Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d6sc01168c

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 10th February 2026
Accepted 17th April 2026

DOI: 10.1039/d6sc01168c

rsc.li/chemical-science

A dual-mode large language model assistant for on-surface reactions *via* fine-tuning and retrieval-augmented generation

Juan Xiang,[a] Qi Huang,[a] Xinyi Zhang,[a] Tairan Yang,[a] Zhiwen Zhu,[a] Chanyu Li,[b] Liangliang Cai[a] and Qiang Sun *[ab]

Surface reactions underpin catalysis, nanomaterials, energy conversion, and molecular-scale fabrication, yet the field suffers from fragmented knowledge dispersed across unstructured literature, hindering systematic analysis and data-driven discovery. Existing chemical databases and language models inadequately capture the domain-specific semantics and experimental parameters unique to on-surface reactions. Here, we present an integrated framework that transforms the dispersed surface-chemistry literature into a structured, machine-readable knowledge and leverages it to develop a domain-specialized large language model (LLM) assistant for on-surface reactions. We curated and semantically screened hundreds of thousands of publications to construct the surface-chemistry corpus, from which we extracted 44 predefined reaction attributes across more than 44 000 studies of surface reactions. These structured records were used to build both a high-quality reaction database and a domain-specific question–answering dataset. On this basis, we developed a dual-mode LLM system that combines a parameter-efficient fine-tuned reasoning model with a dual-source retrieval-augmented generation (RAG) framework, enabling both deep inference and verifiable retrieval of experimental parameters. Evaluations demonstrate that the fine-tuned LLM outperforms existing chemistry-oriented language models on surface-chemistry question–answering, achieving a Bert-*F*1 score exceeding 0.8. Incorporation of the RAG framework further improves factual accuracy, completeness, and reasoning consistency by grounding responses in the retrieved literature and structured reaction data. Latent-space analyses reveal that domain-specific fine-tuning reorganizes internal representations toward task-oriented coherence. This work establishes a scalable pathway for converting fragmented surface-chemistry knowledge into an intelligent platform, paving the way toward data-driven prediction, experimental planning and automated reasoning in on-surface reactions.

Introduction

Surface chemistry plays a foundational role across catalysis, nanomaterials, energy conversion, and molecular-scale fabrication, as it governs how atoms and molecules interact, transform, and assemble at interfaces.[1–3] Advances in experimental techniques, such as scanning probe microscopy and surface-sensitive spectroscopies have enabled increasingly precise characterization of molecular adsorptions, reaction pathways, and structure–property relationships on solid surfaces.[4–8] Despite its central importance, the field remains heavily fragmented: critical experimental details, mechanistic insights, and structure–reactivity trends are dispersed across decades of heterogeneous literature, embedded in unstructured text, and reported using inconsistent terminology. This lack of systematic, machine-readable surface chemistry knowledge has created a pronounced bottleneck, limiting both cumulative scientific understanding and the development of data-driven approaches that could accelerate discovery and enable predictive modeling of relevant surface reactions.

Although AI-driven research paradigms have advanced rapidly in chemistry and materials science,[9–19] their effectiveness remains limited in highly specialized domains such as surface chemistry, where both structured data availability and domain-specific semantic representation remain insufficient. Existing chemical and materials databases are primarily designed for solution-phase chemistry or bulk materials and therefore do not systematically capture the experimental complexity of on-surface reactions.[20–23] Key variables, such as precursor identity, substrate crystallography, activation protocol, and surface coverage, are often dispersed across different sections of the primary literature in a heterogeneous and unstructured form, making reliable extraction, retrieval, and comparative analysis intrinsically difficult. At the same

[a]*Materials Genome Institute, Shanghai University, 200444 Shanghai, China. E-mail: qiangsun@shu.edu.cn*

[b]*Qianweichang College, Shanghai University, 200444 Shanghai, China*

time, recent language models and domain-adapted foundation models have demonstrated strong capabilities in adjacent areas.[24–29] For example, SciBERT and ChemBERT have improved scientific text understanding and chemical language modeling,[30,31] while CrystaLLM and MOFTransformer have shown the potential of domain-specialized architectures for crystal structure modeling and materials property prediction.[32,33] Chemistry-oriented assistants such as Chemma and ChemDFM further illustrate the growing ability of large language models to support synthesis planning and chemical reasoning.[34,35] However, these models are generally developed for broader scientific language understanding, molecular property prediction, bulk-material modeling, or conventional solution-phase chemistry, and thus do not adequately reflect the distinctive experimental logic and knowledge organization of surface chemistry and on-surface reactions.[36–43] As a result, the continued lack of sufficiently structured and domain-specific data remains a central obstacle to the development of robust machine-learning frameworks for reaction prediction and experimental condition optimization in surface chemistry. Indeed, our own preliminary attempt to apply LLMs to automated literature mining in the field of on-surface reactions was limited to fewer than 70 publications, underscoring the difficulty of constructing a scalable and statistically meaningful database under existing data.[44]

To address these limitations, we established an integrated framework that combines large-scale literature curation, structured data construction, and domain-specialized language modeling for surface chemistry and on-surface reactions. Using a multi-stage semantic classification pipeline, we systematically filtered and organized hundreds of thousands of publications to construct a large-scale literature corpus for this field. Building on this corpus, we extracted 44 predefined reaction attributes from more than 44 000 publications to create a structured on-surface reaction database, which was subsequently used to generate a high-quality, domain-specific question–answering dataset covering surface chemistry concepts, synthesis conditions, and mechanistic reasoning. Leveraging these resources, we developed a dual-mode LLM assistant consisting of a fine-tuned reasoning module for mechanistic inference and a dual-source retrieval-augmented generation framework for real-time, verifiable retrieval of experimental parameters. Together, these advances provide a structured and intelligent platform for organizing fragmented surface-chemistry knowledge and support future developments in reaction prediction, condition optimization, and data-driven discovery.

Results and discussion

The initial dataset was compiled from the PubMed database, Web of Science, and the literature repository accumulated by our research group. As shown in Fig. 1a, for the PubMed database, we employed a keyword-based retrieval strategy. From over 39 million publications, we exported nearly 300 000 literature records containing metadata such as title, abstract, DOI, and author information. To overcome the high costs of data processing and uncertainties associated with raw, unstructured data, we used the fine-tuned Text-to-Text Transfer Transformer (T5) model[45] (more details in Section 1.1 of the SI) for completing the structured processing of 291 566 metadata entries. For the WOS dataset, a total of 74 934 samples containing metadata were retrieved using keywords and were stored in a tabular format. The repository from our research group contributed additional related literature. These diverse sources collectively established an initial corpus that supports downstream data mining and model training.

Then, we designed a multi-stage semantic screening process based on the literature metadata to determine their relevance. The first stage of the screening process evaluates whether the semantic content of a literature title and abstract falls within the broader domain of surface chemistry. It should be noted that surface chemistry encompasses the study of physical and chemical phenomena occurring at different interfaces. Core topics include, but are not limited to, surface adsorption, desorption, catalytic reactions, surface reconstruction, defects, nucleation and growth, as well as characterization using typical surface sensitive techniques such as scanning probe microscopy (SPM) and X-ray photoelectron spectroscopy (XPS). Subsequently, a second stage of screening is implemented to determine whether the literature is highly relevant to the core theme of on-surface reactions. This second stage primarily focuses on the "top-down" approach to controlling on-surface reactions at the atomic or molecular scale for synthesizing new substances or functional structures. Examples include work on metal or semiconductor surfaces involving the active and controllable construction of new molecular structures and nanomaterials *via* different activation methods, as illustrated in the right panel of Fig. 1b. We fine-tuned SciBERT models for the aforementioned two-step classification task. Through a precise semantic classification, a robust set of literature closely aligned with the research objectives was ultimately obtained (more details in Section 1.2 of the SI).

