A Dual-Mode Large Language Model Assistant for On-Surface Reaction via Fine-Tuning and Retrieval-Augmented Generation

Juan Xiang[1], Qi Huang[1], Xinyi Zhang[1], Tairan Yang[1], Zhiwen Zhu[1], Chanyu Li[2], Liangliang Cai[1], Qiang Sun[1,2]*

[1]Materials Genome Institute, Shanghai University, 200444 Shanghai, China.

[2]Qianweichang College, Shanghai University, 200444 Shanghai, China.

*E-mail: qiangsun@shu.edu.cn

Table of Content

1. Methods

1.1 T5 model training

Metadata extracted from PubMed records were stored in an unstructured plain text format. Given that traditional regular expression-based parsing methods are susceptible to format variations when processing such a large-scale dataset with heterogeneous text formats, they struggle with deep semantic understanding and inevitably introduce errors in the extraction of metadata (particularly the abstract). To ensure the precise conversion of citation information and data integrity, we adopted a Text-to-Text transformer to transform the original text-formatted metadata into the standard BibTeX format, thereby structuring the metadata for subsequent tasks such as classification. We selected the T5 model[1] as the base model, whose encoder-decoder architecture exhibits superior performance in sequence-to-sequence text transformation tasks. We first constructed a sample dataset comprising 490 instances and performed fine-tuning using this dataset. These 490 samples come from the DeepSeek model with temperature = 0.1. These results are checked by human experts. The key parameters are as follows: a learning rate of 1.13×10^{-4}, a total of 6 training epochs, and a batch size of 1 combined with 4 gradient accumulation steps.

1.2 SciBERT model training

We fine-tuned the SciBERT model for a two-step classification task using 900 human-annotated samples as the training set, employing Bayesian Global Optimization (Bgolearn)[2] to identify the optimal hyperparameters. As detailed in Table S1 (with optimization data in Tables S2 and S3), the model's performance significantly improved after fine-tuning: in the first phase, the macro recall, macro precision, and macro F1 score increased substantially from 0.49, 0.41, and 0.39 to 0.92, respectively; similarly, in the second phase, these three metrics rose from 0.36, 0.38, and 0.13 to 0.89. The optimized hyperparameters of the first-stage included batch size, which determines samples processed per weight update; learning rate, controlling the step size of updates; weight decay, a regularization technique against overfitting; max length for input sequence limits; early stop patience, which prevents overfitting by setting the maximum non-improving epochs on the validation set. In the second-stage classification task, we used the full text as input to enable the model to capture richer linguistic cues and thereby achieve more accurate identification of on-surface reaction literature. To accommodate the long document length, we implemented a sliding-window mechanism that efficiently processes full-text inputs and supports systematic document chunking. Consequently, the optimization process was expanded to include the following key parameters specific to this long-text processing approach:

1. Stride: This parameter represents the distance of movement or the overlap between adjacent windows, which is crucial for controlling the contextual continuity across segmented chunks.
2. Maximum number of windows: This metric constrains the maximum number of segments into which the input text may be partitioned. By leveraging optimization, the system adaptively identifies an optimal trade-off between model performance and computational cost, thereby preventing unnecessary resource consumption and ensuring efficient utilization of computational budget.
3. Aggregator: Following the individual classification of all windows, a strategic aggregator is necessary to summarize the prediction results from each window to yield the final document-level classification. Candidate aggregation methods include selecting the maximum value or calculating the average value of the window-level predictions.

The formula for calculating F1 is as follows:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

The evaluation function used is Macro F1 (Macro-averaged F1 score).[3] The formula for calculating Macro F1 is as follows:

$$\text{Macro F1} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{F1}_i$$

Using a Macro F1 Score means that the evaluation process independently calculates the F1 score (harmonic average of precision and recall) for each class, and then averages the F1 scores for all classes. This ensures that the model does not exhibit artificially high overall performance due to a large sample size in some classes of the dataset, thus guaranteeing high accuracy and reliability in distinguishing all target classification categories. Through these two stages of accurate semantic classification, a highly robust literature dataset that closely matches the research objectives was finally obtained. The main specifications of the computation platform running this program are as follows: i5-13490F @ 2.5GHz CPU, 16GB RAM.

Table S1. Performance comparison of the SciBERT model before (Baseline) and after fine-tuning (Trained) for the two-stage literature classification task.

		Baseline	Trained
Step 1	Macro Recall	0.49	0.92
	Macro Precision	0.41	0.92
	Macro F1	0.39	0.92
Step 2	Macro Recall	0.36	0.89
	Macro Precision	0.38	0.89
	Macro F1	0.13	0.89

Table S2. The Bayesian optimization process of the first-stage classification model; the parameters highlighted in bold are the optimal parameters.

batch size	learning rate	weight decay	focal gamma	max length	early stop patience	Macro recall	Macro precision	Macro F1
4	1.70E-06	2.94E-06	1	128	10	0.905	0.906	0.904
2	2.81E-06	3.75E-05	4	256	5	0.914	0.916	0.914
4	7.50E-06	1.38E-06	5	512	9	0.917	0.917	0.916
8	1.51E-06	3.06E-05	1	128	7	0.909	0.910	0.907
8	2.44E-05	4.84E-04	5	128	5	0.892	0.897	0.889
8	2.60E-06	4.25E-05	1	384	4	0.906	0.905	0.903
4	1.38E-05	1.67E-06	3	256	3	0.902	0.905	0.901
8	2.04E-05	2.61E-05	1	512	6	0.918	0.916	0.916
2	1.11E-06	8.11E-05	3	256	9	0.907	0.908	0.906
8	**2.36E-05**	**2.66E-04**	**5**	**512**	**7**	**0.919**	**0.918**	**0.917**
16	6.80E-06	8.10E-04	7	512	7	0.895	0.903	0.895
8	2.77E-05	1.71E-04	7	512	6	0.909	0.908	0.908
8	1.35E-05	1.07E-05	5	512	8	0.908	0.910	0.907
16	1.34E-05	8.79E-06	3	512	6	0.913	0.913	0.911
8	1.90E-05	2.39E-04	6	384	8	0.910	0.909	0.908
8	9.74E-06	1.22E-05	2	512	5	0.909	0.911	0.907
8	3.55E-06	1.18E-04	4	512	8	0.899	0.901	0.897
2	4.75E-06	2.89E-04	6	512	6	0.888	0.895	0.883
16	2.00E-05	2.09E-05	2	384	4	0.914	0.915	0.913
8	1.07E-05	7.46E-05	4	512	7	0.918	0.919	0.917

