Artificial Intelligence Chemistry 2 (2024) 100076

Contents lists available at ScienceDirect

Artificial Intelligence Chemistry

journal homepage: www.journals.elsevier.com/artificial-intelligence-chemistry

Large-language models: The game-changers for materials science research

Songlin Yu [a,b], Nian Ran [a,b], Jianjun Liu [a,b,c,*]

[a] State Key Laboratory of High Performance Ceramics and Superfine Microstructure, Shanghai Institute of Ceramics, Chinese Academy of Sciences, 1295 Dingxi Road, Shanghai 200050, China
[b] Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[c] School of Chemistry and Materials Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Science, 1 Sub-lane Xiangshan, Hangzhou 310024, China

ARTICLE INFO

Keywords:
Large language models
Agent
LLM for materials
Artificial intelligence
Materials science research
Intelligent laboratory

ABSTRACT

Large Language Models (LLMs), such as GPT-4, are precipitating a new "industrial revolution" by significantly enhancing productivity across various domains. These models encode an extensive corpus of scientific knowledge from vast textual datasets, functioning as near-universal generalists with the ability to engage in natural language communication and exhibit advanced reasoning capabilities. Notably, agents derived from LLMs can comprehend user intent and autonomously design, plan, and utilize tools to execute intricate tasks. These attributes are particularly advantageous for materials science research, an interdisciplinary field characterized by numerous complex and time-intensive activities. The integration of LLMs into materials science research holds the potential to fundamentally transform the research paradigm in this field.

1. Introduction

LLMs are artificial intelligence methods that learn the syntax and semantics of natural language from massive text corpora [1]. Since text is a medium for transmitting human knowledge, through training, LLMs can not only comprehend and produce natural language, but also encode human knowledge into their weight parameters. This enables humans to communicate with LLMs in the most natural way (natural language), and bestows upon LLMs a more extensive knowledge domain than that possessed by an individual human [2].

Materials science research is an interdisciplinary field that aims to design novel materials, discover new phenomena, optimize existing materials, predict properties and behaviors, and elucidate underlying mechanisms [3]. These objectives often necessitate the integration of theory and experiment, as well as the mastery of multidisciplinary knowledge, tools, and methods. This imposes high demands on researchers, and also entails a time-consuming and labor-intensive process. With the rapid development of LLMs, there is evidence that they can effectively enhance research productivity [4]. Therefore, we will investigate the profound changes that LLMs may entail for materials science research from three perspectives: LLMs for knowledge acquisition, LLMs for materials research, and agent for materials research. Beyond that, we also discuss the current challenges of applying LLMs to materials science research and the future direction in the following.

2. LLM for knowledge acquisition

The acquisition and utilization of knowledge play an important role in innovation of new materials, and performance optimization of known materials [5,6]. Nonetheless, materials scientists encounter a significant hurdle when it comes to accessing and distilling pertinent information from the extensive and varied literature, particularly when delving into unfamiliar domains or interdisciplinary subjects. LLMs offer a practical and efficient solution to this predicament, as they are adept at handling extensive text data and generating natural language outputs that can summarize, inquire, respond to questions, and elucidate relevant knowledge [7,8].To illustrate, LLMs have the capability to utilize retrieval functions to extract pertinent papers and data in response to a specific research topic or query. They can also amalgamate the key discoveries and contributions of individual papers or a collection of papers [9]. Furthermore, when confronted with an unfamiliar concept, LLMs can facilitate a multi-level comprehension, bridging the gap from fundamental principles to macroscopic phenomena through interactive exchanges between humans and machines [1]. These applications can

* Corresponding author at: State Key Laboratory of High Performance Ceramics and Superfine Microstructure, Shanghai Institute of Ceramics, Chinese Academy of Sciences, 1295 Dingxi Road, Shanghai 200050, China.
E-mail address: jliu@mail.sic.ac.cn (J. Liu).

https://doi.org/10.1016/j.aichem.2024.100076
Received 30 September 2023; Received in revised form 31 July 2024; Accepted 20 August 2024
Available online 24 August 2024
2949-7477/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

significantly aid researchers in improving their ability to access and comprehend existing literature, as well as in acquiring new knowledge more effectively.

The expeditious acquisition of domain-specific knowledge through the LLMs harbors considerable promise. Nonetheless, a judicious approach is imperative. Such models are susceptible to engendering hallucination, proffering information that, while ostensibly plausible, may be egregiously misleading or unequivocally erroneous [10]. This propensity is markedly pronounced in the context of versatile, LLMs that have been indoctrinated on datasets devoid of rigorous specialization in material expertise [1]. To mitigate the incidence of these deceptive semblances, we next discuss several promising techniques that bolster the veracity of domain knowledge acquisition.

2.1. Fine-tuning of LLMs

In the era of rapid advancements in generalized LLMs, a multitude of open-source LLMs, including Llama [11], ChatGLM [12], etc., have been released and widely adopted for scientific research [2]. These models have acquired substantial foundational knowledge through extensive corpus learning. However, they still exhibit limitations in terms of domain-specific depth and granularity, occasionally leading to disillusionment. To mitigate this issue, fine-tuning these models with domain-specific knowledge becomes essential. By incorporating specialized expertise and accuracy related to a specific domain, the fine-tuning process aims to adapt the model more effectively to the nuances and requirements of that domain. Analogously, it transforms a broadly educated generalist into an expert within a particular field of specialization. LLM fine-tuning typically follows two main approaches: Full Model Fine-Tuning (FFT) [13] and Parameter-Efficient Fine-Tuning (PEFT) [14].

In the context of FFT, all parameters, including weights and biases, of a pre-trained model undergo updates during subsequent training phases. This process enables the model to assimilate nuances, discern patterns, and align with specific domain goals. However, this approach typically demands substantial data, computational resources, and time investment. In contrast, PEFT offers a more resource-efficient alternative [13]. PEFT selectively updates only a small subset of the model's parameters, facilitating the acquisition of domain-specific knowledge. A notable representative of PEFT is Low-rank adaptation (LoRA) [15], which adeptly adjusts to downstream tasks by introducing a low-rank matrix. This matrix effectively reduces the number of trainable parameters. For instance, in Fig. 1A, the LoRA fine-tuning process involves freezing the pre-trained weights W (derived from the base model) to prevent new gradient updates. Instead, the newly introduced low-rank matrices A and B contain trainable parameters that contribute to learning domain-specific knowledge during the training phase. In the realm of materials science fine-tuning, high-quality question-and-answer (QA) pairs are commonly employed to achieve this objective [16].

In the realm of materials science, groundbreaking investigations have harnessed large-scale language models for fine-tuning, yielding remarkable outcomes. Xie et al. meticulously curated over 60,000 corpora from a vast corpus of 6000,000 published articles and 16 FAIR datasets specifically within the materials science domain. Their meticulous efforts culminated in fine-tuning the open-source model Llama2–7B, resulting in the creation of a substantial model christened DARWIN [16]. During the evaluation phase, DARWIN exhibited an impressive 96.9 % accuracy on a selected subset of the SciQ dataset, surpassing the performance of the previous SOTA model. Notably, DARWIN also outperformed GPT-4 in addressing specialized inquiries related to materials science (Fig. 1B). Moreover, Yin et al. further refined the comprehensive model by leveraging an extensive corpus of 26.5 million abstracts and 300,000 full-text full paper, amounting to a staggering 15 billion tokens. The culmination of their efforts resulted in the release of MatGPT, currently the largest foundational model tailored specifically for materials science [2]. MatGPT, based on the Llama or GPT-NeoX architectures, exhibits superior predictive accuracy for material bandgaps compared to existing models such as MatSciBERT [17]. Beyond open-source models, commercial entities like OpenAI provide users with the opportunity to fine-tune closed-source models (e.g., GPT-3, GPT-3.5 Turbo) through API interfaces. Notably, in a Q&A dataset focused on battery materials, the fine-tuned GPT-3 model achieved an impressive 89 % answer accuracy, significantly surpassing the performance of the unfine-tuned GPT-3.5 model (which achieved only 61 %) [5]. These instances underscore the efficacy of domain-specific fine-tuning in mitigating modeling biases and enhancing the reliability of domain knowledge acquisition. Moreover, they underscore the immense potential of LLMs tailored to specific scientific domains.

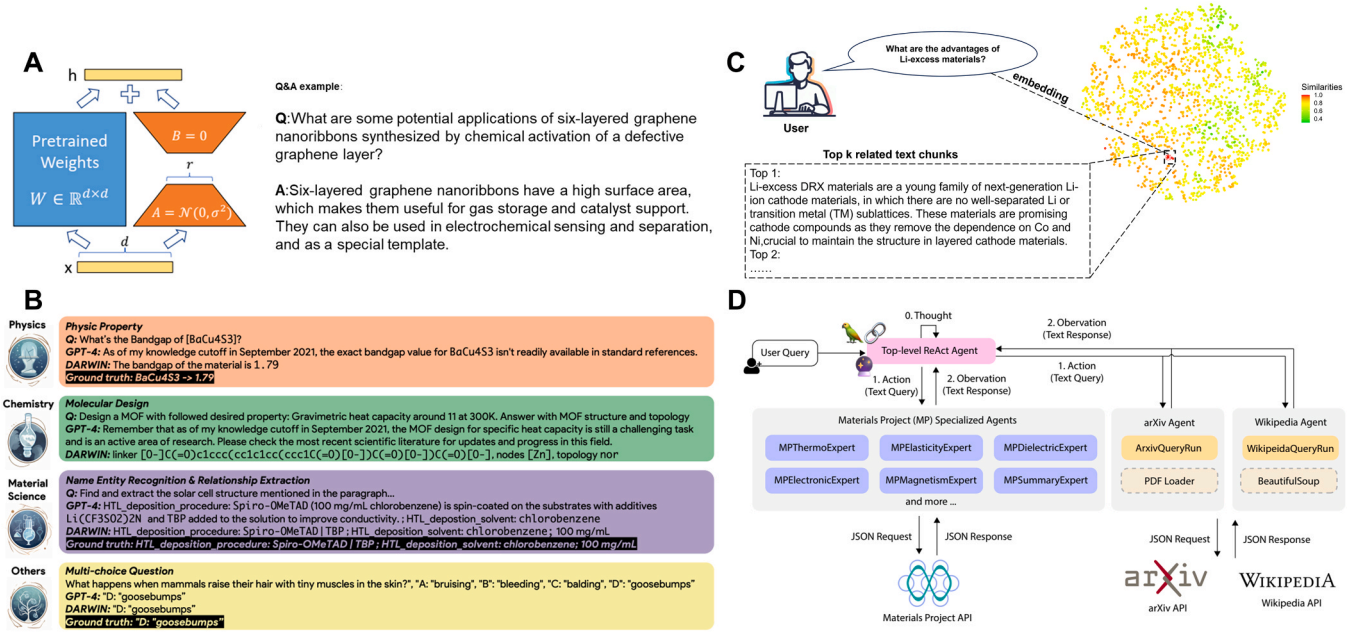

Fig. 1. Fine-tuning and RAG for LLMs. A, LoRA [15] efficient fine-tuning schematic and an example of a Q&A pair used for fine-tuning. B, DARWIN [16] model vs. GPT-4 in answering scientific questions. C, Schematic diagram of RAG. D, Schematic of LLaMP [21] for a multi-modal RAG framework.

