outline how linking this curated corpus to lightweight machine-learning models and even automated experimental synthesis facilities can close the loop from target specification to experimental validation, offering a blueprint for accelerated materials discovery.

1 Introduction

Since the technological breakthrough of blue light-emitting diodes (LEDs) in the 1990s,[1] phosphor-converted white LEDs (pc-wLEDs) have rapidly replaced traditional lighting technologies due to their high energy efficiency, environmental friendliness and long lifespan.[2–4] The pc-wLEDs typically comprise a high-efficiency blue LED chip with one or more phosphors. In this configuration, the blue light emitted by the chip excites the phosphors, which then emit yellow or other visible lights, ultimately producing white light. For this purpose, phosphors activated by lanthanide ions, Eu^{2+} and Ce^{3+}, have attracted great attention due to the tunability of the emitted light wavelength and the wide spectrum, which are particularly desired properties for making pc-wLEDs.[5,6]

The continued evolution of lighting technologies demands the development of new phosphors with advanced features, such as a broader color gamut, high quantum efficiency and excellent thermal stability. Recently, violet-light-excited phosphors have garnered significant attention for surpassing conventional blue-light-excited systems, as they hold promise for improving color rendering[7] and eye protection from a high content of blue light.[8] Traditionally, the search for novel phosphors has been guided by empirical guidelines based on crystal field theory and existing experimental results.[9,10] Recently, the field has been increasingly embracing data-driven discovery, leveraging computational tools and machine learning to accelerate the identification and optimization of next-generation phosphors.[11–16]

However, the design and development of phosphors face many challenges at the level of computational simulation. For example, the 4f–5d electronic transitions of the rare-earth ions are influenced by complex physical processes, such as crystal field splitting, electron–phonon coupling and the Jahn–Teller effect.[17,18] Consequently, the positions of their energy levels are sensitive to the local crystal environment and hard to fully capture by empirical rules. In this sense, density functional theory (DFT) based first-principles calculations have become the workhorse method in this field, but still face the challenge of treating the strong correlation effect of 4f electrons of the lanthanide ions.[13,14,19] Recently, machine learning methods have been adopted to predict properties of materials.[20–22]

With the rapid development of large language models (LLMs), their applications have broken through the scope of traditional text processing. LLMs now demonstrate potential for constructing domain-specific intelligent systems and have attracted increasing attention in interdisciplinary areas such as information mining, knowledge reasoning and scientific discovery workflows.[23–25] Compared with traditional models,

[a] State Key Laboratory of High Performance Ceramics, Shanghai Institute of Ceramics, Chinese Academy of Sciences, Shanghai, 201899, China. E-mail: chenming@mail.sic.ac.cn, yysun@mail.sic.ac.cn
[b] University of Chinese Academy of Sciences, Beijing, 100049, China

LLMs possess multimodal comprehension and language generation capabilities, which provide new possibilities for building domain-specific intelligent systems based on the literature and databases.

However, the direct application of LLMs to precision-driven scientific domains, such as rare-earth-doped luminescent materials, is hindered by several key bottlenecks stemming from their limitations. First, at the data level, these models are confronted with two primary challenges: the lack of high-quality, specialized datasets and the temporal cutoff inherent in their training. The latter means that they lack knowledge of the latest scientific discoveries and experimental data that have emerged after their training was completed, preventing timely updates on the state-of-the-art. Second, at the algorithmic level, the general hallucination problem of LLMs evolves into a more critical challenge in scientific applications: a lack of grounding in physical and chemical laws. Lacking an understanding of underlying physical principles, a model may propose synthesis routes that violate the laws of thermodynamics or physically unstable material compositions.[26] Furthermore, their precision in quantitative prediction is also severely lacking—while LLMs excel at qualitative descriptions, they perform poorly when predicting key performance parameters such as spectral peak positions and quantum yields. In this perspective, we discuss strategies to address these issues and construct a basic framework, as shown in Fig. 1. This framework aims to implement a specialized intelligent agent based on LLMs for the design of rare-earth doped luminescent materials.

2 Comparison of current technical routes

To address the above-mentioned issues of LLMs for applications in specialized (or sometimes referred to as vertical) fields, a key requirement is to expand the corpora of the LLMs by including the scientific literature for these fields. For this purpose, retrieval-augmented generation (RAG)[27] and fine-tuning[28,29] are two accessible approaches, which exhibit distinct trade-offs with respect to data dependency, computational cost, knowledge updating mechanisms and model performance.

(1) In terms of data dependency, RAG relies on external knowledge bases, which consist of purposely prepared documents.[26] The collection of literature in PDF format from a specialized field could be directly used as the knowledge base for RAG. For better performance, however, structured literature files as described in Section 3 could be used. In contrast, fine-tuning typically requires a substantial dataset of task-specific labeled data (*e.g.*, question–answer pairs), similar to training an LLM.[29–31]

(2) In terms of computational cost, RAG does not require adjustment of the LLM parameters. Its main expenses lie in building a vector database *via* embedding models and performing retrieval during inference. Detailed implementation will be introduced in Section 4. RAG demands modest hardware resources, but it incurs lower initial costs. In contrast, fine-tuning offers lower inference costs, but it entails much higher initial training costs than RAG. Moreover, as foundational LLMs are frequently updated, each new version often necessitates repeated fine-tuning, increasing resource consumption and maintenance complexity.[32]