Table S3. The Bayesian optimization process of the second-stage classification model; the parameters highlighted in bold are the optimal parameters.

batch size	learning rate	weight decay	focal gamma	max length	stride	max windows	aggregator	early stop patience	Macro recall	Macro precision	Macro f1
4	1.04E-05	2.94E-06	1	384	256	4	max	6	0.865	0.876	0.868
8	1.73E-06	7.52E-06	2	384	192	4	max	7	0.821	0.819	0.814
8	5.60E-06	2.32E-06	3	384	128	8	max	8	0.838	0.873	0.852
4	2.15E-06	1.37E-06	2	512	256	6	mean	4	0.841	0.844	0.836
4	**1.73E-05**	**2.06E-04**	**1**	**512**	**128**	**8**	**max**	**5**	**0.889**	**0.889**	**0.888**
8	1.88E-06	6.15E-04	5	384	192	6	max	5	0.802	0.819	0.804
4	7.37E-06	1.79E-05	2	512	256	6	max	4	0.853	0.869	0.859
8	7.15E-06	1.43E-06	2	256	192	6	mean	6	0.839	0.855	0.843
2	2.63E-05	9.17E-06	1	512	192	8	max	3	0.847	0.833	0.814
2	3.49E-05	1.08E-04	5	512	128	8	mean	5	0.874	0.882	0.876
2	4.89E-05	9.28E-05	5	512	128	8	mean	5	0.748	0.799	0.742
2	2.43E-05	6.80E-05	4	512	128	8	mean	4	0.862	0.860	0.857
2	2.01E-05	9.32E-04	4	512	128	8	mean	5	0.865	0.880	0.867
4	4.37E-05	2.44E-04	4	512	128	8	mean	3	0.846	0.860	0.850

1.3 LLaMA model training

To develop the specialized domain assistant, the Meta-LLaMA-3.1-8B model was selected as the base large language model[4]. The tokenizer is configured with a maximum sequence length of 4096 tokens. We used a chain-of-thought (CoT) template that (i) set the task within the field of surface chemistry, (ii) includes a "reason before answer" layout with a "Think…" area for intermediate reasoning, and (iii) places the final answer after the reasoning module. See *Supplementary Information* Section 4 for details.

For parameter-efficient domain adaptation, we employed Low-Rank Adaptation (LoRA) within the PEFT[5] framework. In LoRA, the pretrained backbone weights remain frozen, and task-specific adaptation is achieved by introducing trainable low-rank update matrices into selected linear transformations, thereby avoiding full-parameter optimization. In the present study, LoRA modules were applied to the attention projection layers, namely *q_proj*, *k_proj*, *v_proj*, and *o_proj*, so that the model could adapt its attention behavior for chemistry-specific reasoning and response generation while preserving the general knowledge encoded in the pretrained model. Formally, for a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA constrains the parameter update through a low-rank decomposition:

$$W = W_0 + \Delta W = W_0 + BA$$

where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, with the rank $r \ll min(d,k)$. Only A and B are optimized during training, while W_0 remains unchanged. By restricting adaptation to these low-rank components, LoRA enables efficient domain specialization and helps reduce catastrophic forgetting of the pretrained model's general knowledge.

We selected LoRA because it offers a favorable balance between domain adaptation and training efficiency for specialized instruction tuning. In our setting, the goal was not to relearn general language competence from the ground up, but to specialize a strong pretrained model for surface-chemistry concepts, synthesis conditions, and mechanistic reasoning. Compared with full-parameter fine-tuning, LoRA updates only a small subset of parameters, allowing task-specific adaptation while preserving the broad scientific knowledge embedded in the backbone model. In contrast to prompt-only adaptation or in-context learning, LoRA encodes domain-relevant behavior directly into the model parameters, thereby enabling more stable and consistent specialization. Other parameter-efficient approaches, prefix-based methods consume valuable context space; adapter-based methods may introduce additional inference overhead. By comparison, LoRA modifies the model through low-rank reparameterization while preserving the overall architecture, which facilitates straightforward deployment through weight merging.

The LoRA rank was set to 8 and the LoRA scaling factor (alpha) was set to 16. Training was performed using supervised fine-tuning with the trl library and the memory-efficient adamw_8bit optimizer. Training was executed using 4-bit NormalFloat (NF4) quantization via the bitsandbytes[6] library. The training hyperparameters included a learning rate of 2×10^{-4}, cosine learning-rate scheduling, 50 warm-up steps, weight decay of 0.01, and a per-device batch size of 1 with 4 gradient accumulation steps, resulting in an effective batch size of 4. Numerical precision was automatically selected as bf16 when supported by the hardware, otherwise fp16 was used. Model performance was evaluated on the validation split after each training epoch. After fine-tuning, the LoRA weights could be either used directly in PEFT form or merged back into the base model for deployment. The entire training workflow was executed on a single NVIDIA H800 GPU. Overall, this training strategy provided a practical and computationally efficient route for chemistry-specific model

specialization, while retaining sufficient flexibility to improve reasoning quality on domain-relevant tasks.