Following the aforementioned filtering and collection processes, we obtained a reasonable quantity of corpus, specifically 34 906 publications related to surface chemistry and 9246 publications related to on-surface reactions. For the results filtered in the second stage, we conducted further extraction processing. As shown in Fig. 1b, we developed a top-down extraction framework, systematically enumerated 44 potential attributes of on-surface synthesis, and classified into three major categories: Precursors, Reaction Stages, and Final Stages (Fig. S2 in the SI). The Precursors section includes basic information about the precursors, such as IUPAC name, abbreviation, and morphology, along with details on the molecular deposition methods and parameters, and substrate. The Reaction Stages and Final Stages record the characteristics of the intermediates and final products, respectively, including abbreviation, morphology, and coverage, as well as the activation methods (*e.g.*, thermal, light, tip induced, *etc.*). We also recorded the type of reaction. Furthermore, to uniquely identify each publication, a Literature section was established to store the publication metadata. To maximize the accuracy and coverage of the extraction results, we developed a dedicated

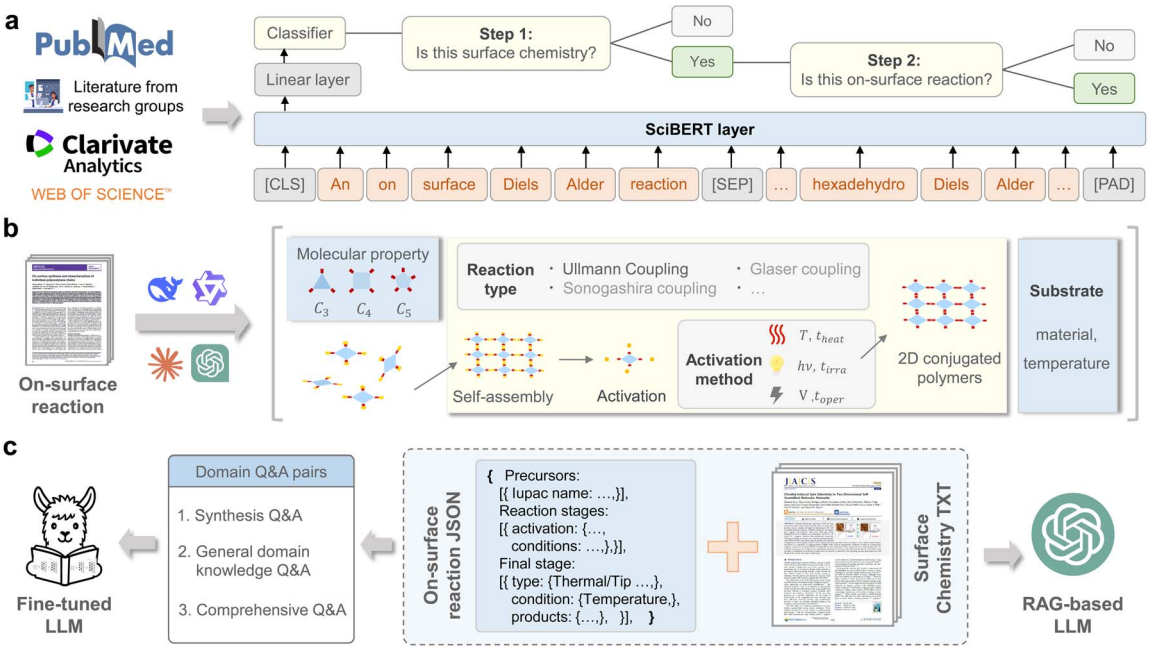

Fig. 1 Workflow for constructing the dual-mode LLM for on-surface reactions. (a) Literature data are processed by a two-stage classification executed by a fine-tuned SciBERT. (b) Literature of on-surface reactions is parsed and converted into a structured JSON file, including information of precursor, reaction type, activation method and substrate. (c) Structured records are integrated with the surface chemistry text database to construct a high-quality corpus for the RAG-based LLM. In parallel, question–answer (Q&A) pairs are generated from these structured records and the text database to fine-tune the LLM.

annotation web interface (Fig. S4 in the SI) that enabled at least five domain experts to annotate the corresponding data within the full-text articles, with results saved in JSON format. We have completed 170 full-text articles with human annotations. We utilized these annotated benchmarks to evaluate mainstream LLMs, including Claude-4, GPT-4.1, Qwen-Plus, and DeepSeek. A carefully crafted prompt was employed to guide the LLMs in extracting on-surface reaction attributes from complete full-text articles. Using full-text documents was essential because descriptions of reaction conditions are often dispersed throughout the literature; for example, parameters for multi-step reactions may be distributed across multiple paragraphs, and the IUPAC name of a molecule typically appears only once, often in the Methods or the Results and discussion sections.

To mitigate the hallucination phenomenon in LLM-based data/information extraction, we designed a prompt with a five-component structure, consisting of Role, Execution Rules, Output Formatting, Reference, and Stress sections (Fig. 2a). The first three components served as system prompts, setting the model's role, guiding it to extract strictly factual values from the original text, and constraining the output to JSON format. The latter two components consisted of an annotated JSON template (designed to enhance the model's understanding of each field) and the full text, with an explicit final instruction emphasizing strict adherence to the prescribed JSON format. The evaluations of extraction results for all models are presented in Fig. 2. For structured fields, such as IUPAC names, the *F*-score criterion was applied (Fig. 2b and c). In contrast, for semantic fields such as morphology of precursor, BertScore-based semantic similarity was introduced as a complementary evaluation metric (Fig. 2d and e). The detailed formulations of the *F*-score evaluation criteria and BertScore-based evaluation criteria are described in Section 2 of the SI. Overall, all models exhibited comparable performance in terms of the *F*1-score, with DeepSeek, Qwen, and Claude achieving scores of 0.71, 0.72, and 0.72, respectively, while GPT performed slightly lower at 0.70. Although overall performance converged, significant differences persisted in specific categories and metrics, primarily stemming from each model's capability to process long-context documents and deeply understand complex, free-form natural language.

At the attribute level, Claude-4 showed the best semantic performance in precursor-related fields. Its *F*1 score in the precursors category was comparable to those of the other models ($R = 0.72$, $P = 0.72$, and $F1 = 0.71$), and it achieved the highest BertScore values, with Bert-recall, Bert-precision, and Bert-*F*1 all reaching 0.74. This pattern suggests greater robustness in handling complex chemical nomenclature and physical-state descriptions. Final stage extraction was the most challenging task overall, with *F*1 scores ranging from 0.61 to 0.67 across all models, reflecting the implicit nature of final product information description in the literature. Additional results for reaction stages and final stages are provided in Fig. S3 (see Section 3 of the SI). It should also be noted that, despite 3–5 iterative rounds of annotation refinement and criterion alignment, some degree of annotation noise was unavoidable. Given the inherent ambiguity of multistage reactions, the intrinsic hallucination risk of LLMs, and the use of zero-shot inference

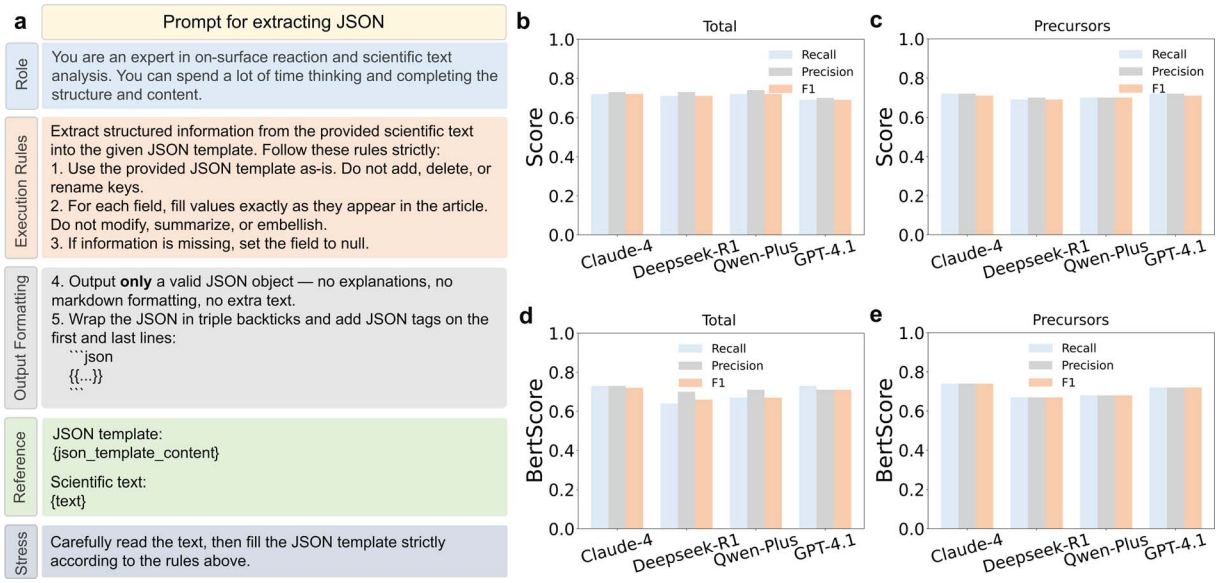

Fig. 2 The prompt framework designed for extracting structured information from the on-surface reaction literature, with comparative performance evaluation across different LLMs. (a) Four key components of the prompt: role definition, execution rules, output format, and template, complemented by an instruction emphasizing strict adherence to the rules. (b and c) Recall, precision, and *F*1-score performance of Claude-4, DeepSeek-R1, Qwen-Plus, and GPT-4.1 across the total dataset and precursor-related extraction tasks. (d and e) Bert-recall, Bert-precision, and Bert-*F*1 score across total and precursor extraction tasks.

on long texts averaging approximately 15 000 tokens, an overall extraction performance exceeding 70% indicates robust performance.