(3) In terms of the knowledge updating mechanism, RAG leverages external knowledge bases built from domain-specific corpora to provide up-to-date and context-relevant information during inference. By contrast, fine-tuning integrates new corpora directly into the parameters (or weights) of the LLMs, allowing the model to internalize and generate knowledge from that vertical field without external retrieval during inference. In short, RAG updates the knowledge of the LLMs through dynamic retrieval, while fine-tuning through static parameter updates.[26,27]

(4) In terms of model performance, by acquiring information from external knowledge bases, RAG reduces the incidence of hallucinations of the LLMs through the retrieval process. It is worth mentioning that RAG generates responses that are traceable to the original literature, which makes it particularly suitable for scientific Q&A.[33] In comparison, by integrating the new corpora into the parameters of LLMs, fine-tuning not only improves the model performance by reducing hallucinations, but also extends the generative capability of the LLMs to the specialized field.[34,35]

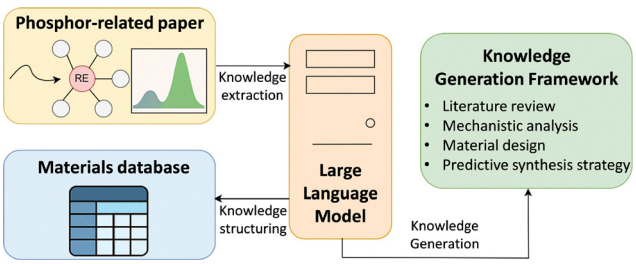

Fig. 1 Schematic of rare-earth doped phosphor agent architecture.

3 Literature structuring preprocessing strategies

The application of LLMs in specialized scientific fields like luminescent materials depends on the availability of a high-quality, structured database. Currently, no such comprehensive public database exists for phosphors, and their key performance metrics—such as luminescence spectra and quantum efficiency—are difficult to obtain through high-throughput, first-principles calculations in the same way as properties like bandgap or formation energy. For this field, the vast body of scattered experimental literature is the primary source of data, making efficient data acquisition a critical bottleneck. Creating structured databases from the ever-expanding corpus of published research remains a complex endeavor. Although LLMs and RAG provide a promising framework, their effectiveness is undermined when raw, unprocessed literature is used as the knowledge base.

Firstly, when processing long, information-dense scientific papers, LLMs often face the problem of context loss. A paper's core arguments, key data, and experimental details—the information needed for a database—are often buried in the middle of the text. During retrieval, the model may excessively focus on the summary content at the beginning and end, thereby overlooking the core evidence that determines the study's validity and reliability. This leads to the extraction of incomplete data and the generation of one-sided or inaccurate insights. Secondly, the inherent rigor and complexity of scientific literature pose a huge obstacle to information extraction. These documents not only contain precise terminology and complex logical relationships, but also rely heavily on non-textual, structured data such as tables, figures, chemical structures, and mathematical equations to present key results. Current LLMs, which are primarily text-based, struggle to directly and accurately parse this multimodal information. This can easily lead to misinterpretation, distortion of data, or even groundless hallucinations, severely compromising the reliability of any database built upon it.

We take Eu^{2+}-doped phosphors as a representative case to illustrate these obstacles. Over 50 years of research on Eu^{2+}-doped phosphors has produced a wealth of experimental data. However, these results are scattered across more than 400 academic journals, as illustrated in Fig. 2, creating significant barriers to systematic integration. The core challenge lies in the extreme heterogeneity of this literature: record formats, terminology, and measurement methods vary widely, leading to severe information fragmentation. This knowledge silos phenomenon hinders the development of comprehensive knowledge in the field. More critically, key performance parameters—such as excitation/emission wavelengths, quantum efficiencies, and thermal quenching temperatures—are rarely presented in a structured format. Instead, they are typically embedded within unstructured text, figure captions, footnotes, or even supplementary information. This severely impedes automated extraction and large-scale analysis. To compound the issue, the reported properties for the same material often vary between publications, further undermining the overall consistency and credibility of the data.

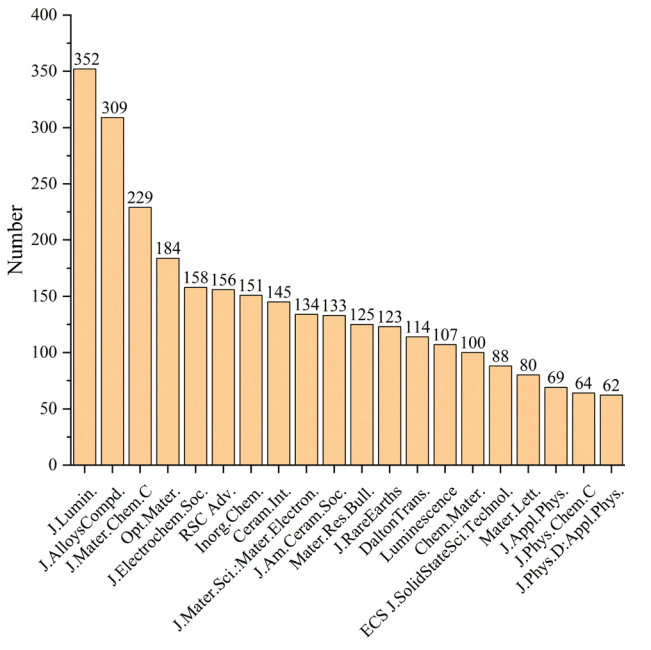

Fig. 2 Top 20 journals ranked by number of publications in Eu^{2+}-doped phosphor research.