1.4 LLaMA model evaluation

As shown in Figure S1, we systematically evaluated the convergence behavior and training dynamics of the LLaMA model across various sizes of the fine-tuning dataset. Figure S1a demonstrates that increasing the dataset scale significantly improves model convergence; the final loss value steadily decreased from approximately 1.30 (for 0.2 K samples) to a minimum of about 1.08 (for 100 K samples). Figure S1b further highlights the impact of data volume: for smaller datasets (e.g., 0.2 K and 0.5 K), the training loss curve remains noticeably higher and exhibits greater volatility, indicating suboptimal learning stability and efficiency. Conversely, as the amount of training data increases, particularly from 5 K up to 100 K (represented by the green and yellow hues), the loss curves descend rapidly, showing a smoother and more consistent downward trend. Collectively, these findings strongly suggest that utilizing larger datasets is essential for enhancing model convergence and improving generalization capability in this specialized domain.

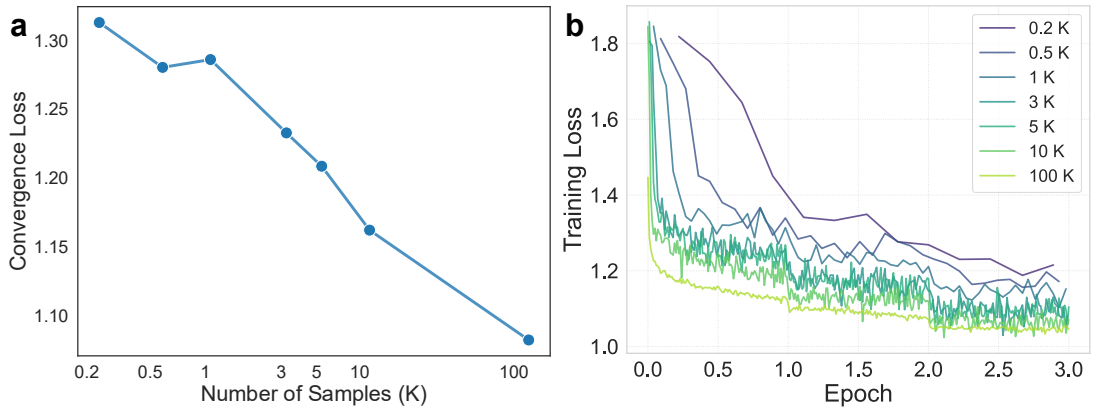

Figure S1. Analysis of the LLaMA model's fine-tuning dynamics as a function of the training dataset size (ranging from 0.2 K to 100 K question-answering pairs). **a,** The convergence loss curve shows the relationship between the final loss value and the number of fine-tuning samples. **b,** Training loss across different training datasets.

2. Evaluation Metrics

To rigorously assess the performance of the dual-LLM framework, we employ a combination of traditional overlap-based metrics and semantic-based evaluation criteria. The detailed formulations of these metrics are provided below.

2.1 F-score metrics

The precision (P) and recall (R) used to evaluate model performance are defined based on the alignment between the generated response and the ground truth. These metrics quantify the exact lexical or structural overlap:

Precision measures the proportion of predicted units that are correct:

$$\text{Precision} = \frac{\text{Ture Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall measures the proportion of actual ground truth units that were correctly predicted:

$$\text{Recall} = \frac{\text{Ture Positive}}{\text{True Positive} + \text{False Negative}}$$

F1-score is the harmonic mean of precision and recall:

$$\mathrm{F1} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

2.2 BERTScore metrics

To provide a robust and context-aware evaluation of the generated responses against the ground truth answers, we employed BertScore[7] as the primary metric, leveraging the semantic power of pre-trained BERT models to compute the similarity between token sets through contextual embeddings. Unlike traditional similarity metrics such as BLEU or ROUGE, which primarily depend on lexical overlap, BERTScore leverages contextualized embeddings to capture paraphrastic variation and deep semantic correspondence, this capability is particularly critical in surface chemistry domains. The metric computes three core components:

1. Bert-precision measures how well the tokens in the generated response (ŷ) are semantically covered by the tokens in the reference answer (y). It essentially checks if the generated output is relevant and focused.

For each token *i* in the generated response *ŷ*, the maximum cosine similarity to any token *j* in the reference *y* is computed:

$$P_{Bert} = \frac{1}{|\hat{y}|}\sum_{i\in|\hat{y}|} \underset{j\in|y|}{max}(X_i^{\top} X_j)$$

Where X_i and X_j are the contextualized BERT embeddings of the tokens *i* and *j*, respectively.

2. Bert-recall measures how well the tokens in the reference answer are semantically covered by the tokens in the generated response. It assesses the completeness of the generated output relative to the ground truth.

For each token *j* in the reference answer *y*, the maximum cosine similarity to any token *i* in the generated response *ŷ* is computed:

$$R_{Bert} = \frac{1}{|y|}\sum_{j\in|y|} \underset{i\in|\hat{y}|}{max}(X_j^{\top} X_i)$$

3. The Bert-F1 score is the harmonic mean of the Precision and Recall, providing a balanced measure of semantic similarity between the generated text and the reference text. This is the primary aggregated performance indicator used in this study.

$$F1_{Bert} = 2 \cdot \frac{P_{Bert} \cdot R_{Bert}}{P_{Bert} + R_{Bert}}$$

In this evaluation, we used the DeBERTa-base-mnli model and set num_layers to 9 to extract contextual embeddings. By relying on these context-aware metrics, we ensure that the evaluation of the dual-LLM accurately reflects its ability to produce semantically correct and comprehensive answers.