To better understand the sources of error underlying these overall scores, we further analyzed the outputs of all four models at the level of individual reaction attributes. The errors were not uniformly distributed across the schema, but were concentrated in a limited set of fine-grained fields. The most error-prone categories were final product abbreviation (average exact match, 0.54) and intermediate abbreviation (0.58), whereas more stable categories included precursor substrate material (0.76), reaction-stage activation type (0.97), and final-stage type (0.92). A closer inspection revealed four representative error modes. The first was over-segmentation, in which the model inferred unsupported intermediate stages from descriptive passages. For example, Qwen-Plus generated three reaction stages with intermediates "I", "II", and "III",[46] whereas the annotation by human contained no reaction stages and recorded the transformation only at the final stage. This discrepancy arises because these intermediates are computationally derived and were not empirically observed during experimental characterization. The second was information omission. In one case, Claude-4 correctly captured the overall transformation but reduced a richer annotated outcome to a simplified record containing only one explicit final product abbreviation ("1"), whereas the annotation by human preserved multiple products, including "7-AGNRs, 1".[47] The third type of error was mixing of reaction conditions, in which GPT-4.1 merged two distinct thermal processes into a single stage with "433 K and 523 K" and compressed multiple final products into a single string, "D1, D2, D3".[48] In the annotation, these thermal events and products are represented with finer stage-level resolution. The fourth type of error was schema-label mismatch, DeepSeek correctly recovered the intermediate "poly-1" and the annealing temperature of 200 °C, but labeled the intermediate-stage reaction type as "radical step-growth polymerization", whereas the annotation maps this step to "Ullmann coupling".[49] Taken together, Claude-4 exhibited the least severe mismatches and the most balanced performance across extraction attributes. It was therefore selected for extraction on the remaining literature of on-surface reactions, achieving both an overall *F*1 score of 0.72 (Fig. 2b) and a Bert-*F*1 score of 0.72 (Fig. 2d). Importantly, downstream Q&A generation in our workflow was not conditioned solely on the extracted JSON, but jointly on the source article text and the matched structured extraction.

Leveraging the high-quality specialized corpus constructed as described above, we proceed to develop an intelligent Q&A (question and answer) system specifically designed for the field of on-surface reactions. The system supports both general surface chemistry knowledge queries and process-level questions related to on-surface reactions, including activation methods, deposition temperatures, and substrates. The training data for the fine-tuned LLM primarily comprise three Q&A categories, namely Synthesis Q&A, General Domain Knowledge Q&A, and Comprehensive Q&A, which are derived from on-surface reaction JSON files and surface chemistry literature. The Synthesis Q&A category focuses on process-oriented knowledge of on-surface reactions, explicitly specifying molecular precursors using IUPAC names, substrates and crystallographic orientations, deposition conditions, activation methods, intermediate structures, and final products, with an

emphasis on reaction pathways and mechanistic interpretation. The General Domain Knowledge Q&A category addresses general and conceptual aspects of surface chemistry. The Comprehensive Q&A category emphasizes integrative and comparative reasoning across different on surface reaction contexts. The fine-tuned LLM employs a parameter-efficient fine-tuning strategy based on Low-Rank Adaptation (LoRA) (Fig. 3), in which trainable low-rank update matrices are introduced into the attention projection layers (Q/K/V) while the pretrained backbone weights remain frozen. Compared with full-parameter fine-tuning, LoRA is more computationally efficient and less prone to overfitting in moderate-scale domain datasets, whereas compared with prompt-only adaptation it allows chemistry-specific behaviors to be more stably encoded into model parameters. The model is trained on the Q&A dataset constructed from the three categories described above, enabling efficient adaptation of the LLaMA-3.1-8B[50] model without modifying most of its parameters (detailed in Sections 1.3 and 1.4 of the SI). This process enables the model to capture complex on-surface reaction knowledge, allowing for inference and responses to specialized questions concerning specific substrates, reaction types, and activation conditions. For example, when asked about the "reaction pathway of the TIPB molecule on the Ag(111) surface", the model simulates expert reasoning by first analyzing the catalytic effect of the Ag(111) substrate on C–I bond cleavage and the formation of radical intermediates, followed by inferring the pathway of Ullmann coupling to deliver a complete, logically chained answer. In addition to the finetuned LLM described above, the Q&A system also incorporates an online RAG based LLM built upon the comprehensive JSON and literature corpus.

In addition to the fine-tuned LLM described above, the Q&A system also incorporates an online RAG based LLM built upon the comprehensive literature corpus. The RAG-based LLM adopts an advanced Retrieval Augmented Generation framework (Fig. 3b), which is specifically designed to mitigate knowledge lag and reduce hallucination effects when processing surface chemistry related information. A more detailed description of the RAG architecture, retrieval pipeline, and implementation is provided in Section 6 of the SI. RAG-based LLM integrates a dual-source external knowledge base, including the surface chemistry (TXT format) and the specially extracted structured reaction conditions database (JSON format), to support Synthesis Q&A and General Domain Knowledge Q&A, respectively. Specifically, a user query is first encoded and then used for similarity search within a vector

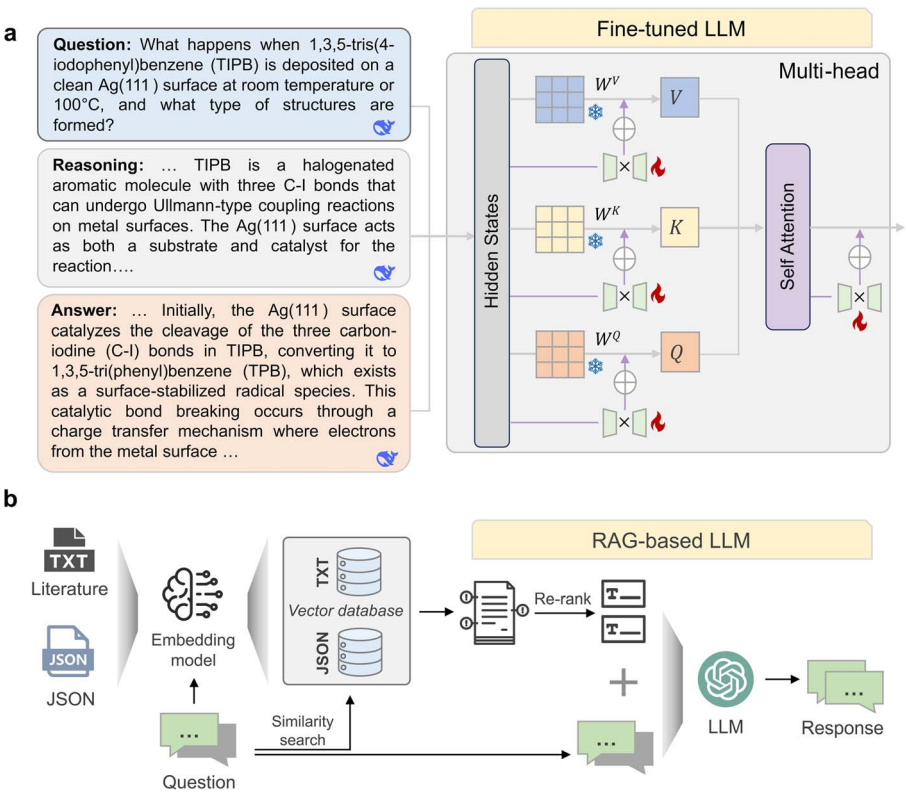

Fig. 3 Dual-mode LLM. (a) The fine-tuned LLM is built on a multi-head self-attention mechanism within the transformer architecture and optimized using a parameter-efficient fine-tuning strategy (LoRA). In this framework, the original W^V, W^K and W^Q weights (represented by blue snowflakes) remain frozen, while low-rank adaptation modules (represented by green trapezoidal blocks and red spark signs) are introduced to perform low-rank updates. The training inputs consist of the Question, Reasoning, and Answer datasets, while the fine-tuned LLM outputs the Reasoning and Answer. (b) RAG-based LLM utilizes an embedding model to encode data (TXT/JSON) and store it in a vector database. When a user proposes a question, it performs a similarity search to retrieve relevant documents, which are then re-ranked and provided to the LLM to generate the response.