Therefore, building an efficient and reliable database from scientific literature cannot be achieved by simply feeding raw documents to a model. A more viable path is to implement a dedicated information extraction and knowledge structuring stage beforehand. By using a data mining approach to transform relationships and core data from text and tables into a structured knowledge base, we can effectively overcome the aforementioned drawbacks and ensure the accuracy, completeness, and reliability of the data foundation for any subsequent RAG system or analysis.

In the past, scientific data mining primarily relied on two approaches: manual annotation and rule-based natural language processing (NLP) systems. Manual annotation is inefficient and prone to subjective bias, making it unsuitable for meeting the growing demand to process high-throughput scientific literature. Rule-based systems, such as ChemDataExtractor,[36] OSCAR4[37] and ChemTagger,[38] possess basic term recognition capabilities. However, they struggle with complex scientific texts that require the interpretation of implicit information, cross-sentence relationships and contextual reasoning. Moreover, these systems depend heavily on domain experts for their construction and maintenance, resulting in high costs and limited portability across domains.

Due to the limitations of traditional methods, generative approaches based on LLMs have emerged as a promising direction for scientific information extraction in recent years.[39–41] Our methodology is built upon this foundation, with a process that begins with the structured preprocessing of the literature. First, we employ optical character recognition (OCR) and layout analysis tools to batch-convert the original PDF documents into Markdown format. This step is crucial as it preserves the document structure—including headings, paragraphs, tables, and lists—providing a high-quality text source for the subsequent precise information extraction. Next, we proceed to the core knowledge extraction phase. We utilize an LLM combined with designed prompt engineering to perform an analysis of the Markdown text. For the phosphor domain, our prompts are designed to automatically extract several key categories of information:

(1) Material compositions: for example, the chemical formula of the host material (*e.g.*, $Y_3Al_5O_{12}$, $CaAlSiN_3$), the activator ions (*e.g.*, Ce^{3+}, Eu^{2+}) and their doping concentrations, as well as any potential co-dopants or sensitizer ions.

(2) Synthesis methods: identifying the specific preparation process, such as the high-temperature solid-state reaction method, co-precipitation, or the sol–gel method, and extracting key process parameters like sintering temperature, holding time, and the use of a reducing or oxidizing atmosphere.

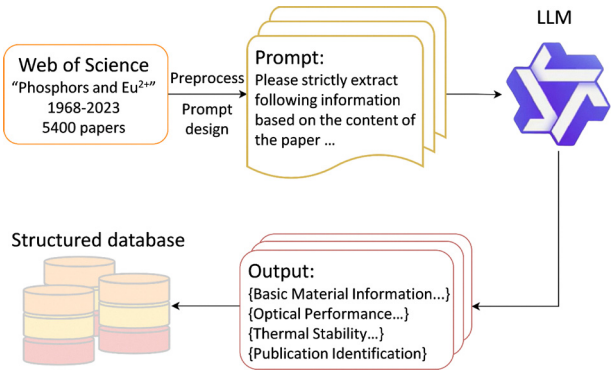

Fig. 3 Framework for building a structured database based on intelligent literature mining.

(3) Performance parameters: precisely capturing core optical and thermal performance data, including the peak wavelengths of the excitation and emission spectra (λ_{ex}, λ_{em}), internal and external quantum efficiency (IQE/EQE), color coordinates (CIE), and thermal quenching behavior (*e.g.*, thermal stability at 150 °C).

Finally, these extracted discrete information elements are systematically organized into standardized structured knowledge units (SKUs). Each SKU can be considered a digital profile for a specific phosphor sample, clearly documenting the material's entire identity–synthesis–performance information chain in a key-value format. These standardized SKUs serve as the cornerstone for building our phosphor knowledge database, enabling efficient support for complex downstream queries and Q&A applications. The overall process is illustrated in Fig. 3.

Thise method not only effectively reduces the interference of irrelevant information by the generative model, but also enhances the readability, controllability and embedding quality of the data. The advantages of structured processing are mainly reflected in two aspects: (1) improving the relevance and precision of information retrieval: similarity calculation based on structured semantic units significantly improves the retrieval recall rate and matching effect; (2) enhancing the contextual support capability: compared with the traditional text input, the structured data provides a clearer contextual context for LLMs to improve the accuracy and rationality of the generated content, which is especially suitable for multi-round Q&A and cross-document integration tasks.

4 Implementations

As discussed above, compared with fine-tuning LLMs to incorporate new knowledge, the RAG architecture offers advantages in cost and scalability. It reduces demands on computational resources and maintenance costs while enabling independent updates of the knowledge base. This allows for more frequent updates, easier customization and better support for data isolation across diverse application scenarios. Here, we suggest an implementation of the phosphor agent, as shown in Fig. 1, by adopting RAG as the core framework to support intelligent Q&A, knowledge retrieval and scientific assistance.