3. JSON extraction

3.1 Annotation workflow and quality control

3.1.1 Annotation Guidelines

To minimize subjectivity in the annotation workflow, annotators followed a predefined set of guidelines aligned with the JSON schema. These guidelines were designed to ensure consistency and faithful representation of the source literature.

1. Direct Extraction

All values were required to be extracted directly from the source text without modification or unsupported inference. If a given attribute was not explicitly reported in the original article, it was recorded as null. For example, the precursor be referred to only as “triyne 1” in the article[8]. In such cases, the abbreviation field was recorded as “triyne 1”, whereas the iupac name field was assigned null if no explicit IUPAC name was provided.

2. Units and Deposition Parameters

Within the deposition block, all numerical parameters were required to be recorded together with their physical units. For example, if a precursor was sublimated at 212 °C, the temperature was recorded as “212 °C” rather than as a unitless value. Likewise, the substrate temperature field was required to preserve the exact experimental description provided in the article, such as “room temperature” or “10 K”.

3. Reaction Stage Assignment

The temporal progression of the reaction was explicitly mapped within the JSON structure. Intermediate transformations were recorded in the reaction stages array, whereas the final transformation outcome was recorded in final stage. For example, if polymer formation occurred at 200 °C and subsequent planarization into NH2-chGNRs[9] occurred at 260 °C, the former was annotated under reaction stages and the latter under final stage. If no intermediate stage was described, reaction stages was assigned null.

4. Activation Condition Mapping

The activation block was used to map the physical trigger of the transformation to the appropriate fields. For thermal activation, only thermally relevant parameters were populated, such as annealingTemperature. In contrast, fields specific to scanning tunneling microscope (STM) manipulation, such as bias or tunnelingCurrent, were left as null unless explicitly reported and relevant.

5. Reaction Type Classification

The reactionType field within the mechanism block was assigned according to a predefined schema ontology. Annotators selected the descriptor that best matched the description in the source text.

6. Molecular Identity Tracking

Annotators were required to track changes in molecular identity and nomenclature throughout the reaction sequence. The intermediates and products fields were used to reflect the stage-specific identity reported in the article. For example, a species described as “polymer” after Ullmann coupling and later as “NH2-chGNRs” after cyclodehydrogenation was recorded accordingly in the corresponding reaction-stage fields.

7. Contextual Attribute Recording

The substrate object was required to specify the surface material precisely, using forms such as “Au(111)” or “Cu(111)”. Any relevant preparative or pretreatment information described in the experimental section was extracted into the associated notes field to preserve contextual detail.

Precursor names and abbreviations were recorded exactly as used by the authors. If a precursor was designated by a numerical label in the article, that designation was retained verbatim. For instance, if 2,2’-dibromo-[9,9’-bianthracene]-10,10’-diamine was explicitly labeled as “(3)” in the source text, the corresponding abbreviation field was recorded as “3”.

The morphology and geometry fields were intended to preserve the authors’ original descriptions as closely as possible. Rather than paraphrasing or synthesizing new descriptions, annotators extracted

the relevant wording directly from the article. For example, a molecular geometry description could be recorded as: “This compound is a twisted biphenylophane with a highly distorted diacetylene moiety.”[8]

3.1.2 Annotation workflow

Based on carefully designed prompts, the LLM was first used to perform an initial round of automated information extraction from the surface chemistry and reaction literature that had already undergone multi-stage semantic screening, thereby generating preliminary structured data in JSON format. This was followed by a rigorous manual verification and correction stage. To maximize annotation consistency and identify potential LLM-generated errors or omissions, each paper was independently reviewed by at least two domain experts. During this process, the annotators followed the guidelines described in Section 3.1.1 to verify, supplement, and revise the preliminary LLM outputs. In cases of disagreement, a third senior researcher served as an adjudicator and made the final decision on the basis of the original text and the predefined annotation criteria. This workflow combines the efficiency of LLM-assisted extraction with multiple layers of expert validation, thereby improving data-construction efficiency while supporting the fidelity and internal consistency of the final corpus.

3.1.3 Annotation quality assessment

To quantitatively evaluate the reliability of the annotation procedure, we used Cohen’s kappa (κ) as the primary metric for measuring agreement between two independent annotators. Cohen’s kappa is more robust than simple percent agreement because it corrects for agreement expected by chance. It is defined as

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o denotes the observed proportion of agreement between annotators, and P_e denotes the expected chance agreement based on their individual annotation distributions. In our workflow, each document was independently reviewed by two experts, each of whom produced a finalized structured JSON record. These paired JSON files were treated as the outputs of two raters. We then performed a programmatic field-by-field comparison across the predefined set of key physicochemical attributes, including precursor information, reaction-stage descriptors, and final-product characteristics. Each field value was treated as an individual categorical annotation unit. This analysis yielded an overall Cohen’s kappa coefficient of 0.81. According to the widely used interpretation proposed by Landis and Koch[10], this value indicates strong inter-annotator agreement. This result suggests that the annotation guidelines were sufficiently clear to minimize ambiguity and that the dual-review process was effective in reducing individual subjective bias.

3.2 Templates

As illustrated in Figure S2, we have constructed a clear, coherent, and highly structured framework dedicated to systematically summarizing and extracting all experimental conditions relevant to on-surface reaction. This framework divides the core information of the entire synthesis process into four main components: Precursors, Reaction Stages, Final Stage, and Literature. Our focus lies in the detailed categorization of experimental conditions for both the precursors and the reaction/final stages, with particular emphasis on different activation types.