Here, we emphasize that supervised fine-tuning is an efficient and cost-effective method that can significantly improve the performance of LLMs in downstream material science tasks. This approach is particularly advantageous in research environments with limited computational resources or smaller data scales. Compared to training a dedicated large model for material science from scratch, the cost of fine-tuning is almost negligible. A notable example is the Llama 2–7B model, which, after being fine-tuned, performs even better on specific material science tasks than the much larger and more advanced GPT-4 model [16]. Moreover, the choice of base model has a decisive impact on the effectiveness of fine-tuning. It is recommended to select a base model that excels in general material science tasks. Such models typically have stronger generalization capabilities and richer material science knowledge, enabling them to better adapt to the data distribution of specific downstream tasks.

2.2. Retrieval augmented generation for LLM

Retrieval Augmented Generation (RAG) is a technique that enhances the capabilities of LLMs by integrating information from external knowledge sources. This approach enables LLMs to generate more accurate and contextually aware responses while reducing the incidence of hallucinations [18]. The process, illustrated in Fig. 1C, operates as follows: when a user submits a query, RAG retrieves the top-k most relevant pieces of information from a pre-vectorized knowledge base or a large corpus of documents using vector search. The knowledge base or documents are pre-encoded using embedding models (e.g., text-embedding-ada-002, BERT), allowing for the computation of cosine similarity between the user query's embedding and the vectorized knowledge base. The top-k results with the highest similarity scores are returned and incorporated into the LLM's contextual framework, thereby enhancing response quality [19]. This mechanism is particularly advantageous in the rapidly evolving field of materials science, where the static nature of the parameter knowledge in LLMs can be a limitation. By leveraging RAG, LLMs can access the most current information directly, without the need for retraining, and generate responses based on the retrieved data [18]. Additionally, RAG offers the benefit of traceability, providing verifiable data sources for the generated answers, which enhances the reliability and credibility of the responses [20].

The integration of RAG with LLMs has been effectively implemented in the field of materials science. Yuan et al. developed a multimodal RAG framework named LLaMP, which incorporates three data-aware reasoning and action (ReAct) agents: the Materials Project agent, the arXiv agent, and the Wikipedia agent (Fig. 1d) [21]. These agents dynamically interact with their respective external data sources. Compared to ChatGPT-3.5, LLaMP significantly reduces errors in material property predictions. For example, it reduces the mean absolute percentage error (MAPE) in bandgap predictions from 5.21 % to nearly negligible levels. Additionally, it corrects substantial errors in formation energy predictions, lowering the MAPE from 1103.54 % to much more acceptable values. By leveraging high-fidelity and up-to-date data from the Materials Project, LLaMP provides responses that are almost free from hallucinations. Furthermore, while understanding and generating three-dimensional atomic structures pose a formidable challenge for unrefined LLMs, LLaMP, aided by RAG, successfully generated a plausible crystal structure (inserting a lithium atom into the tetrahedral position of a diamond cubic silicon structure), whereas ChatGPT-3.5 produced significantly flawed results. These findings demonstrate that RAG can effectively address the inherent limitations in LLMs' knowledge encoding.

Despite the significant improvements RAG technology has made in reducing hallucinations in LLMs, relying solely on vector similarity search can result in responses that lack precision and logical coherence when dealing with complex queries and reasoning tasks [22]. A novel approach, Graph RAG, addresses this issue by constructing graph models from knowledge graphs, representing entities and their relationships as graphs, and then enhancing retrieval using LLMs. Graph RAG treats the knowledge graph as an extensive vocabulary, with entities and relationships analogous to words [22]. By jointly modeling entities and relationships, Graph RAG more accurately interprets query intent and provides precise retrieval results. Buehler et al., leveraging Graph RAG technology, developed the MechGPT model. In their Q&A framework, the graph structure allows for multiple hops, infusing concepts and their interrelationships into the responses [23]. Compared to traditional RAG, Graph RAG-based answers offer more detailed and insightful perspectives. These advancements open new possibilities for tackling complex queries and reasoning tasks, while also helping to overcome the limitations of LLMs.

2.3. Prompt engineering strategy

Prompt engineering constitutes a critical component in mitigating the hallucination propensity of LLMs within the domain of materials science. Through meticulously constructed prompts, researchers can effectively steer LLMs towards generating highly pertinent and factually sound information concerning materials science [24]. Fundamentally, prompts impose robust contextual constraints on LLMs, thereby circumscribing their generative potential and attenuating the likelihood of hallucinated outputs.

Specific and detailed instructions are integral to prompt engineering. Compared to vague inquiries, precise and refined questions can significantly narrow the model's search space, thereby diminishing the probability of generating inaccurate or irrelevant information [1]. For instance, instead of posing a broad question like "What are the properties of new materials?", a more focused query such as "Describe the thermal and electrical properties of graphene and its applications in electronic devices" delineates the scope and depth of the required information, guiding the model to deliver more specialized and in-depth responses.

Providing comprehensive background information is a crucial strategy for enhancing the accuracy of LLM outputs [25]. By embedding specific background knowledge into prompts, researchers can assist LLMs in constructing more complete semantic frameworks, leading to more precise, comprehensive, and insightful responses, and potentially unlocking the model's innovative potential. An effective approach is to incorporate elements of role-playing in the prompts, such as setting the stage: "You are an expert with deep expertise in polymer processing, particularly with extensive experience and professional knowledge in the rubber industry. Please detail the primary processing methods for silicone rubbers and analyze the pros and cons of each method."

Utilizing domain-specific terminology and structured input formats can further enhance the authenticity of LLM outputs [26]. Employing precise scientific terms familiar to materials scientists and standardized data presentation formats can reduce ambiguity and misinterpretation. For example, a prompt like "List the crystal structures and corresponding lattice parameters of commonly used semiconductor materials in photovoltaic cells" encourages the LLM to provide detailed and accurate information according to the conventions of the field.

Incorporating guiding statements into prompts is also a proactive strategy for mitigating hallucinations in large models. One effective method is the Chain-of-Thought (CoT) technique, which encourages the model to reason through intermediate steps before generating the final answer [27]. This not only improves the accuracy of the response but also makes the model's reasoning process more transparent and comprehensible. For example, when faced with a complex multi-step material synthesis problem, researchers can guide the model to first list all possible reaction pathways, then analyze the reaction conditions and products step by step for each pathway, and finally determine the optimal synthesis route. By doing so, LLMs can gain a deeper understanding of the problem, leading to more reliable and accurate answers. Additionally, even without additional training data, appending a prompt such as "let's think step by step" after the question can

significantly enhance the quality of the model's responses. This approach is known as Zero-shot Chain-of-Thought (Zero-shot CoT) [28]. Furthermore, few-shot learning, which involves providing specific examples to clarify the task scope [1], enhance context understanding, and improve model adaptability, also can effectively reduce the likelihood of hallucinations in LLMs.

In summary, strategic prompt engineering not only mitigates the likelihood of hallucinations in LLMs but also leverages their capabilities to provide insightful and reliable information. By employing techniques such as clear instructions, incorporating comprehensive background details, and utilizing domain-specific terminology, researchers can harness LLMs to advance the field of materials science with greater confidence and accuracy.

3. LLMs for materials research

A commonly applied approach in materials research involves conducting iterative experiments to investigate novel materials, a process that is both time-consuming and resource-intensive [29]. Over the past decade, machine learning has exhibited considerable potential in expediting the discovery of new materials. However, it also encounters certain challenges. Firstly, materials science is a multidisciplinary field characterized by limited correlations between its various subdomains, leading to a scarcity of data and inadequate datasets [30]. Secondly, constructing descriptors in the field of materials science presents significant challenges, particularly in scenarios where data are scarce. This process necessitates not only extensive trial-and-error experimentation but also demands that researchers possess profound expertise and comprehensive knowledge of the discipline [31]. Thirdly, materials scientists not only require the ability to predict material performance but also need a profound understanding of the intrinsic relationship between material structure and performance. Traditional statistical-based machine learning algorithms often operate as "black boxes" and lack interpretability in this regard [32]. In this section, we will conduct a detailed examination of the application of LLMs in materials design, focusing on the three perspectives previously outlined.

3.1. LLM for tackling data scarcity

Unlike disciplines such as biology and chemistry, which benefit from universal encoding standards, materials science is intrinsically interdisciplinary and operates across diverse scales, leading to a notable absence of standardized encoding rules [33]. Consequently, there is a paucity of structured, data-driven datasets within this field. The majority of technical material information in contemporary scientific literature is encoded in unstructured natural language, encompassing details on structure, processing history, and performance measurements, all embedded within research articles [34]. Therefore, extracting structured data from this unstructured text is imperative to mitigate the data scarcity in materials science.

Traditionally, manually extracting scientific information from literature has been an arduous, time-consuming, and error-prone process [35]. However, the rapid advancement of LLMs has opened up promising avenues for automating data extraction from scientific literature. This general process is depicted in Fig. 2:

1. Acquisition of Unstructured Text: Data sources include XML and PDF format papers from publishers, as well as open-source scientific text databases [36].

2. Data Cleaning: This stage involves eliminating irrelevant content, such as references and XML syntax, and segmenting the unstructured text into paragraphs or sentences. Some methodologies directly attempt document-level data extraction [37].

3. Preliminary Screening: This step filters out paragraphs or sentences that evidently lack data. Common techniques include manual segmentation [35] and using LLMs to assess whether a segment contains data [37]. Segments with data are retained; otherwise, they are discarded. Regular expressions and trained naive Bayes models are also utilized for filtering specific rules and classification [38].

4. Preliminary Data Extraction: This is a pivotal step where LLMs extract material properties of interest from the filtered text. Given the high density of unit knowledge in scientific texts, this poses a substantial challenge for general LLMs. Advanced LLMs, such as GPT-4, have demonstrated superior extraction performance compared to baseline models (e.g., Vicuna-16k and LongChat models) in extracting polymer nanocomposite data without any specific design [39]. Additionally, leveraging the few-shot learning capabilities of LLMs has shown excellent reasoning and relation extraction abilities in named entity extraction tasks within the SuperMat superconductor corpus. Fine-tuning advanced models like GPT-3.5-turbo has further improved accuracy beyond GPT-4's few-shot learning capabilities [39]. Therefore, further fine-tuning on advanced LLMs may be a better option for data extraction.