Furthermore, a hybrid system with RAG and a model context protocol (MCP) is constructed, which combines the high recall capability of vectorized semantic search with the high-precision matching capability of queries on the structured knowledge base. This enables a layered information retrieval process that transitions from fuzzy matching to precise extraction.

4.1 RAG

We generated a structured knowledge base for the RAG system, which consists of the research articles representing the whole specialized field. Each article is indexed by its DOI number. We extracted the key information from the article, including but not limited to the basic information of materials (*e.g.*, chemical formula, atomic structure, doping element and doping sites), photoluminescence properties (*e.g.*, excitation/emission wavelength, quantum efficiency, and CIE), and synthesis method. Compared with an unstructured knowledge base, the structured approach not only enhances the retrievability and the quality of data embedding, but also provides a more accurate and higher quality input corpora to the LLMs for the intelligent Q&A.

Based on the structured knowledge base, we built a vector database and indexing system. To improve semantic matching, we adopted Alibaba's open-source embedding model Qwen3-Embedding-8B,[42] which is currently the SOTA of open-source embedding models, according to the HuggingFace MTEB leaderboard.[43] During querying, user questions are vectorized using the same model and matched against the vector database, as shown in Fig. 4.

To improve retrieval accuracy, we used a hybrid scoring mechanism that combines keyword similarity and vector cosine similarity through weighted fusion. This approach balances semantic understanding with precise keyword alignment, reducing false positives caused by overgeneralization. A similarity threshold is also applied to filter out irrelevant results, ensuring that the top-k retrieved documents are semantically relevant. The system shows promising recall and efficiency across multiple test cases, suggesting the potential of the RAG framework for scientific Q&A applications.

In the answer generation phase, new prompts from the retrieved document blocks will be constructed to supplement the user queries. Prompt engineering techniques can be used to

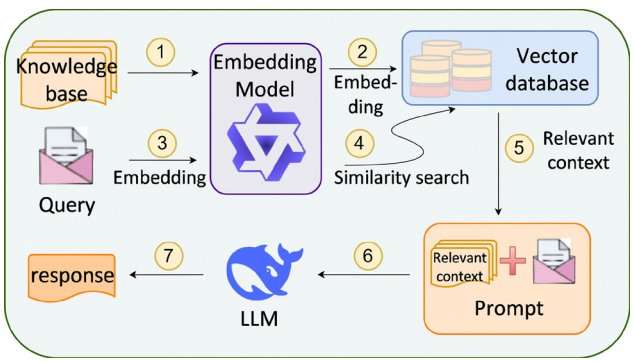

Fig. 4 Schematic of RAG implementation.

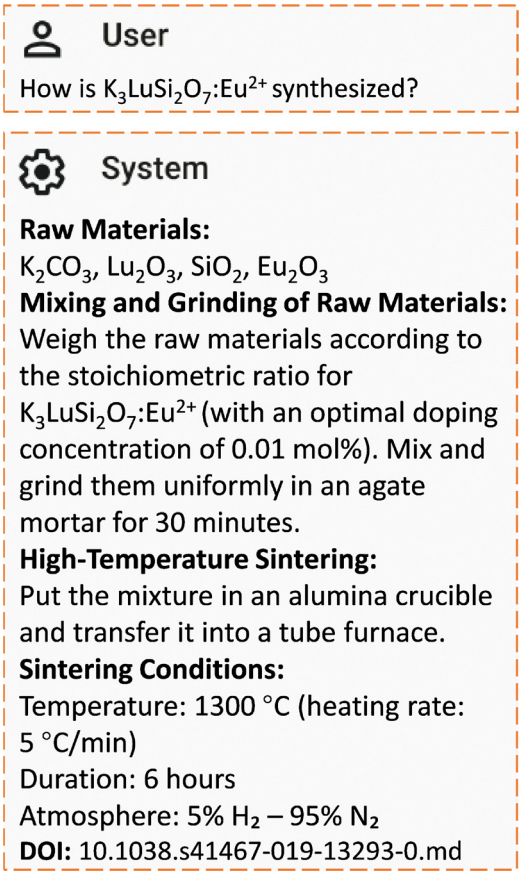

Fig. 5 RAG-based agent for synthesis condition recommendation.

guide the foundational LLM to generate professional answers. Customized outputs can be required. For example, the model can be assigned to play the role of an expert in the field of rare-earth doped phosphors, with an adjustable format and degree of scientific rigor for its answers. After considering the cost and performance, we selected Deepseek-R1[44] as the generative model, considering its relatively strong reasoning capabilities and support for the chain-of-thought mechanism, which can help produce more coherent and insightful responses. An example of the actual Q&A output is illustrated in Fig. 5.

To validate the effectiveness and reliability of our RAG system, we designed a multi-faceted evaluation framework targeting two critical capabilities: novel information processing, and precision of knowledge updates.

(1) To evaluate our RAG system's ability to process novel information, we constructed a specialized test corpus using content published after the baseline LLM's knowledge cutoff. This corpus consists of eight recent phosphor-related papers from 2025, sourced from journals such as advanced optical materials. Against this corpus, we crafted 40 questions meticulously stratified into three types to assess distinct capabilities: precise numerical extraction, recitation of experimental methods, and summarization or inferential tasks. This corpus was ingested into our RAG system, and all 40 questions were posed to both our system and a standalone Deepseek-R1 baseline.