database built by the all-MiniLM-L6-v2 embedding model. This process allows for the retrieval of the most relevant context from two data sources: the text literature corpus provides chemistry and physics descriptions, research backgrounds, and mechanistic explanations from a microscopic perspective to support reasoning and knowledge generalization, while the JSON structured data offers precise parameters, such as specific substrate orientation, activation temperatures, precursor molecules, and reaction types, ensuring the numerical or categorical accuracy of synthesis parameters. The retrieved text segments then undergo a re-ranking step to optimize relevance and reduce redundancy before being input alongside the input question into the LLM to generate the final response. This multi-source RAG framework enables the RAG-based LLM to function as a specialized intelligent tool with improved verifiability and factual reliability. Together, the finetuned-LLM and the RAG-based LLM form a dual-mode intelligent system that supports deep inference and real-time retrieval, thereby enhancing experimental decision efficiency and knowledge discovery in studies of on-surface reactions.

We designed a set of prompts for generating the Q&A, as detailed in Section 4 of the SI, and subsequently applied them across current mainstream large language models for comparative evaluation. Fig. 4a presents the performance evaluation of four widely used LLMs, namely Qwen-Plus, DeepSeek-R1, Claude-4, and GPT-4.1, on the task of Q&A pair generation. The assessment considered five key dimensions:[51] relevance, measuring the fit between the generated Q&A and the source text; agnosticism, which evaluates the degree of context independence by requiring that the Q&A does not reference figures or tables from the source text; accuracy, measuring the factual correctness of the Q&A regarding surface chemistry knowledge; completeness, measuring the comprehensiveness of the information provided in the Q&A and reasonableness, measuring the internal logical coherence of the generated answer and evaluating whether it contains contradictions. To ensure professional rigor and impartiality in the evaluation, we engaged human experts to manually score the generated Q&A sets (more details in Section 5 of the SI). As shown in Fig. 4a, although all models achieved near-maximum scores on the Relevance metric, substantial differences emerged in metrics capturing deeper semantic. Specifically, GPT-4.1, Claude-4 and DeepSeek-R1 demonstrated clear advantages in Accuracy and Completeness, with scores approaching 80%. On the Agnosticism metric, DeepSeek-R1 (approximately 68%) and Claude-4 (approximately 70%) outperformed the other models, indicating stronger capability in maintaining contextual independence. Given that DeepSeek-R1 achieved a well-balanced performance across Agnosticism, Accuracy, Completeness, and Reasonableness, and considering its ease of use and cost effectiveness, we selected DeepSeek-R1 for subsequent large-scale Q&A pair generation to support efficient expansion of the remaining dataset.

To evaluate the performance of the proposed finetuned LLM on surface chemistry question answering tasks, we benchmarked it against a set of existing chemical language models. Performance was assessed using Bert-recall, Bert-precision, and Bert-*F*1, as detailed in Section 1.4 of the SI. The BertScores for the different LLM models reveal the limitations of existing models in the domain of surface chemistry (Fig. 4b). ChemGPT[52] exhibited near-zero performance, as it is primarily trained for molecular structure and molecular formula generation, resulting in a mismatch with the requirements of the evaluated tasks. Other chemistry domain LLMs, including ChemLLM,[53] ChemDFM,[35] and Darwin,[54] as well as the base LLaMA model, achieved similarly low performance on surface chemistry Q&A tasks, with Bert-*F*1 scores around 0.4.

To provide a broader zero-shot reference, we additionally evaluated two general-purpose commercial LLMs, GPT-4o-mini and Gemini 2.5, under the same benchmark. These models achieved Bert-*F*1 scores of 0.6 and 0.63, respectively, substantially outperforming the chemistry-oriented baselines that were not specifically adapted to surface chemistry, still falling short of our proposed domain-adapted model. In contrast, our model achieved a Bert-*F*1 score exceeding 0.8, demonstrating the effectiveness of targeted domain adaptation for the knowledge structure and reasoning demands of surface chemistry.

To examine the impact of training dataset size on model performance, we conducted an ablation study using the base model (0 K) as a reference and systematically increasing the

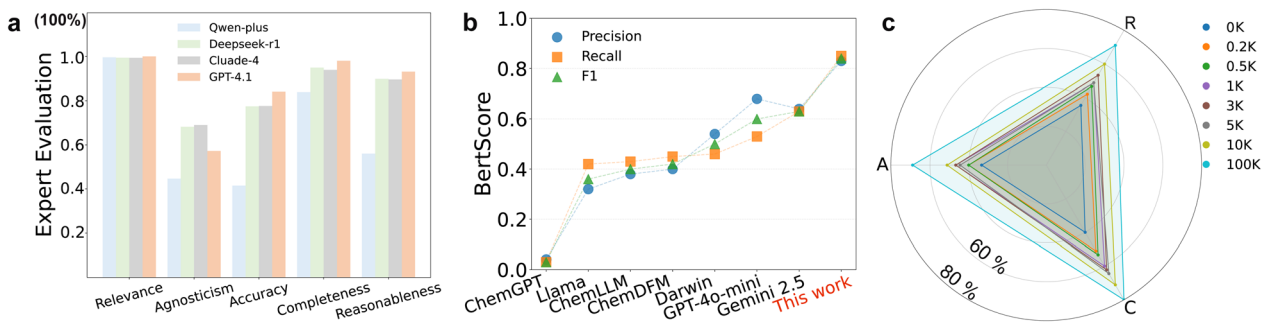

Fig. 4 The evaluation results of the question–answering task. (a) The comparison for Relevance, Agnosticism, Accuracy, Completeness and Reasonableness across four LLMs (Qwen-Plus, DeepSeek-R1, Claude-4, and GPT-4.1). (b) Performances of the model in this work and other chemistry/commercial large language models based on the Bert-recall, Bert-precision, and Bert-*F*1 scores. (c) Performance across different numbers of training samples, evaluated using three dimensions: Accuracy (A), Completeness (C), and Reasonableness (R).