5. Validation: This step is essential for enhancing data collection quality and mitigating hallucinations in large models. A common strategy involves using chain-of-thought or agent-based methods, enabling LLMs to self-critique and rectify errors, thereby improving accuracy [34,37]. Notably, Polak's research indicates that incorporating reflective historical records into memory can further enhance the accuracy of data extraction [37].

6. Application to Downstream Tasks: The structured data collected can be applied to data-driven materials research. For instance, Yang et al. utilized GPT-4 with prompt engineering to extract a bandgap dataset larger and more diverse than the largest existing manually curated experimental bandgap database. Models trained on the extracted database achieved a 19 % reduction in mean absolute error for bandgap prediction compared to those trained on manually curated

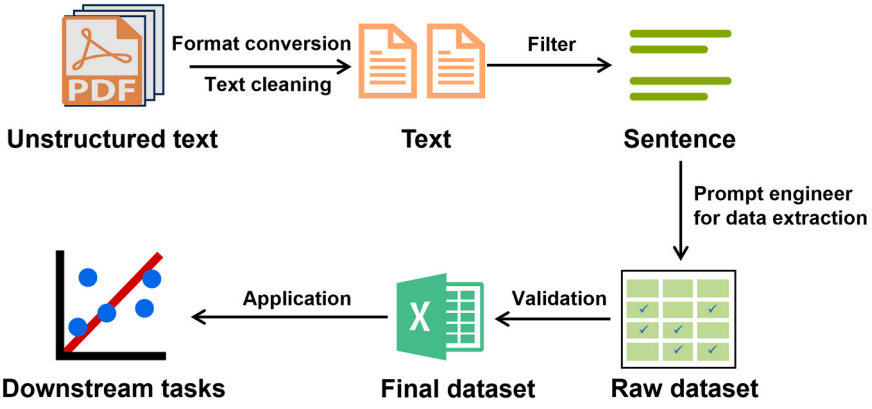

Fig. 2. Schematic of the general workflow of LLMs for structured data extraction in scientific texts.

databases [34]. Similarly, Lee et al. extracted a comprehensive and readily available metal-organic framework (MOF) dataset from over 40, 000 papers, achieving an R^2 value of 0.80 with models trained on experimental data, significantly outperforming the 0.48 R^2 value for models trained solely on computational data [36]. This underscores that experimental data extracted from literature provides more valuable guidance for materials design.

In conclusion, the rapid development of LLMs has significantly mitigated the issue of data scarcity in materials science. Presently, LLMs greatly alleviate researchers from the burdensome and labor-intensive task of data collection. As datasets continue to expand, this advancement is expected to produce a ripple effect, thereby accelerating data-driven materials discovery and advancing the field of materials science. Furthermore, the data extracted from literature closely mirrors the real physical world, and its integration with relatively accessible theoretically derived data promises to bridge the substantial gap between theoretical materials design and experimental validation [34].

3.2. LLMs for materials feature extraction

In data-driven materials science research, the scarcity of data necessitates meticulous construction of descriptors, imposing significant demands on researchers [31]. Additionally, researchers with diverse backgrounds may focus on different aspects of the same material, resulting in poor compatibility between datasets from various sources [33]. The rapid development of LLMs offers a promising solution to these challenges. Research indicates that word embeddings can capture latent knowledge in materials science, enabling LLMs to extract material features effectively [40]. Compared to traditional feature engineering, representing materials in natural language has three distinct advantages: Firstly, LLMs can autonomously extract features, eliminating the need for complex feature engineering due to their ability to directly comprehend natural language [33]; Secondly, LLMs can efficiently leverage the growing corpus of unstructured scientific texts to acquire materials science knowledge [41]; Thirdly, LLMs can adeptly handle features that are difficult to quantify (such as varying synthesis methods) and address missing values across different data sources [33].

Recent studies have demonstrated the extensive application of LLMs in material feature extraction. For instance, Liu et al. showcased that LLMs, trained on a large corpus and guided by meticulously designed prompts, can function as encyclopedic entities, aligning datasets from disparate sources (Fig. 3A) [33]. The meticulously curated material datasets were used to fine-tune MatSciBERT, a base model trained on extensive materials literature [17]. The resultant MgBERT model outperformed traditional logistic regression by 44 % in metal glass classification accuracy and improved classification accuracy by 463 % in unbalanced samples. [33] Moreover, it is noteworthy that LLMs, typically considered to lack spatial imagination when trained on textual data, however, the LLM-Prop model trained after textualization of the 144 K crystal structure outperforms the current state-of-the-art graph neural network (GNN)-based crystal performance predictor by about 4 % in measuring the bandgap, by 3 % in classifying whether the bandgap is direct or indirect, and by 66 % in predicting the unit cell volume [42]. This indicates that, through judicious structural descriptions, LLMs can extract complex spatial structure features. In another study, Huang et al. introduced an innovative textual representation scheme for crystal structures (Fig. 3B), decomposing material crystals into a combination of space group (geometric shape of the crystal), informatics (type of crystal material), and chemical formula (chemical composition of the crystal material) [43]. They pre-trained a feature extraction model specifically designed for material property prediction, MatInformer. Compared to structure-based models (such as CGCNN), MatInformer outperformed CGCNN in 6 out of 8 tasks. Furthermore, textual features extracted by LLMs can also be integrated with other modalities. In Yin's work, LLMs extracted features of chemical formulas, and GNNs extracted features of crystal 3D structures (Fig. 3C) [44]. These were then concatenated for performance prediction, achieving superior results in band gap prediction compared to MF-CGNN (the SOAT model). These examples illustrate that, beyond extensive feature engineering based on atomic distances, angles, and other attributes utilized by many structure-based GNNs, there is much latent information that can be utilized by models, which can be expressed in text and extracted by LLMs.

The material features extracted by LLMs can also guide material design. Qu et al. proposed a novel materials discovery framework that employs embeddings from language models to represent compositional and structural features of materials (Fig. 3D) [41]. Utilizing feature similarity search, they successfully identified several promising thermoelectric materials, which were validated by first-principles calculations. Romas et al. explored the use of LLMs for Bayesian optimization of catalysts through In-Context Learning (ICL) [45]. The principle is that, with carefully designed prompts, predictions of material properties can be achieved. Given that autoregressive LLMs are essentially classification tasks, the softmax values of the predicted outputs can be interpreted as uncertainties. Therefore, leveraging these uncertainties for Bayesian optimization led to faster convergence in drug molecule water solubility tasks compared to Gaussian Process Regression (GPR), and comparable results in the C2 yield dataset. One advantage of this approach is that ICL does not necessitate retraining the model in each iteration, only modifying the prompts, significantly reducing computational costs and time, whereas GPR requires retraining in each iteration. Additionally, Kristiadi's team further utilized LLMs as feature extractors input to Bayesian Neural Networks, achieving superior inference performance in 4 out of 6 molecular discovery tasks compared to using chemical fingerprints (Figs. 3E and 3F) [46]. This demonstrates that LLMs fine-tuned with domain knowledge can provide more detailed information than chemical fingerprints, which can be used to accelerate the research process.

In brief, LLMs trained on extensive materials science knowledge in an unsupervised manner can implicitly encode vast amounts of materials science knowledge. These encoded features can assist researchers in downstream classification and regression tasks. Unlike the hallucinations from LLMs that could lead to serious consequences, this embedding operation is a low-risk application. Despite the high training cost of LLMs, they can be applied to multiple tasks after a single training session, and the domain knowledge is transferable, potentially addressing the issue of data scarcity in materials science.

3.3. LLMs for interpretability materials research

In data-driven materials research, machine learning models that operate as black-box systems have been criticized for their lack of interpretability [47]. LLMs possess significantly more parameters than traditional machine learning algorithms, rendering their internal workings and decision-making processes opaque and often inscrutable [48]. However, interpretability in materials science is not always about the pursuit of strict mathematical expressions for structure-property relationships but rather about achieving a constructive perspective for materials design. From this vantage point, LLMs offer distinct advantages over traditional machine learning algorithms.

Firstly, LLMs trained on scientific corpora encapsulate extensive knowledge from various domains. When queried in natural language, they provide plausible explanations derived from their inherent knowledge encoding (an example see Fig. 4 A). Despite potential biases or inaccuracies, these interpretations can still yield valuable guidance, akin to a novice receiving insights from an experienced mentor whose answers, while not infallible, are often enlightening and useful [48].

Secondly, in terms of generalized interpretability methods, the outputs of LLMs can be elucidated using model-agnostic interpretability techniques. Pre-trained LLMs, possessing a holistic view, may excel in downstream tasks via knowledge transfer, enhancing model feature importance assessments [44]. Traditional machine learning depends heavily on feature engineering to meticulously construct descriptors, but