Table 1 Summary of model evaluation results

Model	Accuracy	Faithfulness	Completeness
Baseline model	0.625	—	2
RAG system	1.825	1.825	2

A panel of domain experts then conducted a blind review of all outputs. Each response was scored on the following three core metrics using a three-point scale: accuracy score (0–2): assesses if the core information in the answer is correct. A score of 2 indicates complete correctness, 1 for partial correctness, and 0 for an incorrect answer. Faithfulness score (0–2): measures if the answer is fully based on the provided literature. A score of 2 indicates the answer is entirely based on the source text, 1 for being partially based on some extrapolation, and 0 for fabrication or contradiction. Completeness score (0–2): evaluates if the answer comprehensively addresses all aspects of the question. A score of 2 is given for a complete answer, 1 for a partial answer, and 0 for missing key information. The results, listed in Table 1, demonstrated a marked performance advantage for our system; our RAG model achieved an average score of 1.825 in both accuracy and faithfulness, significantly outperforming the baseline model's accuracy of 0.625. The fact that both systems provided complete answers indicates that the baseline model can understand our questions.

(2) To evaluate the dynamic update capability of our system's knowledge base, we conducted an assessment experiment. The methodology involved augmenting the system's vector knowledge base with multiple synthetic knowledge entries to test its capacity for persistent knowledge integration. Each entry, representing a distinct fictitious fact, was injected as a standalone document. The evaluation was performed by querying the system with two sets of ten questions each: a relevant set directly related to the injected knowledge, and an irrelevant set on unrelated topics. We measured system performance using two core metrics: update success rate and knowledge stability rate.

The experimental results show that the system can absorb new knowledge, achieving an update success rate of 90%. The failure occurred when the system was asked about a recent technology; it presented both the old and new answers simultaneously, indicating a lack of definitive decision-making capability when handling potentially conflicting or outdated information. On the other hand, the system's knowledge stability was excellent. It was not influenced by the new information in any of the tests with irrelevant questions, achieving a knowledge stability rate of 100%.

In summary, the system possesses the ability to integrate new knowledge, and the introduction of this knowledge does not contaminate the pre-existing knowledge corpus. However, the experiment revealed the system's shortcomings in managing knowledge version conflicts and timeliness issues. To address this limitation, we plan to implement a more sophisticated arbitration mechanism in our future work. This will involve incorporating metadata such as publication dates and

impact factors for knowledge sources and performing weighted calculations, thereby enabling the system to automatically identify and select the most authoritative or current information.

Although a structured knowledge base is effective in improving retrieval accuracy and system efficiency, there are still limitations in handling numerical exact matches, as it mainly relies on semantic similarity search. To balance accuracy and flexibility, we constructed a service system based on MCP and kept the complete PDF database as a supplementary resource. Through the hybrid search strategy of the structured knowledge base and original documents, the system not only supports precise queries but also can cope with complex scientific Q&A scenarios that require divergent thinking or contextual reasoning.

4.2 MCP

MCP is a standardized communication protocol for LLMs, enabling dynamic access to external resources for enhanced task execution. It adopts a three-tier "host–client–server" architecture, streamlining integration between models and external tools or data sources *via* a unified interface. While RAG mitigates general knowledge gaps in LLMs, it struggles with the high-precision numerical retrieval required in specialized scientific domains. To address this limitation, we have developed a service architecture based on MCP. This framework equips LLMs with capabilities for precise querying and multi-dimensional perception through two core services.

(1) Precise query service: this service is engineered to overcome the numerical inaccuracies of traditional RAG, especially for querying specific data in fields like rare-earth doped phosphors (*e.g.*, excitation/emission wavelengths). We selected MongoDB, a document-based NoSQL database, for its flexible schema and high scalability, which support real-time data updates. Its native JSON-like format is perfectly compatible with our structured semantic units, simplifying data parsing and manipulation. We encapsulated the database within an MCP server using the mcp–mongo–server module. This architecture enables highly accurate, database-level queries based on specific numerical ranges. Unlike conventional vector search, our approach transcends the top-k limitation, returning a complete set of all results that satisfy the query conditions.

(2) Multimodal visualization service: this service primarily provides visualization for crystal structures, implemented by integrating the JSMol tool. We have designed a dual-call process: when a user queries a crystal structure using a chemical formula, the model first calls the query server to precisely match the formula to its corresponding ICSD (inorganic crystal structure database) number. Subsequently, it calls the resource server to retrieve the CIF (crystallographic information file) for that number and completes the 3D visualization rendering. This not only significantly enhances the model's cross-modal understanding capabilities (from text to 3D images) but also establishes a standardized interface for integrating more modalities in the future, such as spectral diagrams and electron microscopy images.

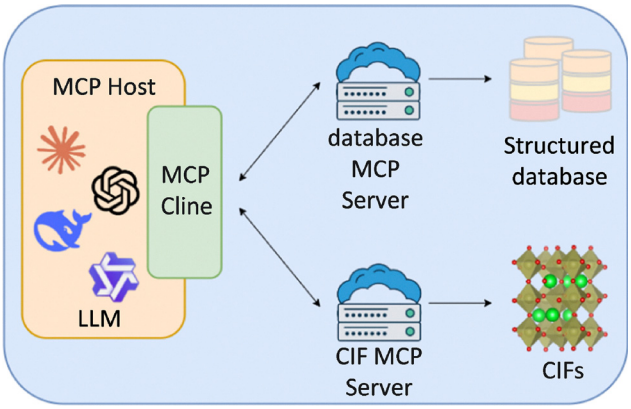

Fig. 6 Schematic of MCP implementation.