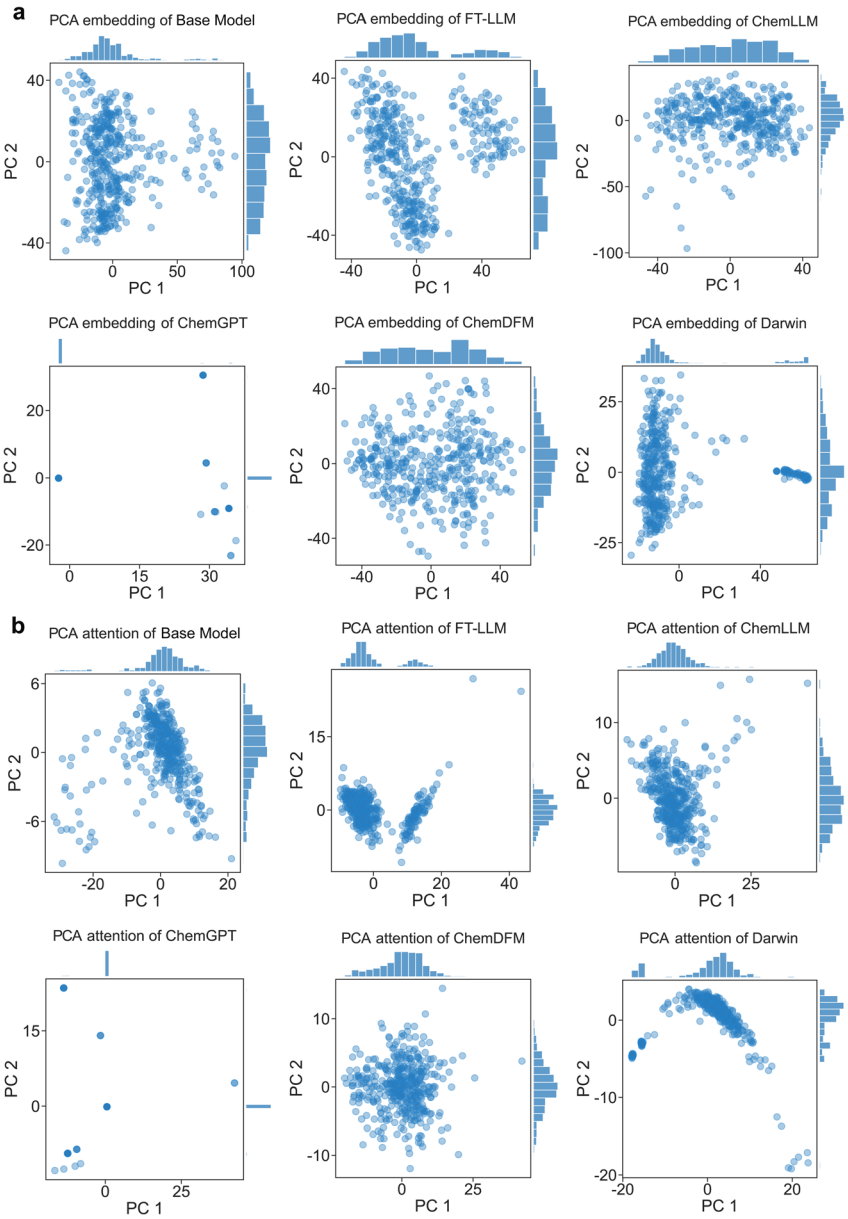

Fig. 5 PCA of embeddings and attention patterns for six language models. (a) PCA of embeddings. (b) PCA of attention patterns. Marginal histograms along the top and right of each panel show the distribution density along the corresponding axes.

number of training samples from 0.2 K to 100 K question–answer pairs. We utilized three metrics for evaluation: Accuracy, Completeness, and Reasonableness. As displayed in Fig. 4c, the model performance exhibited a steady upward trend with increasing training dataset size. The base model achieved scores of approximately 34%, 40%, and 36% on Accuracy, Completeness, and Reasonableness, respectively. In comparison, the fine-tuned model with 100 K samples reached substantially higher scores of 70%, 80%, and 72%, respectively. Relative to the base model, Accuracy increased by approximately 105.9%, while Completeness and Reasonableness each improved by about 100.0%. Even when compared with the model fine-tuned using 0.2 K samples (Accuracy, Completeness, and Reasonableness of 40%, 50%, and 42%, respectively), notable performance gains were observed, with improvements of approximately 75.0% in Accuracy, 60.0% in Completeness, and 71.4% in Reasonableness. These performance gains demonstrate that large-scale and high-quality question–answering datasets are a prerequisite for achieving substantial improvements in model performance.

To visualize differences of distributions in the latent representations (including embeddings and self-attention patterns) across the base model, the fine-tuned LLM, and other chemistry specific LLMs (ChemLLM, ChemGPT, ChemDFM, and Darwin), we applied principal component analysis (PCA) and uniform manifold approximation and projections (UMAP). Note that these analyses were not performed for the commercial baselines, as comparable representation-level access is not available

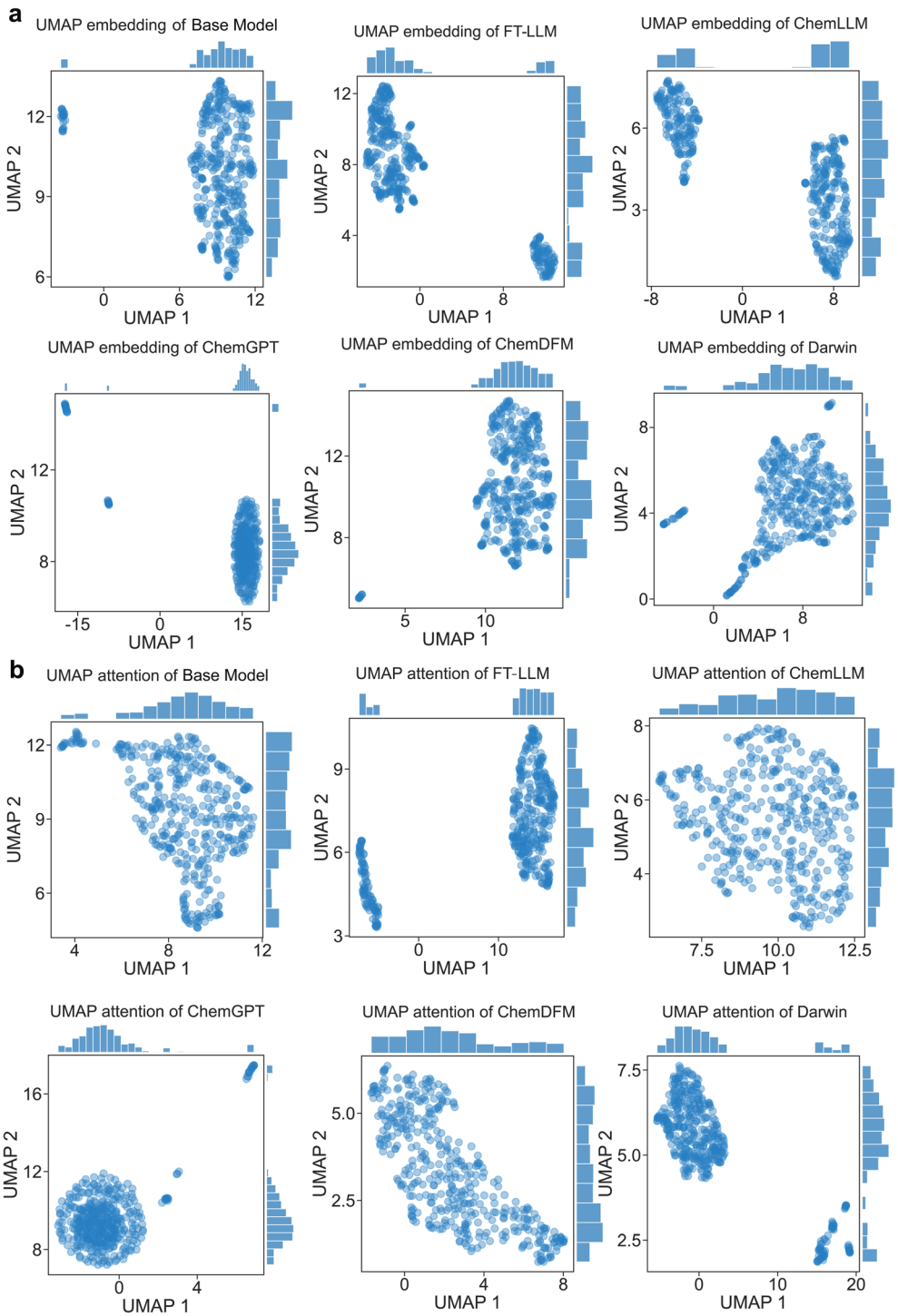

Fig. 6 UMAP results of embeddings and attention patterns for six language models. (a) UMAP of embeddings. (b) UMAP of attention patterns. Marginal histograms along the top and right of each panel show the distribution density along the corresponding axes.

for closed-source models. Fig. 5 and 6 present the latent space for the models, while a more comprehensive comparative analysis involving different training dataset sizes is provided in SI Section 5. These projections illustrate the distribution of embeddings and attention patterns in latent spaces, highlighting inter-model differences in representational diversity and clustering behavior. Compared to the base LLaMA model, the fine-tuned LLM (FT-LLM) exhibits favorable separation in both embedding and attention spaces. While the base model shows broadly dispersed representations, fine-tuning appears to promote the organization of internal representations for the evaluated tasks. This tighter clustering in the latent space reflects more compact internal representations under the evaluated conditions. In contrast, ChemGPT shows a compressed