Specifically, the extraction of precursor information covers several key parameters, including the

international union of pure and applied chemistry (IUPAC) name, deposition method, substrate, morphology, and geometry. For the Intermediate/Product parameters, we extract their abbreviation, morphology, coverage, and reaction type. It is important to note that the reaction type is strictly extracted and labeled according to our predefined classification standards.

The activation types primarily include four major modes: thermal, tip-induced, light-induced, and other. All parameters required for these activation methods (such as temperature, time, bias voltage, etc.) have been explicitly predefined, ensuring consistency in data extraction. Finally, the extraction of literature information is primarily utilized for the effective tracking and sourcing of all experimental data, thereby guaranteeing the reliability and verifiability of the information.

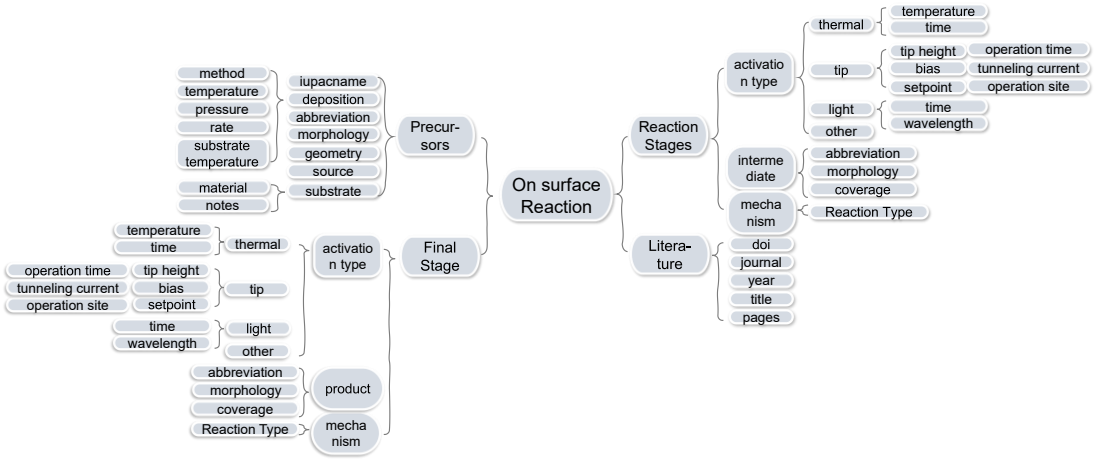

Figure S2. Template for extracting experimental parameters of on-surface reaction.

3.3 Metrics

Figure S3 illustrates the performance evaluation results of four commercial large language models (Deepseek, Qwen, Claude, and GPT) in the information extraction task of on-surface reaction, across four key categories: Precursors, Reaction Stages, Final Stage, and Literature.

Regarding the standard F1 metric, which measures exact match capability, all models demonstrated very high accuracy in extracting literature information. Claude led the pack with an F1 score of 0.83, closely followed by Deepseek and Qwen, both achieving 0.80. For the Stage-related metrics, model performance began to diverge significantly. In the parameter extraction for Reaction Stages, the Qwen model performed the best, exhibiting an F1 score of 0.73, which demonstrates its advantage in precisely extracting process-related details. However, the F1 scores for the Final Stage were generally lower across all models (with Qwen's 0.67 being the highest), clearly indicating that the extraction of final product information is the most challenging task among all categories.

It is noteworthy that the trend showed a subtle shift when we turned to the Bert F1 metric, which places a greater emphasis on semantic understanding. On the Bert F1 metric for Reaction Stages, the Claude model scored 0.67, slightly surpassing Qwen's 0.66. This suggests that although Qwen holds an advantage in literal, exact matching, Claude exhibited greater robustness in capturing the semantic similarity and contextual accuracy of the extracted reaction parameters.