Both servers access the MCP client, which is realized by the Cline plugin of VS Code and the overall system architecture is shown in Fig. 6. In the actual invocation, the LLM will judge the task type according to the system prompts and select the corresponding MCP server to initiate the request. An example of the actual Q&A output is illustrated in Fig. 7.

5 Discussion

We introduced above a working Q&A agent for the field of rare-earth-doped phosphors, which has basic functionalities: integrating existing knowledge, performing preliminary reasoning and correlating information from multiple sources. The system in its current form falls short of being a true intelligent agent. It

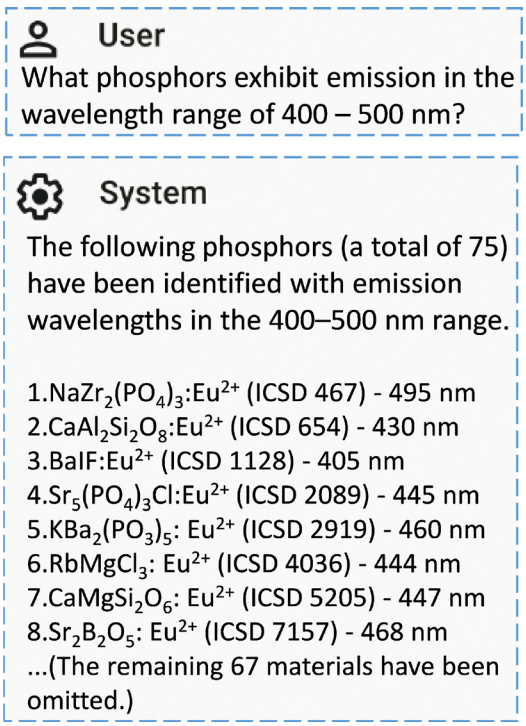

Fig. 7 MCP-based agent for phosphor luminescence performance.

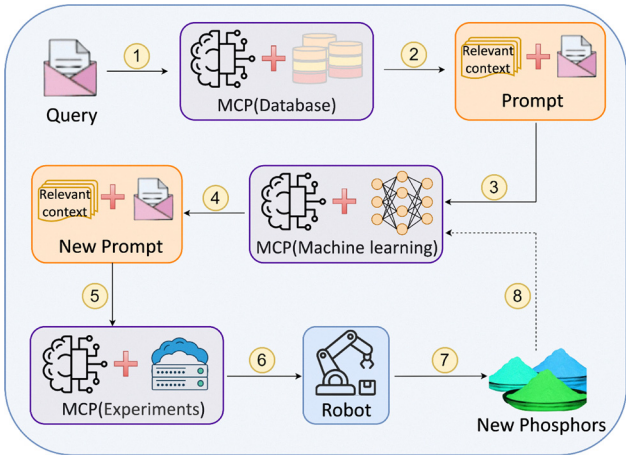

Fig. 8 Future perspectives on intelligent agents for discovering new phosphors.

operates more like a specialized search engine with a Q&A interface, offering a consolidated view of established knowledge. A key limitation is the inability to auto-verify the accuracy of its predictions, and it lacks the mechanisms for continuous learning and self-correction. The next direction of development is to further integrate several key components, including a vectorized knowledge base, machine learning models and an automated experimental system, based on the MCP technology. Through this integration, an intelligent agent with autonomous exploration and synthesis capabilities, as illustrated in Fig. 8, can be expected.

The operational flow of the system is as follows: (1) the user inputs the target performance requirements for the phosphors. (2) The system combining LLM, RAG and MCP queries the knowledge base for materials meeting these requirements; if none are found, it recommends potential candidate materials. (3) Lightweight machine learning models are employed to carry out performance predictions on the candidate materials, serving as a correction to the LLM-RAG-MCP system. (4) The system predicts possible synthesis pathways and integrates them with an experimental protocol and some attached equipment to enable intelligent material synthesis. If the synthesized material meets the target performance, the process concludes. Otherwise, the experimental results are fed back into the agent for further optimization, enabling a closed loop encompassing material prediction, design and execution of experimental synthesis, as well as a feedback mechanism.

The implementation above depends on several key infrastructures and resources, all seamlessly integrated *via* MCPs: (1) a database of phosphor literature, which supports semantic search and knowledge extraction, enabling an understanding of existing research findings; (2) an experimental protocol library, which incorporates standardized process templates to support LLMs in automatically generating experimental procedures; (3) an automated experimental platform, which integrates transport robots with intelligent laboratory equipment to enable end-to-end automation from sample transfer to experimental execution and result collection; (4) a machine learning model library, which brings together both proprietary and open-source models. These models are designed to perform rapid screening and preliminary performance prediction of candidate materials in the early stages of discovery.