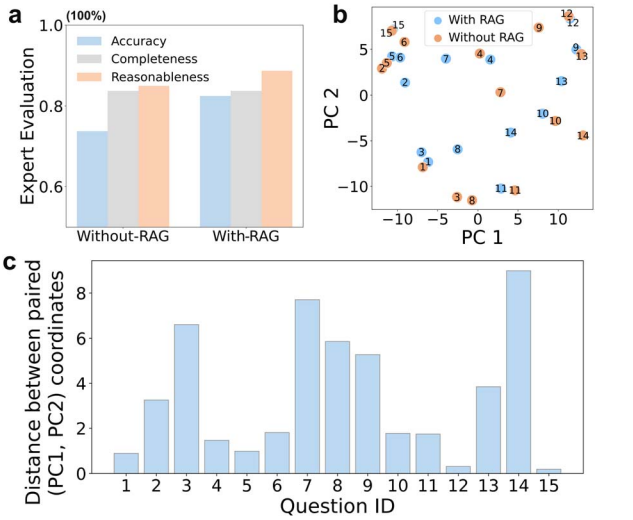

Fig. 7 Performance of the RAG-LLM. (a) The performance comparison of models with and without Retrieval Augmented Generation (RAG). (b) Responses generated with and without RAG are visualized in a two-dimensional principal component space constructed using PCA. (c) Distances between paired responses generated with and without RAG for the same question in a two-dimensional PCA space.

distribution in PCA and UMAP projections, with data points concentrated in a small portion of the latent space. Such compression may reflect reduced diversity in the generated representations. In contrast, ChemGPT exhibits a compressed distribution in PCA. This pattern is consistent with the repetitive nature of SMILES-style representations, which may limit the effective diversity of latent representations. Furthermore, models such as ChemLLM, ChemDFM, and Darwin, which have been fine-tuned on general chemical knowledge, exhibit partial clustering behavior across the projected spaces. These results indicated that fine-tuning on general chemistry alone does not fully translate to improved representation organization for the specific surface chemistry tasks considered here.

The effectiveness of the RAG framework was validated across 15 questions using expert human judgment and embedding-based metrics. Fig. 7a illustrates a paired comparison between the standalone GPT-4.1 backbone and its retrieval-augmented counterpart. This comparison highlights the impact of RAG on three key human-centric dimensions: Accuracy, Completeness, and Reasonableness. The results indicate that incorporating RAG leads to performance improvements across all evaluation metrics. In particular, the accuracy score increased from approximately 74% for the model without RAG to approximately 84% for the RAG-based LLM. This improvement suggests that retrieval augmentation contributes to reduced hallucination in large language models by grounding responses in retrieved external information. The Reasonableness score increased from approximately 85% for the model without RAG to approximately 89%. This gain indicates that providing precise, fact-based contextual information helps improve the logical coherence of the model's reasoning process and the consistency of its final conclusions. We employed PCA to project the embeddings onto the first two principal components, enabling visualization and quantitative confirmation of embedding shifts as shown in Fig. 7b. Responses were projected as points in the embedding space, with a consistent label used to denote responses derived from the same question. Blue circles denote responses generated with RAG, while orange circles represent responses generated without RAG. The distance observed between paired responses in the two-dimensional PCA space indicates structural differences in the generated outputs. To quantify the representation changes introduced by the RAG framework, we computed Euclidean distances between paired responses in the PCA projected space (Fig. 7c).

This evaluation follows a paired design, in which each validation question serves as its own control. For each question, the distance was calculated between the response generated with the RAG framework and its corresponding response generated without it, based on their coordinates in the first two principal components. While the magnitude of these distances varies across questions, reflecting heterogeneity in the impact on individual responses, the embeddings overall exhibit consistent separation between the two conditions (Fig. 7c). It should be noted that the two-dimensional PCA projection captures only a portion of the total variance in the embedding space. Nevertheless, the observed separation indicates systematic shifts in representation induced by retrieval augmentation. In summary, our analysis indicates that incorporating a retrieval-augmented generation approach improved factual consistency and robustness in domain-specific question–answering. Detailed methodology and retrieval pipeline design are provided in SI Section 6.

To assess performance across different knowledge tiers, Q&A pairs are categorized into general and synthesis types, contrasting conceptual and synthesis-oriented reasoning, as shown in Fig. 8 for both the fine-tuned and RAG based LLMs. For general surface-chemistry questions, both models demonstrate strong capabilities in conceptual explanation and domain-knowledge response. The fine-tuned LLM example addresses a foundational question regarding activation mechanisms in surface reactions, clearly distinguishing between thermal activation, photoexcitation, and tunneling electron excitation. The response follows a structured explanatory format, beginning with a conceptual decomposition of the problem and subsequently providing concise definitions of each activation pathway.

By integrating the retrieved literature with parametric knowledge, the model explains whether the metal substrate acts as a passive template or an active catalyst, and further compares Au(111), Ag(111), and Cu(111) in terms of their relative catalytic strengths. Together, these general examples confirm that the system reliably handles conceptual queries that require accurate, scientific descriptions. In contrast, synthesis-oriented questions assess the model's ability to reason about multi-step reaction pathways, precursor behavior, and structure–property relationships that are central to on-surface synthesis. The fine-tuned LLM example describes the hierarchical three step activation of DP-DBBA on Au(111), detailing the molecular structures formed at each thermal stage, including

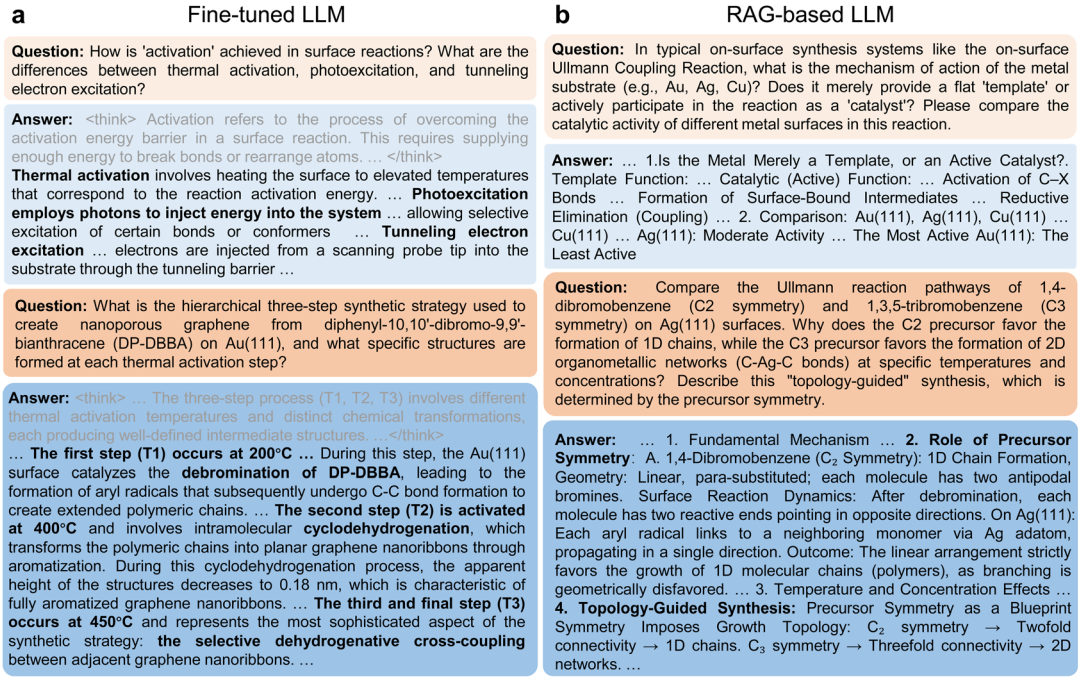

Fig. 8 Responses of two modes of the LLM to different types of questions. (a) The fine-tuned LLM; (b) RAG-based LLM. Light background represents general-type questions, and dark background represents synthesis-type questions.

dehalogenation, cyclodehydrogenation, and final polymer aromatization. This example illustrates the model's ability to summarize temperature dependent reaction intermediates within a complex synthetic sequence. The RAG-based LLM example further extends the depth of reasoning by comparing the Ullmann coupling pathways of C_2 symmetric 1,4 dibromobenzene and C_3 symmetric 1,3,5 tribromobenzene on Ag(111). The response establishes a clear causal relationship linking precursor symmetry, radical coupling geometry, and the dimensionality of the resulting surface networks, while also accounting for the effects of temperature and molecular concentration. This structured, mechanistic reasoning indicates that retrieval augmented generation enables the model to relate microscopic precursor topology to macroscopic reaction products. Collectively, these four Q&A examples indicate that the proposed dual LLM architecture can support both reliable conceptual responses to general scientific questions and more detailed mechanistic reasoning for synthesis level problems, aligning with the requirements of an expert oriented system for on-surface reaction analysis.