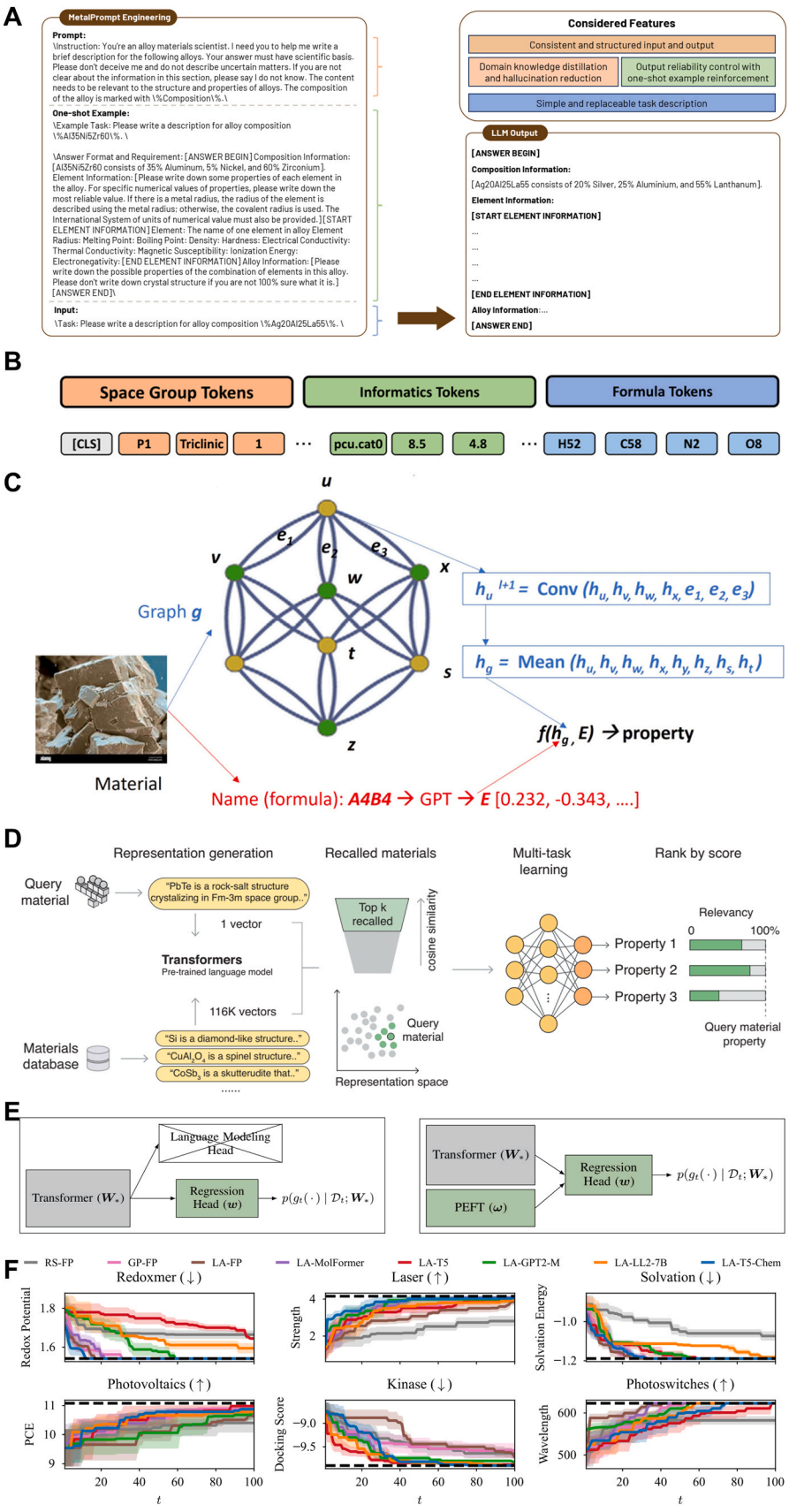

Fig. 3. LLM for feature extraction. A, Schematic diagram of MetalPromt for data alignment from different sources [33]. B, Schematic of material token composition for MatInFormer [43], including spatial group tokens, informatics tokens, and formula tokens. C, Schematic diagram of combining LLM embedding with GNN for material property prediction [44]. D, A workflow for screening candidate materials using the material representation of LLMs [41]. E, Schematic illustration of the feature embedding of LLMs used as input to Bayesian neural networks [46]. F, Evaluation of BO convergence speed with different feature inputs [46].

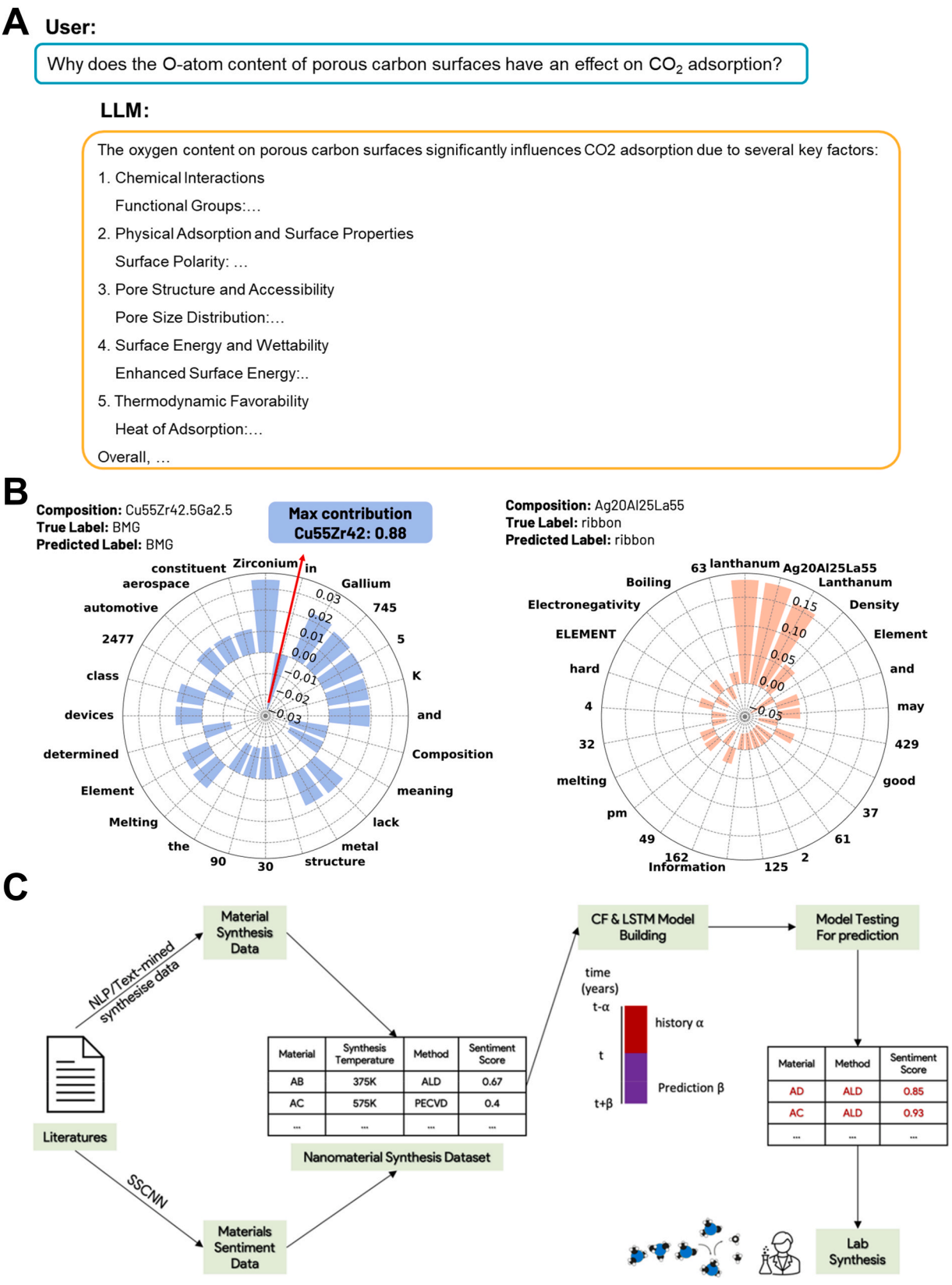

Fig. 4. LLMs for interpretability. A, The interpretability of the LLM in itself via interaction. B, A case where the LIME method is used to explain the relationship between the inputs and outputs of LLMs [33]. C, Architecture of nanomaterial synthesis method prediction system based on NLP and LSTM techniques [49].

in the data-scarce field of materials science, such operations can reduce the number of features and obscure detailed information. Aggregation of features might generate new, relevant descriptors but can also obscure physical and chemical meanings [43]. Conversely, textual inputs to LLMs can encapsulate detailed, meaningful physical and chemical information, offering insights often overlooked by traditional machine learning algorithms. For instance, Liu et al.'s study utilizing the Locally Interpretable Model-Agnostic Explanations (LIME) revealed that in the Cu55Zr42.5Ga2.5 alloy (Fig. 4B) [33], the Cu55Zr42 component predominantly determines its classification as a metallic glass, whereas in the Ag20Al25La55 alloy, this component's influence is weaker. Additionally, textual descriptions capture nuanced feature details, such as the "745" term in the CuZr alloy referring to the ionization energy of copper, which positive impacts predictions but might be disregarded in conventional feature engineering.

From an application perspective, the ability of LLMs to comprehend natural language endows them with substantial potential for text mining, leading to meaningful guided insights. For example, Wang's team utilized ChatGPT to analyze data from 603 articles on ammonia synthesis catalysts, identifying activation pressure, duration, temperature, and heating rate as critical multidimensional parameters influencing catalyst synthesis [9]. These insights are pivotal for subsequent catalyst parameter design using Bayesian optimization. Furthermore, Xie et al. developed the Scientific Sentiment Network (SSNet) employing natural language processing techniques to extract expert opinions on materials from the literature [49]. Combined with a recommendation algorithm and a Long Short-Term Memory (LSTM) network (Fig. 4C), SSNet shows promise in predicting potential synthesis methods for specific materials, highlighting the promising applications of natural language processing in materials science research. Lastly, the integration of LLMs' text comprehension with the symbolic reasoning capabilities of knowledge graphs can produce more coherent explanations, guiding materials design effectively. This integrated approach not only augments the model's explanatory power but also introduces innovative concepts for the research and development of complex material systems.

4. Agent for materials research

In the previous section, advanced LLMs have demonstrated significant potential in the field of materials science. However, is this their sole utility? When relying solely on the intrinsic capabilities of these models, LLMs resemble erudite scholars who possess extensive knowledge yet lack the practical ability to take action. Materials constitute the fundamental basis of the physical world, and research in this domain is notably labor-intensive. A typical materials research workflow encompasses knowledge acquisition, materials design, materials synthesis, materials characterization, and performance verification. This process necessitates iterative cycles until the desired task requirements are fulfilled. Consequently, it is extremely tough to revolutionize the materials research landscape by depending solely on LLMs.

Agents, as derivatives of LLMs, diverge from LLMs in their capacity for action and interaction with the environment, including the physical world [50]. Thus, Agents are poised to play a pivotal role in various facets of the materials research process, potentially reshaping the methodologies and protocols in this field. In this section, we will examine the potential of Agents in materials science research through four lenses: the composition and architecture of Agents, their application in materials design, their role in materials characterization, and their contribution to materials synthesis.

4.1. Agent composition

An agent can be conceptualized as a synthesis of cognitive and instrumental capabilities, enabling it to autonomously comprehend, plan, make decisions, and execute intricate tasks. The operational mechanism of an agent can be delineated as a cyclic process encompassing three fundamental steps: perception, planning, and action. Perception pertains to the agent's ability to gather information from the environment and extract pertinent knowledge. Planning involves the strategic decision-making process aimed at achieving specific objectives. Action entails executing operations based on environmental data and the formulated plan [51].

When tackling complex tasks, an agent integrates its memory with its understanding of the task to facilitate planning, decision-making, and action. Each action influences the environment and generates feedback. The agent then utilizes this feedback to inform subsequent decisions, iterating this cycle until the task is accomplished. In LLM-based agent systems, the LLM serves as the "brain," supported by several essential components (as illustrated in Fig. 5):

I. Planning: A complex task typically encompasses multiple steps that the agent must identify and plan in advance. The planning process consists of four key components:

1. Task Decomposition: The agent divides a large, intricate task into smaller, manageable sub-goals to handle the task efficiently.