Based on the above resources, there is a clear division of functions within the agent: the MCP-database is responsible for extracting information related to the experimental objectives from the literature and recommending potential candidate materials; the MCP-machine-learning is responsible for predicting the key performance indexes of the candidate materials by invoking the model and completing the preliminary screening; the MCP-experiment automatically generates synthesis schemes for the candidate materials and translates them into commands that can be recognized by the experimental equipment; and finally, the robotics automated experimental hardware to carry out the synthesis and characterization processes. Although the current work is still focused on the phosphor system, database and knowledge integration, the architecture shows versatility and scalability. In the future, once extended to a wider range of materials, the system is expected to reshape the materials research and development process: starting from the target properties, the AI agent will generate the material structure, predict the properties, plan the synthesis pathway and execute the experiments automatically, thus truly facilitating the realization of a "robotic scientist" and accelerating the progress towards unmanned laboratories and autonomous materials discovery. Some groups have already made efforts in this direction.[24,45–48]

6 Final remarks

We demonstrate that LLMs can serve not only as tools for information processing but also as integral components deeply embedded within the complex and dynamic process of scientific discovery. The aim is not to replace experiments or traditional computational simulations, but to foster a smarter, more efficient and collaborative research ecosystem. Along this journey, we recognize the challenges ahead, such as enhancing the depth of scientific reasoning within LLMs, improving model interpretability and result reliability and exploring ways to seamlessly interface virtual screening with real-world experiments in the future. Nevertheless, this study offers evidence of the profound application of AI for materials science. It underscores that only through the deep integration of data, algorithms and domain expertise can the full potential of AI in scientific research be unleashed, accelerating the discovery of fundamental materials laws and driving rapid breakthroughs in critical technological fields.

Conflicts of interest

There are no conflicts to declare.

Data availability

Data for this article are available at Eu^{2+} doped phosphors database at **https://doi.org/10.57760/sciencedb.29992**.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2021YFB3500501.

References

1 S. Nakamura, T. Mukai and M. Senoh, *J. Appl. Phys.*, 1994, **76**, 8189–8191.
2 G. Li, Y. Tian, Y. Zhao and J. Lin, *Chem. Soc. Rev.*, 2015, **44**, 8688–8713.
3 L. Wang, R.-J. Xie, T. Suehiro, T. Takeda and N. Hirosaki, *Chem. Rev.*, 2018, **118**, 1951–2009.
4 N. C. George, K. A. Denault and R. Seshadri, *Annu. Rev. Mater. Res.*, 2013, **43**, 481–501.
5 S. Lai, M. Zhao, Y. Zhao, M. S. Molokeev and Z. Xia, *ACS Mater. Au*, 2022, **2**, 374–380.
6 X. Qin, X. Liu, W. Huang, M. Bettinelli and X. Liu, *Chem. Rev.*, 2017, **117**, 4488–4527.
7 D. Steigerwald, J. Bhat, D. Collins, R. Fletcher, M. Holcomb, M. Ludowise, P. Martin and S. Rudaz, *IEEE J. Sel. Top. Quantum Electron.*, 2002, **8**, 310–320.
8 S. M. Pauley, *Med. Hypotheses*, 2004, **63**, 588–596.
9 P. Dorenbos, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2000, **62**, 15640–15649.
10 P. Dorenbos, *J. Lumin.*, 2003, **104**, 239–260.
11 A. Canning, A. Chaudhry, R. Boutchko and N. Grønbech-Jensen, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **83**, 125115.
12 J. Andriessen, E. Van Der Kolk and P. Dorenbos, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2007, **76**, 075124.
13 A. Chaudhry, R. Boutchko, S. Chourou, G. Zhang, N. Grønbech-Jensen and A. Canning, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2014, **89**, 155105.
14 Y. Jia, A. Miglio, S. Poncé, M. Mikami and X. Gonze, *Phys. Rev. B*, 2017, **96**, 125132.
15 Y. Zhuo, A. Mansouri Tehrani, A. O. Oliynyk, A. C. Duke and J. Brgoch, *Nat. Commun.*, 2018, **9**, 4377.
16 L. Jiang, X. Jiang, Y. Zhang, C. Wang, P. Liu, G. Lv and Y. Su, *ACS Appl. Mater. Interfaces*, 2022, **14**, 15426–15436.
17 P. Dorenbos, *ECS J. Solid State Sci. Technol.*, 2013, **2**, R3001–R3011.
18 S. Wang, Z. Song, Y. Kong, Z. Xia and Q. Liu, *J. Lumin.*, 2018, **194**, 461–466.
19 C. Ji, Q. Chen, C.-K. Duan and M. Yin, *Phys. Rev. B*, 2024, **109**, 165152.
20 T. Xie and J. C. Grossman, *Phys. Rev. Lett.*, 2018, **120**, 145301.
21 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**(9), 3564–3572.
22 S.-Y. Louis, Y. Zhao, A. Nasiri, X. Wang, Y. Song, F. Liu and J. Hu, *Phys. Chem. Chem. Phys.*, 2020, **22**, 18141–18148.
23 S. Ouyang, Z. Zhang, B. Yan, X. Liu, Y. Choi, J. Han and L. Qin, *Structured Chemistry Reasoning with Large Language Models*, 2024, **https://arxiv.org/abs/2311.09656**, arXiv:2311.09656 [cs].
24 Y. Kang and J. Kim, *Nat. Commun.*, 2024, **15**, 4705.
25 O. Zhang, H. Lin, H. Zhang, H. Zhao, Y. Huang, C.-Y. Hsieh, P. Pan and T. Hou, *J. Am. Chem. Soc.*, 2024, **146**, 31357–31370.
26 Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang and H. Wang, *Retrieval-Augmented Generation for Large Language Models: A Survey*, 2024, **https://arxiv.org/abs/2312.10997**, arXiv:2312.10997 [cs] version: 5.
27 P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel and D. Kiela, *Advances in Neural Information Processing Systems*, 2020, pp. 9459–9474.
28 Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, E. Pathak, G. Karamanolakis, H. G. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, M. Patel, K. K. Pal, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. K. Sampat, S. Doshi, S. Mishra, S. Reddy, S. Patro, T. Dixit, X. Shen, C. Baral, Y. Choi, N. A. Smith, H. Hajishirzi and D. Khashabi, *Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks*, 2022, **https://arxiv.org/abs/2204.07705**, arXiv:2204.07705 [cs].
29 L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike and R. Lowe, *Advances in Neural Information Processing Systems*, 2022, pp. 27730–27744.
30 H. Soudani, E. Kanoulas and F. Hasibi, Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, Tokyo Japan, 2024, pp. 12–22.
31 A. Balaguer, V. Benara, R. L. D. F. Cunha, R. D. M. E. Filho, T. Hendry, D. Holstein, J. Marsman, N. Mecklenburg, S. Malvar, L. O. Nunes, R. Padilha, M. Sharp, B. Silva, S. Sharma, V. Aski and R. Chandra, *RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture*, 2024, **https://arxiv.org/abs/2401.08406**, arXiv:2401.08406 [cs].
32 C. Wang, Z. Yang, S. Gao, C. Gao, T. Peng, H. Huang, Y. Deng and M. Lyu, *RAG or Fine-tuning? A Comparative Study on LCMs-based Code Completion in Industry*, 2025, **https://arxiv.org/abs/2505.15179**, arXiv:2505.15179 [cs].
33 A. Asai, Z. Wu, Y. Wang, A. Sil and H. Hajishirzi, The Twelfth International Conference on Learning Representations, 2024.
34 H. Li, H. Cao, B. Feng, Y. Shao, X. Tang, Z. Yan, L. Yuan, Y. Tian and Y. Li, *Beyond Chemical QA: Evaluating LLM's Chemical Reasoning with Modular Chemical Operations*, 2025, **https://arxiv.org/abs/2505.21318**, arXiv:2505.21318 [cs].
35 S. M. Narayanan, J. D. Braza, R.-R. Griffiths, A. Bou, G. Wellawatte, M. C. Ramos, L. Mitchener, S. G. Rodriques and A. D. White, *Training a Scientific Reasoning Model for Chemistry*, 2025, **https://arxiv.org/abs/2506.17238**, arXiv:2506.17238 [cs].
36 M. C. Swain and J. M. Cole, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.
37 D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy and P. Murray-Rust, *J. Cheminf.*, 2011, **3**, 41.