Conclusions

In this work, we aim to address a long-standing bottleneck in surface chemistry: the absence of systematic, machine-readable knowledge capable of supporting cumulative understanding and data-driven discovery. By integrating large-scale literature curation, structured data construction, and domain-specialized language modeling, we establish a comprehensive framework that bridges the unstructured surface-chemistry literature and an intelligent question–answer system. Through multi-stage semantic screening, we assembled a high-quality corpus encompassing tens of thousands of surface chemistry and on-surface reaction studies. From this corpus, we extracted a rich set of reaction attributes that capture precursor identity, substrate crystallography, activation pathways, and reaction products, forming a structured database tailored to the unique characteristics of on-surface synthesis. Leveraging these resources, we developed a dual-mode LLM system that unifies parameter-efficient fine-tuning for reasoning with a dual-source retrieval-augmented generation framework for accurate and verifiable access to experimental details.

Comprehensive evaluations show that domain-specific fine-tuning enhances performance over existing chemistry-oriented language models, while retrieval augmentation further reduces hallucination and improves logical coherence by grounding responses in literature-derived evidence. Analyses of latent representations reveal that targeted training reorganizes the model's internal space toward a task-relevant structure, underscoring the importance of domain-aligned data in specialized scientific reasoning. Beyond providing an effective question–answering assistant, this work offers a generalizable paradigm for transforming fragmented scientific literature into structured knowledge and actionable intelligence. The resulting platform offers a promising basis for future progress in on-surface reaction prediction, experimental condition optimization, and autonomous research workflows.[15,55]

Author contributions

Juan Xiang: conceptualization, methodology, data annotation, writing original draft. Qi Huang: data annotation, investigation.

Xinyi Zhang: data annotation, investigation. Tairan Yang: data annotation, investigation. Zhiwen Zhu: data annotation, investigation, visualization. Chanyu Li: data annotation, investigation. Liangliang Cai: data annotation, investigation. Qiang Sun: conceptualization, supervision, data annotation, writing – review & editing.

Conflicts of interest

The authors declare no competing financial interest.

Data availability

The source code used in this study is publicly available at GitHub: https://github.com/juanxiang-shu/OSSAssitant. The fine-tuned model checkpoints are available at Hugging Face: https://huggingface.co/JuanXiang-SHU/SurfaceScienceAssistant. The training data used for LLM fine-tuning are available at: https://huggingface.co/datasets/JuanXiang-SHU/Surface_Chemistry. In addition, answers to the sample questions are available at: https://github.com/juanxiang-shu/OSSAssitant/tree/main/Q&A/Sample.

Supplementary information (SI): detailed model training procedures, prompt designs and RAG framework details. See DOI: https://doi.org/10.1039/d6sc01168c.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 22302120). J. X. acknowledges Hao Jiang for helpful discussions on attributes extracted from the literature of on-surface reactions. The authors would like to thank the developers of Bgolearn (https://doi.org/10.48550/arXiv.2601.06820) for providing the Bayesian optimization framework used to optimize the hyperparameters during the SciBERT model training process.

References

1 G. A. Somorjai and Y. Li, *Introduction to surface chemistry and catalysis*, John Wiley & Sons, 2010.
2 A. Cucinotta, J. A. Davies, K. S. Mali and S. De Feyter, Scanning probe microscopy of metal–organic coordination systems: characterization of monolayers, single crystals, discrete architectures, *Chem. Soc. Rev.*, 2025, **54**, 10654–10689.
3 S. Clair and D. Oteyza, Controlling a Chemical Coupling Reaction on a Surface: Tools and Strategies for On-Surface Synthesis, *Chem. Rev.*, 2019, **119**, 4717–4776.
4 L. L. Patera, S. Fatayer, J. Repp and L. Gross, Probing Molecular Properties at Atomic Length Scale Using Charge-State Control, *Chem. Rev.*, 2025, **125**, 5830–5847.
5 R. K. Houtsma, J. de la Rie and M. Stöhr, Atomically precise graphene nanoribbons: interplay of structural and electronic properties, *Chem. Soc. Rev.*, 2021, **50**, 6541–6568.
6 Q. Sun, R. Zhang, J. Qiu, R. Liu and W. Xu, On-Surface Synthesis of Carbon Nanostructures, *Adv. Mater.*, 2018, **30**, 1705630.
7 W. Xiong, G. Zhang, D.-L. Bao, J. Lu, L. Gao, Y. Li, H. Zhang, Z. Ruan, Z. Hao, H.-J. Gao, L. Chen and J. Cai, Visualizing stepwise evolution of carbon hybridization from sp3 to sp2 and to sp, *Nat. Commun.*, 2025, **16**, 690.
8 H. Jiang, Z. Zhu, X. Zhang, S. Yuan, K. Guo, J. Li and Q. Sun, Exploring Selective Photochemistry on Metal Surfaces through Wavelength-Dependent Light Excitation, *Nano Lett.*, 2025, **25**, 9597–9604.
9 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, Autonomous chemical research with large language models, *Nature*, 2023, **624**, 570–578.
10 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, Augmenting large language models with chemistry tools, *Nat. Mach. Intell.*, 2024, **6**, 525–535.
11 S.-H. Yoo, A. K. Y. Low, J. Recatala-Gomez, H. Sahu, C. Kim, J. F. Joung, H. Chun, K. A. Christofidou, J. Berry, M. Minotakis, K. Kang, K.-s. Kim, G. Shin, H. Jang, S. Lee, M. Park, B.-H. Kim, K. Shin, J. Shin, A. Soon, J. Schrier and W. Jang, Exploring materials data through collaboration: 2024 KRICT ChemDX Hackathon, *J. Mater. Inf.*, 2025, **5**, 54.
12 Z. Zhu, S. Yuan, Q. Yang, H. Jiang, F. Zheng, J. Lu and Q. Sun, Autonomous Scanning Tunneling Microscopy Imaging via Deep Learning, *J. Am. Chem. Soc.*, 2024, **146**, 29199–29206.
13 T. Song, M. Luo, X. Zhang, L. Chen, Y. Huang, J. Cao, Q. Zhu, D. Liu, B. Zhang, G. Zou, G. Zhang, F. Zhang, W. Shang, Y. Fu, J. Jiang and Y. Luo, A Multiagent-Driven Robotic AI Chemist Enabling Autonomous Chemical Research On Demand, *J. Am. Chem. Soc.*, 2025, **147**, 12534–12545.
14 Z. Zhu, J. Lu, S. Yuan, Y. He, F. Zheng, H. Jiang, Y. Yan and Q. Sun, Automated Generation and Analysis of Molecular Images Using Generative Artificial Intelligence Models, *J. Phys. Chem. Lett.*, 2024, **15**, 1985–1992.
15 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff, M. Seifrid and A. Aspuru-Guzik, Self-Driving Laboratories for Chemistry and Materials Science, *Chem. Rev.*, 2024, **124**, 9633–9732.
16 H. Tian, Y. Hu, Z. Ding, J. He and J. Xiong, Dynamic physics-guided neural network for predicting hot deformation behavior of TiAl-based intermetallic alloys, *Mater. Genome Eng. Adv.*, 2025, **3**, e70033.
17 J. Okuyama, Z. Diao, H. Yamashita and M. Abe, Integrated AI Framework for Room-Temperature Atom Manipulation in Scanning Probe Microscopy, *Nano Lett.*, 2025, **25**(51), 17771–17777.
18 L. Hawizy, D. M. Jessop, N. Adams and P. Murray-Rust, ChemicalTagger: A tool for semantic text-mining in chemistry, *J. Cheminf.*, 2011, **3**, 17.
19 M. C. Swain and J. M. Cole, ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.
20 https://scifinder.cas.org.