2. Chain of Thought (CoT): This standardized prompting technique enhances the model's logical reasoning during decision-making by fostering sequential thinking [27].

3. Self-Critique: This involves real-time critical evaluation of current thoughts and choices to ensure decisions are grounded in sound reasoning and sufficient information. During the planning phase, self-critique helps identify potential biases or errors, prompting the agent to reassess its strategy.

4. Reflection: Post-action review and assessment of events and actions undertaken. Reflection enables the agent to recognize and learn from mistakes or deficiencies, thereby refining future planning.

II. Memory: Agents utilize both long-term memory, for storing past experiences or knowledge, and short-term memory, for holding temporary information necessary for immediate tasks. Short-term memory includes task goals, planning details, current progress, and environmental feedback, all of which assist the agent in making decisions for subsequent actions [52].

III. Tools: These are resources available to the agent for executing actions. The tools integrated into an agent vary according to its specific function. Common tools include search functions (to retrieve real-time information from the internet) and code interpreters (to execute code generated by LLMs), etc. Other agents can also be considered tools, facilitating the construction of multi-agent architectures.

IV. Action: Based on the LLM's output instructions, the agent selects and utilizes the appropriate tools to act on the environment and obtain feedback. Action serves as the conduit between the LLM and its interaction with the environment.

4.2. Agent for materials design

Material design is not only a fundamental aspect of materials science but also a crucial driver of technological innovation and industrial development. In this process, scientists and engineers must consider the chemical composition, structural properties, processing techniques, and final application scenarios of materials. They need to accurately acquire a wide array of relevant knowledge and employ advanced simulations to predict material properties and optimize functionality to meet specific technological requirements and commercial objectives.

In Section 3, we explore the potential of LLMs in accelerating knowledge acquisition, while also highlighting the traceability challenges and risks of hallucinations associated with relying solely on LLMs. Conversely, agents can leverage tools to access real-time information from diverse sources, thereby enhancing the accuracy and traceability of knowledge acquisition. For instance, Chaing et al. developed LLaMP, a multi-intelligent, knowledge-aware agent based on LLMs, which dynamically interacts with data sources such as the Materials Project, Wikipedia, and arXiv (Fig. 1D) [21]. This agent not only rectifies the inherent knowledge errors in GPT-3.5 but also aids researchers in

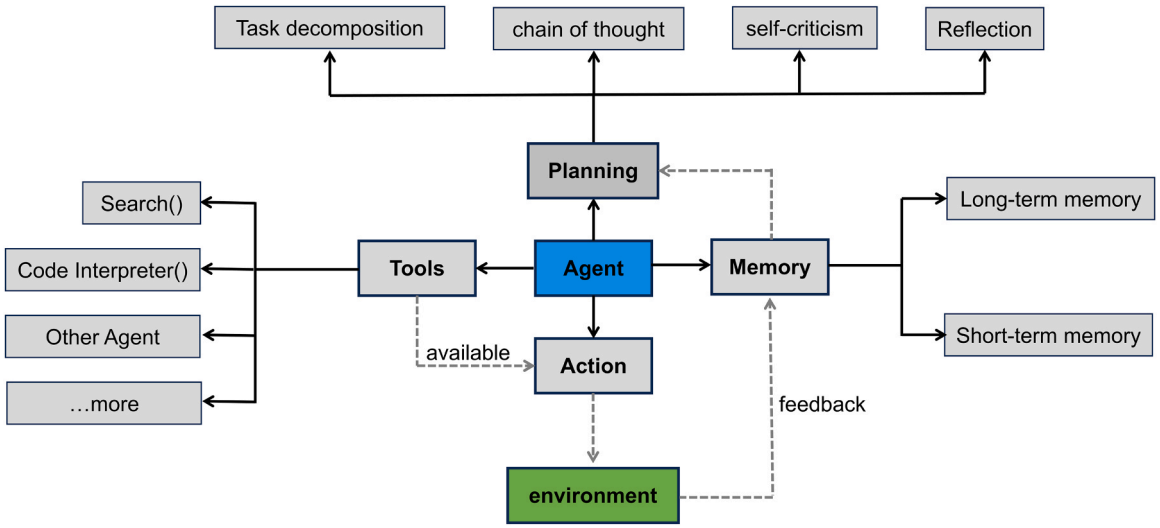

Fig. 5. Schematic of agent composition.

designing accurate crystal structures.

Furthermore, computational simulation plays a pivotal role in materials design. Despite the increasing prevalence of studies that integrate theoretical calculations with experimental work, a skill gap persists among researchers from different backgrounds in bridging experiments and computations. This gap presents a challenge in adopting a comprehensive approach that encompasses both experimental and computational perspectives. Agents offer a promising solution to this challenge. Researchers have already begun to develop intelligent simulation tools. For example, in the study by Buehler et al., an Agent was employed to demonstrate quantum chemical computations, tasked with identifying the lowest energy configuration of an oxygen molecule [23]. The Agent successfully completed this task by writing and executing code, yielding fitted energy maps and force-field parameters (Fig. 6A). Similarly, Liu et al. developed the Lang2Sim framework, which aims to translate human language into executable simulation instructions [53]. This framework is particularly focused on materials simulation, which involves complex scenarios requiring precise translation of natural language into simulation commands. Fig. 6B showcases a Lang2Sim use case where a complex simulation process is accomplished through simple natural language interaction. These examples underscore the significant potential of Agents in materials design, and the judicious use of Agent tools can help researchers to comprehensively think problems from multiple perspectives.

4.3. *Agent for material characterization*

Materials characterization is foundational for understanding the intrinsic properties and behaviors of materials, involving a suite of techniques to elucidate their microstructure, composition, and properties. This process is pivotal for the development of novel materials and the enhancement of existing ones. The advent of advanced

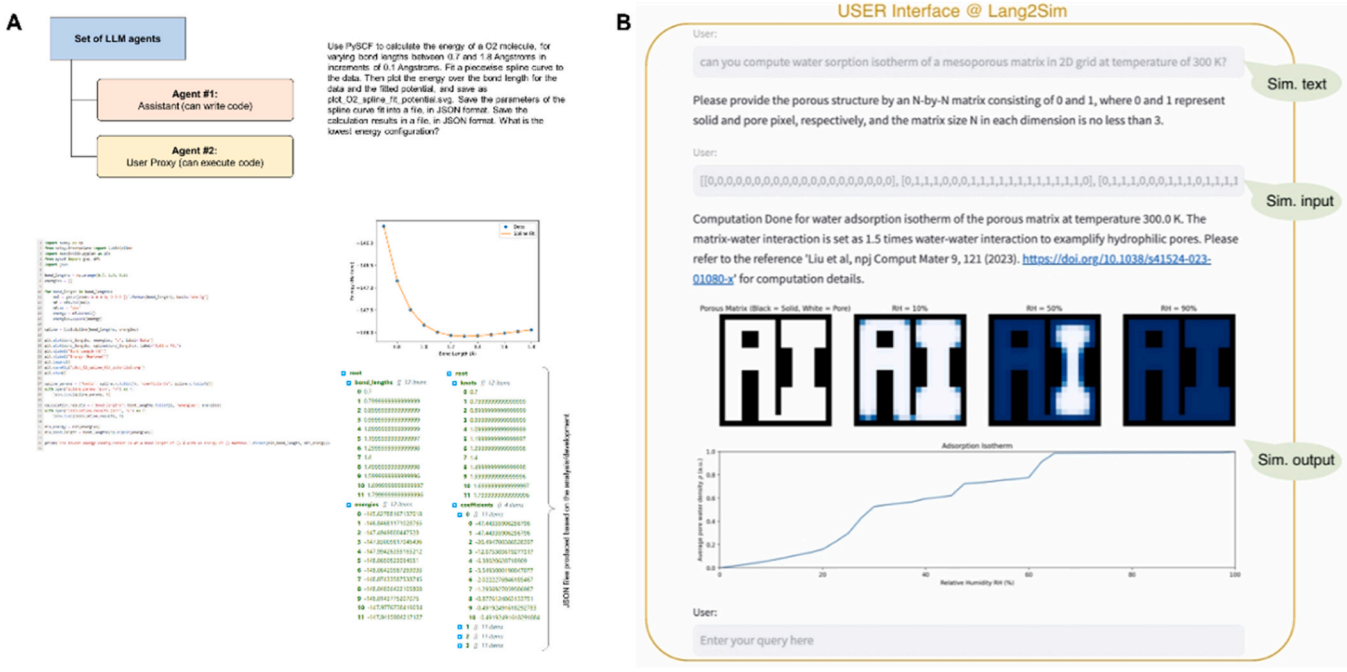

Fig. 6. Agent for materials simulation. **A**, an example of performing quantum chemical calculations using Agents [23]. **B**, a simple Lang2Sim case show for water adsorption simulation via simple human-computer interaction [53].

characterization tools, such as next-generation X-ray sources and neutron facilities, is revolutionizing our comprehension of materials across diverse fields, from life sciences to microelectronics. However, these technological advancements have substantially increased the complexity of operations and data analysis, posing significant challenges for materials scientists in designing and executing experiments.

In this context, Agents based on advanced LLMs exhibit the capacity to manage complex information retrieval, assist in knowledge-intensive tasks, and provide guidance on tool usage and experimental execution. Integrating Agent-based AI assistants with sophisticated characterization instruments emerges as an efficacious solution to these challenges.