38 L. Hawizy, D. M. Jessop, N. Adams and P. Murray-Rust, *J. Cheminf.*, 2011, **3**, 17.
39 W. Zhang, Q. Wang, X. Kong, J. Xiong, S. Ni, D. Cao, B. Niu, M. Chen, Y. Li, R. Zhang, Y. Wang, L. Zhang, X. Li, Z. Xiong, Q. Shi, Z. Huang, Z. Fu and M. Zheng, *Chem. Sci.*, 2024, **15**, 10600–10611.
40 X. Jiang, W. Wang, S. Tian, H. Wang, T. Lookman and Y. Su, *npj Comput. Mater.*, 2025, **11**, 79.
41 S. Gupta, A. Mahmood, P. Shetty, A. Adeboye and R. Ramprasad, *Commun. Mater.*, 2024, **5**, 269.
42 Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang and J. Zhou, *Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models*, 2025, **https://arxiv.org/abs/2506.05176**, arXiv:2506.05176 [cs].
43 MTEB Leaderboard - a Hugging Face Space by mteb, **https://huggingface.co/spaces/mteb/leaderboard**.
44 DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang and Z. Zhang, *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*, 2025, **https://arxiv.org/abs/2501.12948**, arXiv:2501.12948 [cs].
45 J. Hu, H. Nawaz, Y. Rui, L. Chi, A. Ullah and P. O. Dral, *Aitomia: Your Intelligent Assistant for AI-Driven Atomistic and Quantum Chemical Simulations*, 2025, **https://arxiv.org/abs/2505.08195**, arXiv:2505.08195 [physics].
46 W. Yuan, G. Chen, Z. Wang and F. You, *Adv. Mater.*, 2025, 2502771.
47 A. Slattery, Z. Wen, P. Tenblad, J. Sanjosé-Orduna, D. Pintossi, T. Den Hartog and T. Noël, *Science*, 2024, **383**, eadj1817.
48 C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, Z. Wang, A. Shysheya, J. Crabbé, S. Ueda, R. Sordillo, L. Sun, J. Smith, B. Nguyen, H. Schulz, S. Lewis, C.-W. Huang, Z. Lu, Y. Zhou, H. Yang, H. Hao, J. Li, C. Yang, W. Li, R. Tomioka and T. Xie, *Nature*, 2025, **639**, 624–632.