21 https://www.scopus.com.
22 https://www.reaxys.com.
23 https://next-gen.materialsproject.org/materials/.
24 Z. Zheng, N. Rampal, T. J. Inizan, C. Borgs, J. T. Chayes and O. M. Yaghi, Large language models for reticular chemistry, *Nat. Rev. Mater.*, 2025, **10**, 369–381.
25 W. Takahara, Y. Yamaguchi, M. Ogano, F. Kakami, Y. Harashima, T. Takayama, S. Takasuka, A. Kudo and M. Fujii, Materials Dual-Source Knowledge Retrieval-Augmented Generation for Local Large Language Models in Photocatalysts, *J. Chem. Inf. Model.*, 2025, **65**, 13098–13114.
26 Z. Diao, H. Yamashita and M. Abe, Leveraging large language model and social network service for automation in scanning probe microscopy, *Meas. Sci. Technol.*, 2025, **36**, 047001.
27 H. Chen, H. Liu, Y. Tew, X. Ren, X. Tang and X. Wang, Distilling Knowledge from Catalysis Literature with Long-Context Large Language Model Agents, *ACS Catal.*, 2025, **15**, 18244–18254.
28 J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics*, 2020, **36**, 1234–1240.
29 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, A generative model for inorganic materials design, *Nature*, 2025, **639**, 624–632.
30 I. Beltagy, K. Lo and A. Cohan, *SciBERT: A pretrained language model for scientific text*, 2019, pp. 3615–3620.
31 S. Chithrananda, G. Grand and B. Ramsundar, ChemBERTa: large-scale self-supervised pretraining for molecular property prediction, *arXiv*, 2020, preprint, arXiv:2010.09885, DOI: 10.48550/arXiv.2010.09885.
32 L. M. Antunes, K. T. Butler and R. Grau-Crespo, Crystal structure generation with autoregressive large language modeling, *Nat. Commun.*, 2024, **15**, 10570.
33 Y. Kang, H. Park, B. Smit and J. Kim, A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks, *Nat. Mach. Intell.*, 2023, **5**, 309–318.
34 Y. Zhang, Y. Han, S. Chen, R. Yu, X. Zhao, X. Liu, K. Zeng, M. Yu, J. Tian, F. Zhu, X. Yang, Y. Jin and Y. Xu, Large language models to accelerate organic chemistry synthesis, *Nat. Mach. Intell.*, 2025, **7**, 1010–1022.
35 Z. Zhao, D. Ma, L. Chen, L. Sun, Z. Li, Y. Xia, B. Chen, H. Xu, Z. Zhu, S. Zhu, S. Fan, G. Shen, K. Yu and X. Chen, Developing ChemDFM as a large language foundation model for chemistry, *Cell Rep. Phys. Sci.*, 2025, **6**, 102523.
36 Z. Ruan, J. Schramm, J. B. Bauer, T. Naumann, L. V. Müller, F. Sättele, H. F. Bettinger, R. Tonner-Zech and J. M. Gottfried, On-Surface Synthesis and Characterization of Pentadecacene and Its Gold Complexes, *J. Am. Chem. Soc.*, 2025, **147**, 4862–4870.
37 A. Kinikar, X. Xu, T. Onishi, A. Ortega-Guerrero, R. Widmer, N. Zema, C. Hogan, L. Camilli, L. Persichetti, C. A. Pignedoli, R. Fasel, A. Narita and M. Di Giovannantonio, On-surface synthesis of tailored organic platforms for single metal atoms, *Nat. Commun.*, 2025, **16**, 10597.
38 K. Biswas, A. García-Frutos, B. Álvarez, M. Lozano, A. Barragán, J. Janeiro, J. Bello-García, D. Soler-Polo, K. Lauwaet and J. M. Gallego, Driving Multi-Step Regioselectivity in On-Surface Polymer Synthesis by Molecular Coverage, *Angew. Chem.*, 2025, **137**, e202512575.
39 D. Li, N. Cao, M. Metzelaars, O. J. Silveira, J. Jestilä, A. Fumega, T. Nishiuchi, J. Lado, A. S. Foster, T. Kubo and S. Kawai, Frustration-Induced Many-Body Degeneracy in Spin −1/2 Molecular Quantum Rings, *J. Am. Chem. Soc.*, 2025, **147**, 26208–26217.
40 H. Zhu, J. Wang, K. Niu, Y. Zhang, Y. Zhang, C. Deng, P. Huang, D. Li, P. Liu, J. Lu, J. Rosen, J. Björk, J. Cai, L. Chi and Q. Li, Real-space investigations of on-surface intermolecular radical transfer reactions assisted by persistent radicals, *Sci. Adv.*, 2025, **11**, eadu9436.
41 Y. Li, Q. Huang, T. Yang, Z. Zhu, S. Yuan, Q. Yang, X. Zhang and Q. Sun, Self-Driving Laboratory for Accelerated On-Surface Synthesis under Ultrahigh Vacuum, *Nano Lett.*, 2025, **25**, 11609–11617.
42 Y. He, H. Jiang, S. Yuan, J. Lu and Q. Sun, On-surface photo-induced dechlorination, *Chin. Chem. Lett.*, 2024, **35**, 109807.
43 H. Jiang, Y. He, J. Lu, F. Zheng, Z. Zhu, Y. Yan and Q. Sun, Unraveling the Mechanisms of On-Surface Photoinduced Reaction with Polarized Light Excitations, *ACS Nano*, 2024, **18**, 1118–1125.
44 J. Xiang, Y. Li, X. Zhang, Y. He and Q. Sun, Local large language model-assisted literature mining for on-surface reactions, *Mater. Genome Eng. Adv.*, 2025, **3**, e88.
45 C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.*, 2020, **21**, 1–67.
46 C.-H. Shu, M.-X. Liu, Z.-Q. Zha, J.-L. Pan, S.-Z. Zhang, Y.-L. Xie, J.-L. Chen, D.-W. Yuan, X.-H. Qiu and P.-N. Liu, On-surface synthesis of poly(p-phenylene ethynylene) molecular wires via in situ formation of carbon-carbon triple bond, *Nat. Commun.*, 2018, **9**, 2322.
47 R. Zuzak, P. Dabczynski, J. Castro-Esteban, J. I. Martínez, M. Engelund, D. Pérez, D. Peña and S. Godlewski, Cyclodehydrogenation of molecular nanographene precursors catalyzed by atomic hydrogen, *Nat. Commun.*, 2025, **16**, 691.
48 Q. Du, X. Su, Y. Liu, Y. Jiang, C. Li, K. Yan, R. Ortiz, T. Frederiksen, S. Wang and P. Yu, Orbital-symmetry effects on magnetic exchange in open-shell nanographenes, *Nat. Commun.*, 2023, **14**, 4802.
49 D. J. Rizzo, G. Veber, T. Cao, C. Bronner, T. Chen, F. Zhao, H. Rodriguez, S. G. Louie, M. F. Crommie and F. R. Fischer, Topological band engineering of graphene nanoribbons, *Nature*, 2018, **560**, 204–208.
50 A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten and A. Vaughan,

The llama 3 herd of models, *arXiv*, 2024, preprint, arXiv:2407.21783, DOI: 10.48550/arXiv.2407.21783.

51 Y. Wan, Y. Liu, A. Ajith, C. Grazian, B. Hoex, W. Zhang, C. Kit, T. Xie and I. Foster, SciQAG: A framework for auto-generated science question answering dataset with fine-grained evaluation, *arXiv*, 2024, preprint, arXiv:2405.09939, DOI: 10.48550/arXiv.2405.09939.

52 Y. Zhang, W. Liu, Y. Zhang, D. Xiong, J. Zhai, H. Hao, Y. Gu, H. Yang, S. Gao and L. Hu, A Large Language Model for Chemistry and Retrosynthesis Predictions, *arXiv*, 2025, preprint, arXiv:2507.01444, DOI: 10.48550/arXiv.2507.01444.

53 D. Zhang, W. Liu, Q. Tan, J. Chen, H. Yan, Y. Yan, J. Li, W. Huang, X. Yue, D. Zhou, S. Zhang, M. Su, H.-S. Zhong, Y. Li and W. Ouyang, ChemLLM: A Chemical Large Language Model, *arXiv*, 2024, preprint, abs/2402.06852, DOI: 10.48550/arXiv/2402.06852.

54 T. Xie, Y. Wan, W. Huang, Z. Yin, Y. Liu, S. Wang, Q. Linghu, C. Kit, C. Grazian, W. Zhang, I. Razzak and B. Hoex, DARWIN Series: Domain Specific Large Language Models for Natural Science, *arXiv*, 2023 preprint, abs/2308.13565, DOI: 10.48550/arXiv/2308.13565.

55 T. Dai, S. Vijayakrishnan, F. T. Szczypiński, J.-F. Ayme, E. Simaei, T. Fellowes, R. Clowes, L. Kotopanov, C. E. Shields, Z. Zhou, J. W. Ward and A. I. Cooper, Autonomous mobile robots for exploratory synthetic chemistry, *Nature*, 2024, **635**, 890–897.