For instance, Prince et al. developed a scientific context-aware language model (CALMS), an intelligent agent that integrates tools such as the Materials Project API and instrument control software [6]. This agent successfully answered user queries about initiating a tomography scan, showcasing the potential of Agents to navigate users through the operation of complex characterization instruments. Additionally, when connected to instrument control software, the Agent autonomously operated the instrument upon a reasonable request (see Fig. 7A), significantly streamlining a previously cumbersome process. Similarly, in the study by Liu et al., the Agent demonstrated its applicability in operating a scanning probe microscope [54]. The API integration enabled users to execute experiment-specific workflow scripts directly through the API, minimizing the need for manual, repetitive

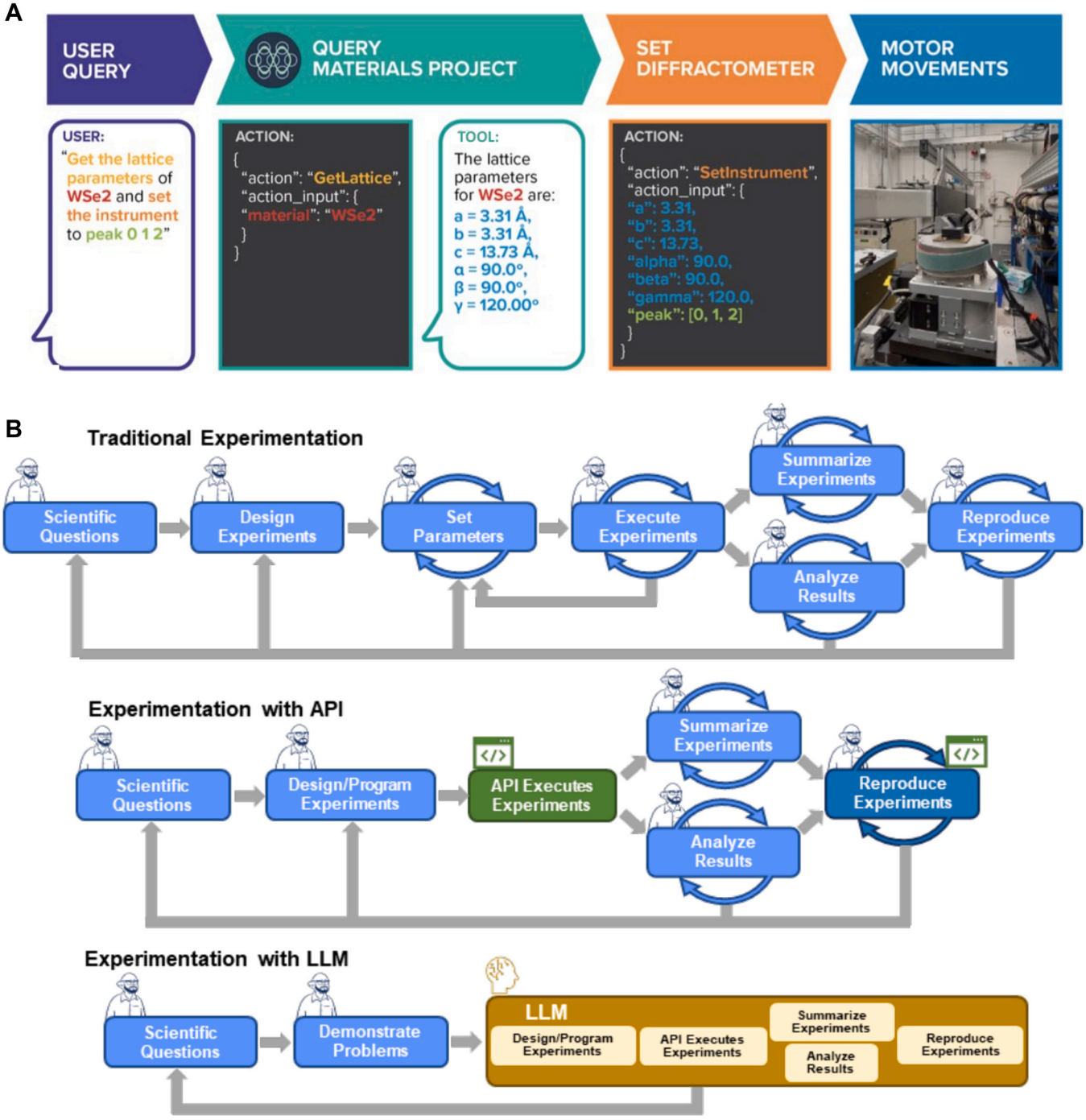

Fig. 7. Agent for material characterization. A, Example of an agent autonomously operating a diffractometer to accomplish material characterization at the reasonable request by a user [6]. B, Comparison of three different experimental models [54].

tasks (see Fig. 7B). The Agent further demonstrated capabilities in parsing user requirements, automating experiments, and assisting in data analysis.

Nevertheless, LLMs currently exhibit limitations in data analysis. For example, untrained models like GPT-4 can only perform general analyses of microscope images and struggle to provide in-depth technical experimental design insights [54]. Despite this, the rapid advancement of data analysis tools, such as AI-driven analysis of XRD [55] and Raman spectroscopy [56], suggests that future integration of these tools into Agents could substantially enhance their effectiveness in materials characterization.

4.4. Agent for materials preparation

Material preparation is pivotal in transforming design concepts into tangible materials, holding a significant position in materials science research. The performance of a material is influenced not only by its intrinsic physical and chemical properties but also by the manner in which it is processed. Traditional materials science research often involves laborious and time-consuming preparation processes, contributing to high costs and inefficiencies. To mitigate these challenges, automated materials and chemistry laboratories utilizing active learning have emerged as a promising solution. Active learning algorithms enhance the efficiency of exploring the material space by selecting the most informative experiments based on real-time experimental feedback, thus reducing experimental costs and aiding in the discovery of materials with desirable properties [57,58].

Despite its potential, this approach encounters several critical challenges: selecting the initial point for active learning is crucial for the algorithm's efficacy [9]. Determining the optimal dimensionality for optimization is also vital, as the complexity of optimization increases exponentially with higher dimensions [59]. Additionally, bridging the gap between material design and synthesis, as well as determining the appropriate preparation processes, remains challenging [60].

Although the application of agents in materials synthesis is relatively underreported, pioneering efforts have demonstrated their significant potential in intelligent materials synthesis. Unlike traditional automated laboratories, agents possess autonomous reasoning capabilities and adeptness in tool usage, marking a transition from mere automation to true intelligence. For instance, agents can leverage LLMs to extract the relationship between material structures and properties from the literature, identifying valuable initial candidates via machine learning algorithms. Wang's team utilized LLMs for text mining to uncover key process steps influencing catalysts, which was instrumental in determining optimization dimensions [9]. Furthermore, advanced generative algorithms can be employed to generate stable material structures [26,60], and LLMs can be used to mine preparation processes from literature, recommending suitable processing solutions. For example, Ceder et al. utilized an automated materials synthesis laboratory to synthesize 41 new materials in just 17 days [61], showcasing the tremendous potential of agents in overcoming barriers between material design and synthesis and in identifying appropriate preparation processes.

4.5. Agent driven materials intelligence platform

LLMs and agent have already demonstrated their potential in the realm of experimental design, and in the future of AI-driven materials intelligence platform, they may serve as the central intelligent hub, seamlessly integrating every facet of the materials research operations (Fig. 8). For instance, LLMs can harness their encoded multidisciplinary knowledge, expert experience, scientific data and theoretical calculation tools to generate novel material designs or structures that meet specified requirements. They also could establish communication with all experimental instruments, enabling real-time monitoring and direct command issuance. Furthermore, they can dynamically adjust the experimental plan based on feedback received during the ongoing experiment and conduct comprehensive analyses of experimental results to make optimal decisions for the subsequent steps.

Furthermore, LLMs holds potential promising in seamlessly integrated with a diverse array of materials science tools, including first-

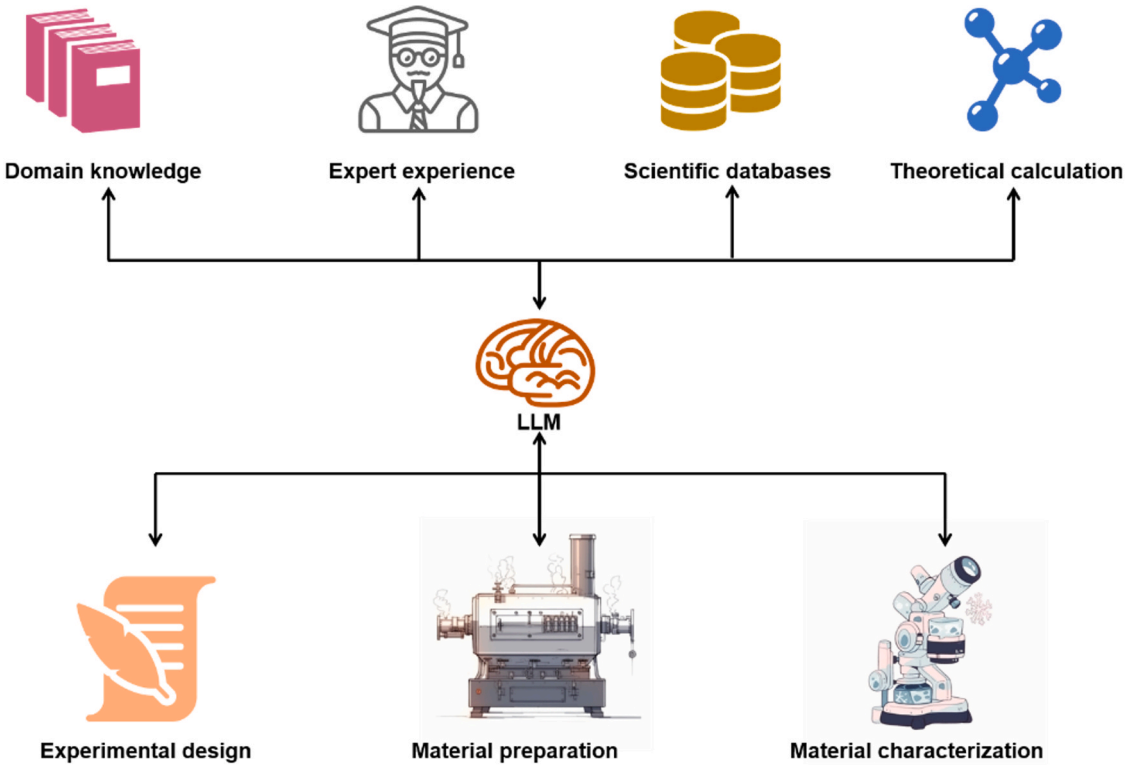

Fig. 8. Schematic of the AI-driven Materials Intelligence Platform.

principles calculation tools, molecular dynamics calculation tools, finite element analysis tools, and even advanced characterization instruments such as scanning electron microscopy and transmission electron microscopy, and others. This integration will empower scientific researchers to bypass the complexities of consulting intricate and opaque tool or instrument manuals. Instead, they will simply need to make reasonable requests to the LLMs that the tool carries in natural language. The tools will automatically fulfill the human needs under the guidance of the LLMs and provide detailed analysis and feedback on the results. Realizing this vision certainly necessitate collaborative efforts among tool developers, instrument manufacturers, material scientists, and artificial intelligence researchers.

5. Current challenges

In the previous discussion, we demonstrated the potential of LLMs to transform the paradigm of materials research. However, this is an emerging field, and to fully realize its advantages, there are still some challenges that need to be addressed. This section will discuss the current limitations of LLMs for materials.