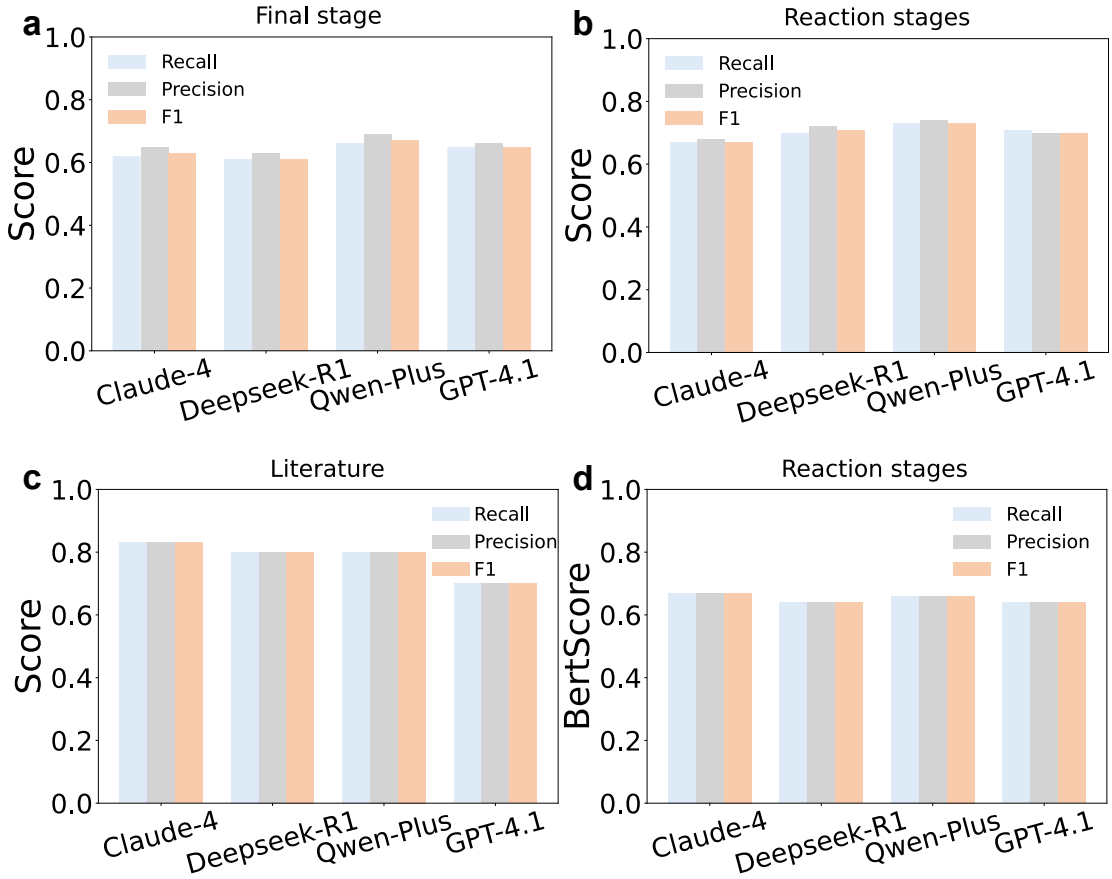

Figure S3 Recall, precision, and F1 scores of Claude-4, Deepseek-R1, Qwen-Plus, and GPT-4.1 in **a,** Final stage; **b,** Reaction stage. **c,** Literature **d,** Bert-recall, Bert-precision, and Bert-F1 scores in reaction stage.

3.4 WebPage

To systematically extract structured data from unstructured literatures, we developed a web-based annotation tool called “Surface Reaction Record Builder” to assist human experts in data extraction. (Figure S4)

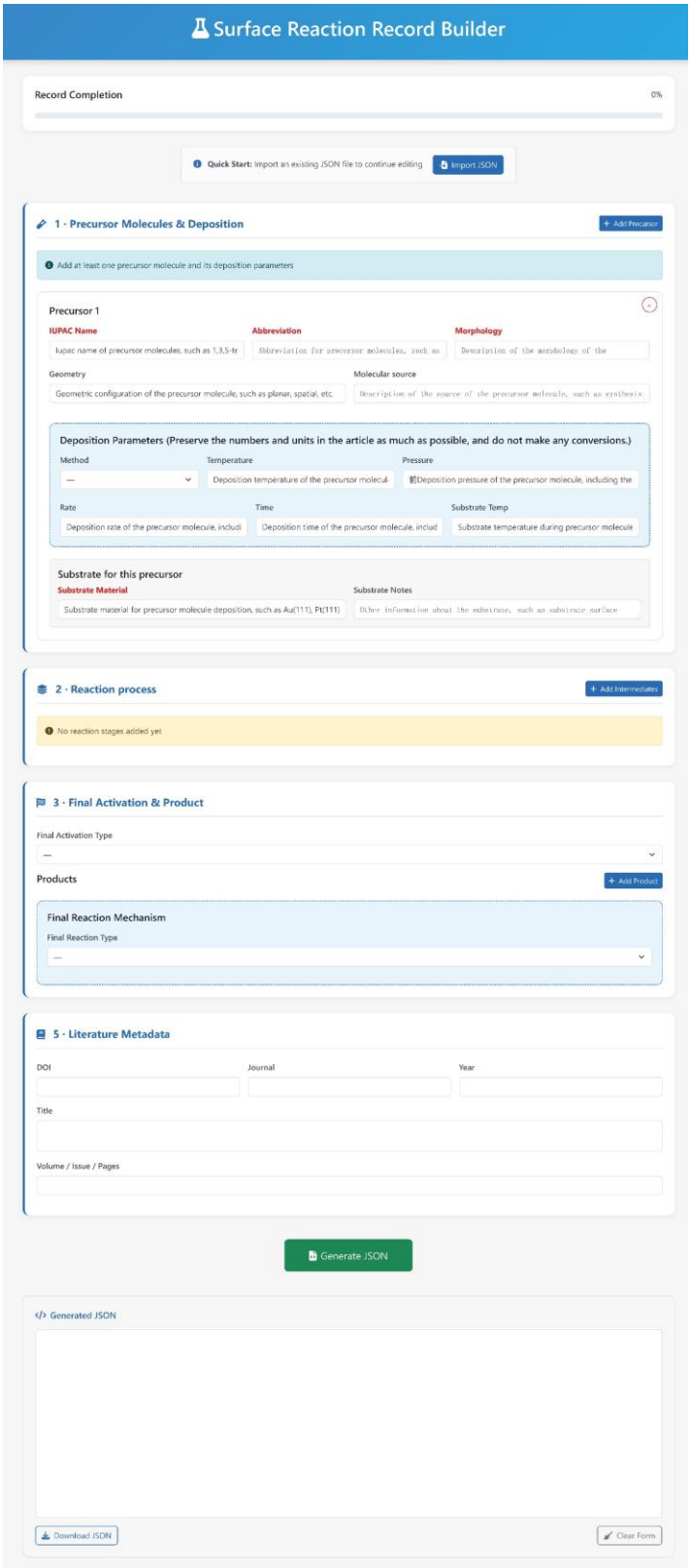

Figure S4 The web interface for data annotations and JSON file export.

4. Prompts

4.1 Prompt for Q&A pair generation

You are a research assistant tasked with generating a high-quality set of question-answer pairs based on a given literature text and its corresponding JSON extraction file. Please strictly adhere to the following requirements.

You must generate Q&A pairs distributed across the following three categories:

1. Synthetic and Retrosynthetic Questions: Focus on specific reaction pathways. For example, "What kind of reaction does precursor XXX undergo on surface XXX? What are the products?" or "How can product XXX be synthesized on surface XXX?"
2. Generalization Questions: Address fundamental surface science. Focus on adsorption thermodynamics, diffusion, self-assembly, electronic/geometric surface effects, and the underlying physics of reaction activation and selectivity. Questions should focus on the system itself (molecules using IUPAC names, reactions, surfaces, etc.) rather than using "this paper" as the subject.
3. Comprehensive Questions: Compare the current system with other surfaces/conditions or use general concepts to explain why a specific reaction behavior occurs. Questions should focus on the system itself (molecules using IUPAC names, reactions, surfaces, etc.) rather than using "this paper" as the subject.

Question Formulation: Independent and Complete: Each question must include necessary background information (experimental subject, stage, conditions) and should not rely on context from other questions.

Avoid Metadata: Do not include content unrelated to research findings, such as DOI, title, or authors.

Avoid Ambiguous References: Do not use vague references like "this paper," "this study," or "this."

Content Design

Must Cover Key Elements of the Article: Precursor characteristics, deposition conditions, initial adsorption behavior, intermediate structures, final covalent products, comparisons between different surfaces, and mechanistic insights.

Each question should be independently meaningful and reflect the core content of the article on its own.

Answer Formulation

Layered Explanation: Answers should be logically structured: state facts → explain reasons → provide inferences or implications.

Avoid Brief Answers: Answers must not be overly concise; they should include detailed explanations and reasoning chains.

Reasoning Field: Between the question and answer, a reasoning field must be included to demonstrate the reasoning or thought process. This may incorporate the model's own knowledge of chemistry/physics and is not limited to the content of the literature.

Natural Expression: Answers do not need to follow a rigid "fact/reason/significance" format but should be expressed in natural paragraphs.

Format Requirements

Output in JSONL format (one object per line), with the following fields:

{
"question": "...",
"reasoning": "...",
"answer": "..."
}

Preferred format: JSONL. Output each QA object on a separate line. If you return multiple items, DO NOT use a JSON array; just output multiple lines.

Quantity and Quality

Generate 5–10 question-answer pairs each time.

Each question-answer pair must be independently meaningful and able to summarize the core information of the article when read in isolation.

Length Requirement: Each question-answer pair (including question + reasoning + answer) must contain no fewer than 4096 tokens.

Note: In the Q&A generation pipeline, each generation task was conditioned jointly on two aligned inputs: the source article text and its corresponding matched JSON extraction. The prompt was designed to require self-contained and independently meaningful questions, to avoid metadata and ambiguous references, and to encourage coverage of multiple scientific aspects of each study, including precursor characteristics, deposition conditions, adsorption behavior, intermediate structures, final products, surface-dependent comparisons, and mechanistic insights. The generated outputs were required to follow a structured JSONL format with three predefined fields: question, reasoning, and answer.

To improve the robustness of the generated dataset, quality-control procedures were implemented at both the input and output levels. At the input level, only samples with matched text–JSON pairs were processed. At the output level, the generated results were subjected to multi-step JSON/JSONL validation, followed by field normalization to retain only the required structured entries. Malformed outputs were excluded from the final dataset and recorded through an error-logging procedure for traceability. The prompt was designed to elicit a range of QA forms, including synthesis-oriented, retrosynthesis-oriented, mechanistic, and general explanatory questions. Factual reliability was supported primarily through source-grounded generation based on the article text and the matched JSON extraction.

4.2 Prompt for training fine-tuned-LLM

Below are instructions describing the task, along with input to provide more context.

Please write a response that appropriately completes the request.

Before generating your final response, please think through the steps and ensure your answer is logical and scientifically rigorous.

Instructions:

You are a seasoned expert in surface science and physical chemistry, specializing in surface physics, chemical reactions, and the characterization of various materials. Please provide a detailed and scientifically rigorous answer to the question you provided.

Question:

{}

Answer:

<think>

{}

</think>
{}

5. Q&A evaluation

5.1 Q&A evaluation webpage

To objectively and quantitatively evaluate the performance of LLMs in professional domain Q&A, we have developed the “(Q/A Rating Panel)” web page shown in the figure. This web page is designed to assess the quality of model-generated question-answer pairs across multiple dimensions through the intervention of human experts. The evaluation panel is clearly divided into three main areas: the PDF Preview Area on the left, which supports importing PDF files containing original literature, allowing review experts to cross-reference the source document in real-time to ensure the factual accuracy of the model’s answer; the Q&A Content Area in the center, used to load the question-answer pair to be evaluated, including three core components: QUESTION, REASONING, and ANSWER, making it convenient for review experts to comprehensively examine the model's reasoning process and result; and the Rating Panel Area on the right, where human experts are required to provide a quantitative assessment from 1 point (Low) to 5 points (High) based on five key metrics for the current Q&A pair, We use percentage values in the assessment chart instead.. These five metrics are: Relevance, Agnosticism, Completeness, Accuracy, and Reasonableness. This evaluation framework facilitates our assessment of LLM Q&A results.