5.1. Limitations of computing resources

The training, fine-tuning, and deployment of LLMs like GPT-4 require substantial computational resources, including high-performance GPUs and extensive data storage facilities [1]. This presents a significant challenge for small research teams and institutions due to their limited access to such infrastructure.

Compared to full-parameter fine-tuning, which can be resource-intensive, parameter-efficient fine-tuning strategies like LoRA offer a more economical alternative [15]. LoRA reduces resource usage and speeds up training by optimizing only a small portion of the model's parameters. However, this approach may require some compromise on the accuracy of downstream tasks. It's worth noting that LoRA is not suitable for all downstream tasks, as its underlying principle is to approximate the high-dimensional representations of large models using low-rank matrices, which can lead to suboptimal performance when dealing with tasks significantly different from the pre-training data distribution [13].

Model quantization (storing model weights at lower floating-point precision) and model distillation are also effective methods for reducing computational requirements, but they come with a trade-off in terms of model accuracy [62]. Another viable option for alleviating computational resource constraints is to leverage commercial LLM APIs, such as those offered by OpenAI, for materials science research. However, uploading sensitive or proprietary data to cloud-based LLMs can raise privacy concerns for research teams and businesses.

5.2. Risk of hallucinations

One of the significant challenges in applying LLMs to materials science is the issue of "hallucinations." Hallucinations occur when LLMs generate harmful content, false information, or otherwise incorrect responses, often due to insufficient training data or ambiguous instructions during human-computer interactions. To mitigate the occurrence of hallucinations, researchers have proposed and validated several strategies: prompt engineering strategy [25], COT strategy [27], RAG strategy [22], model fine-tuning strategy [16], agent strategy [6]. Although these strategies have yielded some progress, the issue of hallucination in LLMs remains a significant challenge. Further research and innovative solutions are required to address this problem to ensure the reliable application of LLMs in materials science research.

5.3. Scarcity of high-quality data

In the preceding text, we discussed how LLMs have shown great potential in alleviating the problem of data scarcity in materials science. However, this potential is contingent upon the LLM being adequately trained with a deep understanding of materials science [2]. While LLMs can perform unsupervised learning from vast amounts of unstructured text, which helps to offset the challenges faced by traditional machine learning methods due to the lack of high-quality labeled data in materials science, obtaining such unstructured text data remains challenging. For example, large-scale collection of published papers often involves complex copyright issues. Furthermore, supervised fine-tuning is crucial for improving the performance of LLMs on downstream materials science tasks, but high-quality domain-specific datasets are scarce. Moreover, the diversity and comprehensiveness of training datasets directly impact the model's ability to generalize across various subfields of materials science and provide accurate insights, therefore, constructing high-quality datasets and designing effective training strategies are key challenges in current research.

5.4. Absence of credibility assessment mechanisms

LLMs operate as black-box models with extremely high potential spatial dimensions and complex internal semantic structures, making it challenging to explain and elucidate their decision-making processes [24]. Although LLMs trained on natural language data can generate self-explanatory outputs, this does not eliminate all risks. Compared to human experts, LLMs lack the inherent ability to evaluate the reliability of generated or retrieved data. Human experts can critically assess information sources and judge the credibility of information, a capability that LLMs do not possess. This deficiency can lead to unverified or low-quality information being disseminated widely, potentially posing risks to critical research applications. Therefore, establishing a credibility evaluation mechanism for LLMs in the field of materials science and developing a comprehensive credibility evaluation framework that can automatically assess and flag potentially unreliable outputs is crucial to mitigating such risks.

5.5. Risks of security

The utilization of agents in materials science introduces additional risks. Unlike LLMs, agents are autonomous and capable of interacting with the physical world [63]. This capability raises concerns that users might intentionally exploit LLM agents to create hazardous situations, such as synthesizing dangerous chemicals. Even in legitimate scientific tasks, agents might inadvertently produce toxic byproducts or unstable compounds. These potential risks underscore the necessity for stringent regulation of agent usage to ensure they are cognizant of the associated dangers. Additionally, it is crucial to align the goals and actions of agents with human values and environmental contexts to prevent unintended harm. As agent technologies rapidly advance in materials science, it is imperative that regulatory measures are rigorously developed and enforced to ensure these technologies are deployed safely and beneficially in research and practical applications.

6. Future direction of LLM for materials

LLM-driven Agent is expected to become independent materials scientists. Compared to human experts, LLMs possess a broader knowledge base, enabling them to develop a more comprehensive understanding of complex problems in materials research and the potential to propose more innovative solutions. In addition, LLM's superior learning ability enables them to rapidly acquire new knowledge and skills beyond those of individual humans. More importantly, LLM's knowledge is stored digitally, which facilitates rapid replication and sharing, thus accelerating knowledge dissemination and innovation. These advantages ensure LLM's core competence as a future materials scientist. To achieve this goal, LLM's development in materials science should focus on the following three points: developing powerful fundamental models

suitable for materials science research, improving the materials reasoning capability of large models, and establishing a perfect ecosystem of LLM materials science tools.

The primary task in applying LLMs to the field of materials science is to equip these models with more cutting-edge knowledge in materials science. While existing general-purpose LLMs excel in natural language processing and general knowledge reasoning, further optimization and customization are needed for them to play a significant role in the highly specialized field of materials science. Specifically, this involves adjustments to model architecture and training strategies: improving the model architecture to better understand and handle specific terms and concepts in materials science, and adopting more targeted training strategies that enable the model to learn professional knowledge in materials science more effectively. Expansion of training data is also crucial, requiring the collection and integration of more materials science-related literature, experimental data, and patent information to enrich the training dataset, ensuring that the model is exposed to the latest research findings and technological advancements, thus enhancing its understanding of professional knowledge in the field of materials science. Through continuous iteration and optimization, future LLMs will demonstrate more accurate and efficient performance in materials science. These optimization measures not only improve the performance of LLMs in materials science but also promote progress in materials science, providing strong support for the development of new materials.

Improving the material reasoning capabilities of LLMs is another important direction for development. Material reasoning ability is critical for materials research, directly impacting the efficiency and success rate of research. Taking the discovery of new materials as an example, this often involves out-of-distribution generalization problems, which pose significant challenges to data-driven traditional machine learning methods. In the past, the discovery of new materials by human largely relied on knowledge-driven methods, meaning the design of materials was guided by the induction and analysis of scientific laws. Therefore, to fundamentally address the issues of scarce data and out-of-distribution generalization in the discovery of new materials in materials research, LLMs need to shift from shallow reasoning to deep reasoning. Existing strategies such as COT and its variants, Monte Carlo Tree Searc, and Graph RAG have shown initial effectiveness, but there is still a long way to go before achieving deep reasoning in LLMs. Therefore, more methods need to be developed to enhance the material reasoning capabilities of LLMs.

Finally, the capability of LLM-driven materials scientists is not only determined by the strength of the LLM itself but is also limited by the availability of tools. Therefore, there is an urgent need to establish a complete ecosystem of LLM materials science tools. Firstly, developing more specialized tools dedicated to the field of materials science will further enhance the application value of LLMs. These proprietary tools include, but are not limited to, materials simulation software, high-throughput computing platforms, automated laboratories, and materials design tools. Additionally, establishing standards for LLM materials-specific tools is also a pressing need. Standardizing can ensure compatibility and interoperability between different tools, reducing the complexity of system integration. Standardization can also promote the standardization of tool development, improving the quality and reliability of tools, accelerating the promotion and application of technology, and providing researchers with a unified and efficient working platform, further promoting innovation and development in the field of materials science. By establishing these standards, LLMs will be able to utilize specialized tools more effectively, achieving higher research efficiency and outcomes.

Funding

This work is financially supported by National Key R&D Program of China (2022YFB3807700), National Natural Science Foundation of China, NSFC (22133005, 21973107, and 11804351), the Program of Shanghai Academic Research Leader (20XD1404100), Project funded by China Postdoctoral Science Foundation (2022M723276) , Sponsored by Shanghai Sailing Program (23YF1454900), and Shanghai Post-doctoral Excellence Program (2022660).

Author contributions

Jianjun Liu designed the project. Songlin Yu and Nian Ran wrote the manuscript.

CRediT authorship contribution statement

Nian Ran: Investigation. **Songlin Yu:** Writing – review & editing, Writing – original draft, Validation, Methodology. **Jianjun Liu:** Supervision.

Declaration of Competing Interest

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

References

[1] OpenAI. GPT-4 Technical Report (OpenAI, *2023*).
[2] M.J. Buehler, MechGPT, A language-based strategy for mechanics and materials modeling that connects knowledge across scales, disciplines, and modalities, Appl. Mech. Rev. 76 (2024), https://doi.org/10.1115/1.4063843.
[3] Miret, S. & Anoop Krishnan, N.M. Are LLMs Ready for Real-World Materials Discovery? arXiv e-prints, arXiv:2402.05200-arXiv:02402.05200 (2024).
[4] S. Noy, W. Zhang, Experimental evidence on the productivity effects of generative artificial intelligence, Science 381 (2023) 187–192, https://doi.org/10.1126/science.adh2586.
[5] J. Choi, B. Lee, Accelerating materials language processing with large language models, Commun. Mater. 5 (2024), https://doi.org/10.1038/s43246-024-00449-9.
[6] Prince, M.H. *et al.* Opportunities for Retrieval and Tool Augmented Large Language Models in Scientific Facilities. arXiv e-prints, arXiv:2312.01291-arXiv:02312.01291 (2023).
[7] Zhao, W.X. *et al.* A Survey of Large Language Models. arXiv e-prints, arXiv:2303.18223-arXiv:12303.18223 (2023).
[8] Saad-Falcon, J. *et al.* PDFTriage: Question Answering over Long, Structured Documents. arXiv e-prints, arXiv:2309.08872-arXiv:02309.08872 (2023).
[9] N.S. Lai, et al., Artificial intelligence (AI) workflow for catalyst design and optimization, Ind. Eng. Chem. Res. 62 (2023) 17835–17848, https://doi.org/10.1021/acs.iecr.3c02520.
[10] Zhang, Y. *et al.* Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv e-prints, arXiv:2309.01219-arXiv:02309.01219 (2023).
[11] Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. arXiv e-prints, arXiv:2302.13971-arXiv:12302.13971 (2023).
[12] Zeng, A. *et al.* GLM-130B: An Open Bilingual Pre-trained Model. arXiv e-prints, arXiv:2210.02414-arXiv:02210.02414 (2022).
[13] Sun, X., Ji, Y., Ma, B. & Li, X. A Comparative Study between Full-Parameter and LoRA-based Fine-Tuning on Chinese Instruction Data for Instruction Following Large Language Model. arXiv e-prints, arXiv:2304.08109-arXiv:02304.08109 (2023).
[14] Z. Han, C. Gao, J. Liu, J. Zhang, S. Qian ZhangParameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. arXiv e-prints, arXiv:2403.14608-arXiv:12403.14608 (2024).
[15] Hu, E.J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. arXiv e-prints, arXiv:2106.09685-arXiv:02106.09685 (2021).
[16] Xie, T. *et al.* DARWIN Series: Domain Specific Large Language Models for Natural Science. arXiv e-prints, arXiv:2308.13565-arXiv:12308.13565 (2023).
[17] T. Gupta, M. Zaki, N.M.A. Krishnan, Mausam, MatSciBERT: a materials domain language model for text mining and information extraction, NPJ Comput. Mater. 8 (2022), https://doi.org/10.1038/s41524-022-00784-w.
[18] Zhao, P. *et al.* Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv e-prints, arXiv:2402.19473-arXiv:12402.19473 (2024).
[19] Lei, G., Docherty, R. & Cooper, S.J. Materials science in the era of large language models: a perspective. arXiv e-prints, arXiv:2403.06949-arXiv:02403.06949 (2024).
[20] Lewis, P. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv e-prints, arXiv:2005.11401-arXiv:12005.11401 (2020).
[21] Chiang, Y., Chou, C.-H. & Riebesell, J. LLaMP: Large Language Model Made Powerful for High-fidelity Materials Knowledge Retrieval and Distillation. arXiv e-prints, arXiv:2401.17244-arXiv:12401.17244 (2024).
[22] Edge, D. *et al.* From Local to Global: A Graph RAG Approach to Query-Focused Summarization. arXiv e-prints, arXiv:2404.16130-arXiv:12404.16130 (2024).

[23] Buehler, M.J. Generative retrieval-augmented ontologic graph and multi-agent strategies for interpretive large language model-based materials design. arXiv e-prints, arXiv:2310.19998-arXiv:12310.19998 (2023).
[24] Y. Liu, et al., Generative artificial intelligence and its applications in materials science: current situation and future perspectives, J. Mater. 9 (2023) 798–816, https://doi.org/10.1016/j.jmat.2023.05.001.
[25] Feldman, P., Foulds, J.R. & Pan, S. Trapping LLM Hallucinations Using Tagged Context Prompts. arXiv e-prints, arXiv:2306.06085-arXiv:02306.06085 (2023).
[26] Flam-Shepherd, D. & Aspuru-Guzik, A. Language models can generate molecules, materials, and protein binding sites directly in three dimensions as XYZ, CIF, and PDB files. arXiv e-prints, arXiv:2305.05708-arXiv:02305.05708 (2023).
[27] Wei, J. *et al.* Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv e-prints, arXiv:2201.11903-arXiv:12201.11903 (2022).
[28] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. arXiv e-prints, arXiv:2205.11916-arXiv:12205.11916 (2022).
[29] D. Lee, W.W. Chen, L. Wang, Y.C. Chan, W. Chen, Data-driven design for metamaterials and multiscale systems: a review, Adv. Mater. 36 (2024) e2305254, https://doi.org/10.1002/adma.202305254.
[30] S. Yu, et al., Studying complex evolution of hyperelastic materials under external field stimuli using artificial neural networks with spatiotemporal features in a small-scale dataset, Adv. Mater. 34 (2022) e2200908, https://doi.org/10.1002/adma.202200908.
[31] Z. Li, et al., Interpreting chemisorption strength with AutoML-based feature deletion experiments, Proc. Natl. Acad. Sci. USA 121 (2024) e2320232121, https://doi.org/10.1073/pnas.2320232121.
[32] J.A. Esterhuizen, B.R. Goldsmith, S. Linic, Uncovering electronic and geometric descriptors of chemical activity for metal alloys and oxides using unsupervised machine learning, Chem. Catal. 1 (2021) 923–940, https://doi.org/10.1016/j.checat.2021.07.014.
[33] S. Liu, T. Wen, A.S.L. Subrahmanyam Pattamatta, D.J. SrolovitzA Prompt-Engineered Large Language Model, Deep Learning Workflow for Materials Classification. arXiv e-prints, arXiv:2401.17788-arXiv:12401.17788 (2024).
[34] Yang, S.J. *et al.* Accurate Prediction of Experimental Band Gaps from Large Language Model-Based Data Extraction. arXiv e-prints, arXiv:2311.13778-arXiv:12311.13778 (2023).
[35] Z. Xiao, et al., Generative artificial intelligence GPT-4 accelerates knowledge mining and machine learning for synthetic biology, ACS Synth. Biol. 12 (2023) 2973–2982, https://doi.org/10.1021/acssynbio.3c00310.
[36] Lee, W., Kang, Y., Bae, T. & Kim, J. Harnessing Large Language Model to collect and analyze Metal-organic framework property dataset. arXiv e-prints, arXiv:2404.13053-arXiv:12404.13053 (2024).
[37] M.P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, Nat. Commun. 15 (2024) 1569, https://doi.org/10.1038/s41467-024-45914-8.
[38] S. Li, et al., Extracting the synthetic route of Pd-based catalysts in methanol steam reforming from the scientific literature, J. Chem. Inf. Model 63 (2023) 6249–6260, https://doi.org/10.1021/acs.jcim.3c01442.
[39] Khalighinejad, G., Circi, D., Brinson, L.C. & Dhingra, B. Extracting Polymer Nanocomposite Samples from Full-Length Documents. arXiv e-prints, arXiv:2403.00260-arXiv:02403.00260 (2024).
[40] V. Tshitoyan, et al., Unsupervised word embeddings capture latent knowledge from materials science literature, Nature 571 (2019) 95–98, https://doi.org/10.1038/s41586-019-1335-8.
[41] J. Qu, et al., Leveraging language representation for materials exploration and discovery, NPJ Comput. Mater. 10 (2024), https://doi.org/10.1038/s41524-024-01231-8.
[42] Niyongabo Rubungo, A., Arnold, C., Rand, B.P. & Bousso Dieng, A. LLM-Prop: Predicting Physical And Electronic Properties Of Crystalline Solids From Their Text Descriptions. arXiv e-prints, arXiv:2310.14029-arXiv:12310.14029 (2023).
[43] Huang, H., Magar, R., Xu, C. & Barati Farimani, A. Materials Informatics Transformer: A Language Model for Interpretable Materials Properties Prediction. arXiv e-prints, arXiv:2308.16259-arXiv:12308.16259 (2023).
[44] Yin, J., Bose, A., Cong, G., Lyngaas, I. & Anthony, Q. Comparative Study of Large Language Model Architectures on Frontier. arXiv e-prints, arXiv:2402.00691-arXiv:02402.00691 (2024).
[45] Caldas Ramos, M., Michtavy, S.S., Porosoff, M.D. & White, A.D. Bayesian Optimization of Catalysts With In-context Learning. arXiv e-prints, arXiv:2304.05341-arXiv:02304.05341 (2023).
[46] Kristiadi, A. *et al.* A Sober Look at LLMs for Material Discovery: Are They Actually Good for Bayesian Optimization Over Molecules? arXiv e-prints, arXiv:2402.05015-arXiv:02402.05015 (2024).
[47] H. Chen, I.C. Covert, S.M. Lundberg, S.-I. Lee, Algorithms to estimate Shapley value feature attributions, Nat. Mach. Intell. 5 (2023) 590–601, https://doi.org/10.1038/s42256-023-00657-x.
[48] H. Zhao, et al., Explainability for large language models: a survey, ACM Trans. Intell. Syst. Technol. 15 (2024) 1–38, https://doi.org/10.1145/3639372.
[49] T. Xie, et al., Opinion mining by convolutional neural networks for maximizing discoverability of nanomaterials, J. Chem. Inf. Model 64 (2024) 2746–2759, https://doi.org/10.1021/acs.jcim.3c00746.
[50] Xi, Z. *et al.* The Rise and Potential of Large Language Model Based Agents: A Survey. arXiv e-prints, arXiv:2309.07864-arXiv:02309.07864 (2023).
[51] Zhou, W. *et al.* Agents: An Open-source Framework for Autonomous Language Agents. arXiv e-prints, arXiv:2309.07870-arXiv:02309.07870 (2023).
[52] Zhang, Z. *et al.* A Survey on the Memory Mechanism of Large Language Model based Agents. arXiv e-prints, arXiv:2404.13501-arXiv:12404.13501 (2024).
[53] Liu, H. & Li, L. On Languaging a Simulation Engine. arXiv e-prints, arXiv:2402.16482-arXiv:12402.16482 (2024).
[54] Liu, Y., Checa, M. & Vasudevan, R.K. Synergizing Human Expertise and AI Efficiency with Language Model for Microscopy Operation and Automated Experiment Design. arXiv e-prints, arXiv:2401.13803-arXiv:12401.13803 (2024).
[55] A. Leitherer, B.C. Yeo, C.H. Liebscher, L.M. Ghiringhelli, Automatic identification of crystal structures and interfaces via artificial-intelligence-based electron microscopy, NPJ Comput. Mater. 9 (2023), https://doi.org/10.1038/s41524-023-01133-1.
[56] Z. Zou, et al., A deep learning model for predicting selected organic molecular spectra, Nat. Comput. Sci. 3 (2023) 957–964, https://doi.org/10.1038/s43588-023-00550-y.
[57] Q. Zhu, et al., Automated synthesis of oxygen-producing catalysts from Martian meteorites by a robotic AI chemist, Nat. Synth. (2023), https://doi.org/10.1038/s44160-023-00424-1.
[58] H. Zhao, et al., A robotic platform for the synthesis of colloidal nanocrystals, Nat. Synth. 2 (2023) 505–514, https://doi.org/10.1038/s44160-023-00250-5.
[59] Z. Rao, et al., Machine learning-enabled high-entropy alloy discovery, Science 378 (2022) 78–85, https://doi.org/10.1126/science.abo4940.
[60] A. Merchant, et al., Scaling deep learning for materials discovery, Nature 624 (2023) 80–85, https://doi.org/10.1038/s41586-023-06735-9.
[61] N.J. Szymanski, et al., An autonomous laboratory for the accelerated synthesis of novel materials, Nature 624 (2023) 86–91, https://doi.org/10.1038/s41586-023-06734-w.
[62] Wang, H. *et al.* BitNet: Scaling 1-bit Transformers for Large Language Models. arXiv e-prints, arXiv:2310.11453-arXiv:12310.11453 (2023).
[63] Tang, X. *et al.* Prioritizing Safeguarding Over Autonomy: Risks of LLM Agents for Science. arXiv e-prints, arXiv:2402.04247-arXiv:02402.04247 (2024).