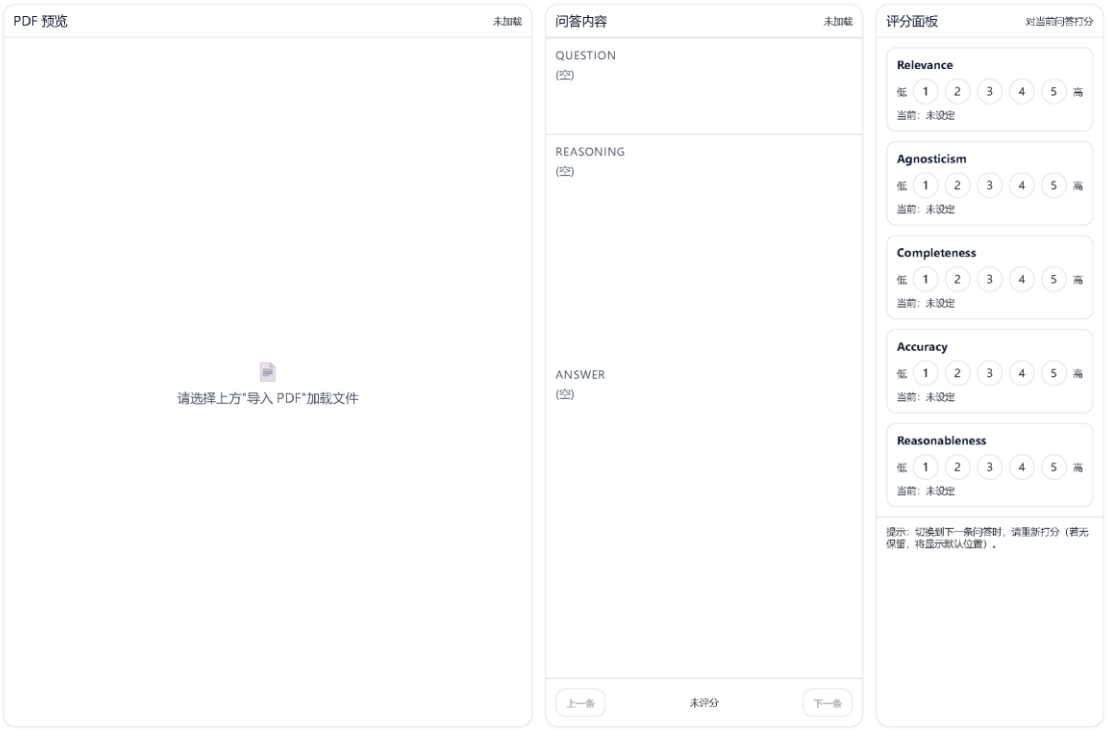

Figure S5 The web interface for QA evaluation by human experts.

5.2 Fine-tuned Q&A evaluation

To assess the sensitivity of the model’s internal representations to training data scale, we visualized

PCA and UMAP projections of embeddings and attention patterns across six dataset sizes, namely 0.2K, 0.5K, 1K, 3K, 5K, and 10K (Figures S6–S9). As illustrated in Figures S6–S9, the visualizations reveal a gradual transition from disordered to structured representations. At early training stages (0.2K–1K), data points are diffusely distributed without clear boundaries, indicating limited acquisition of domain specific structure. As the training dataset increases, a pronounced topological reorganization becomes apparent, with representations exhibiting more localized grouping and clearer structural patterns at the 10K scale. This trend indicates that sufficient training data are necessary for fine tuning to yield more organized and distinguishable latent representations of surface chemical concepts.

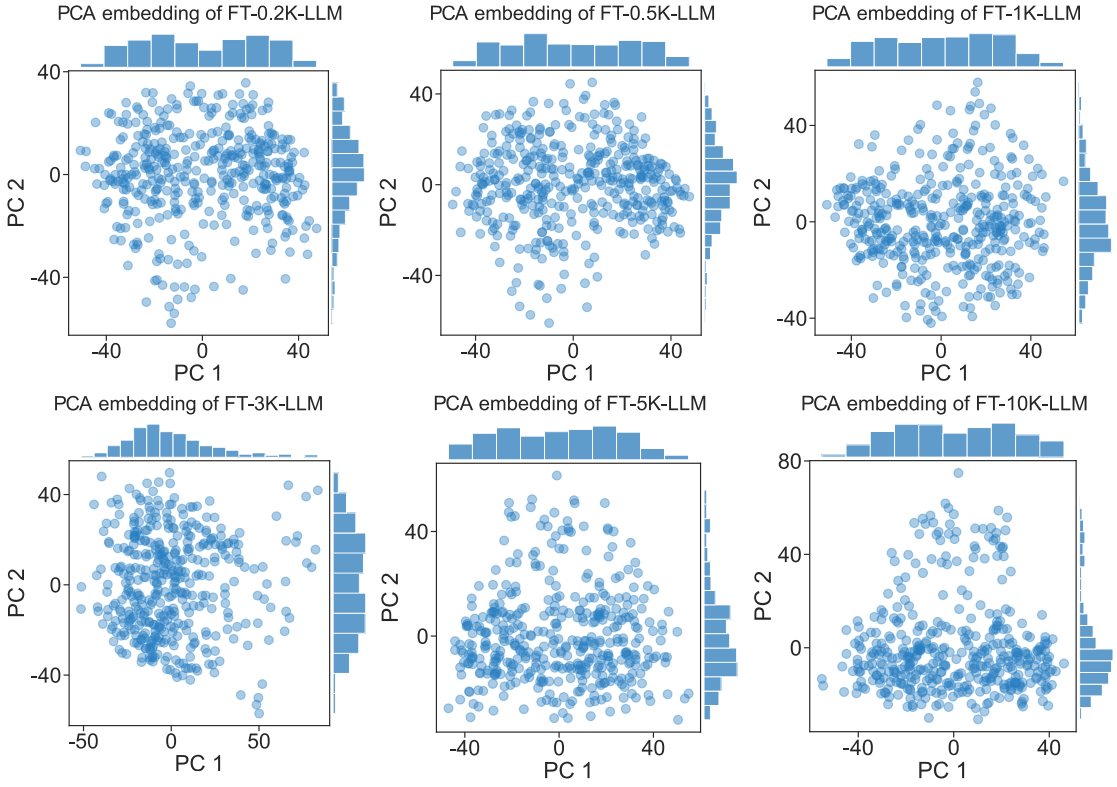

Figure S6 Principal component analysis of embeddings for six models of different dataset size.

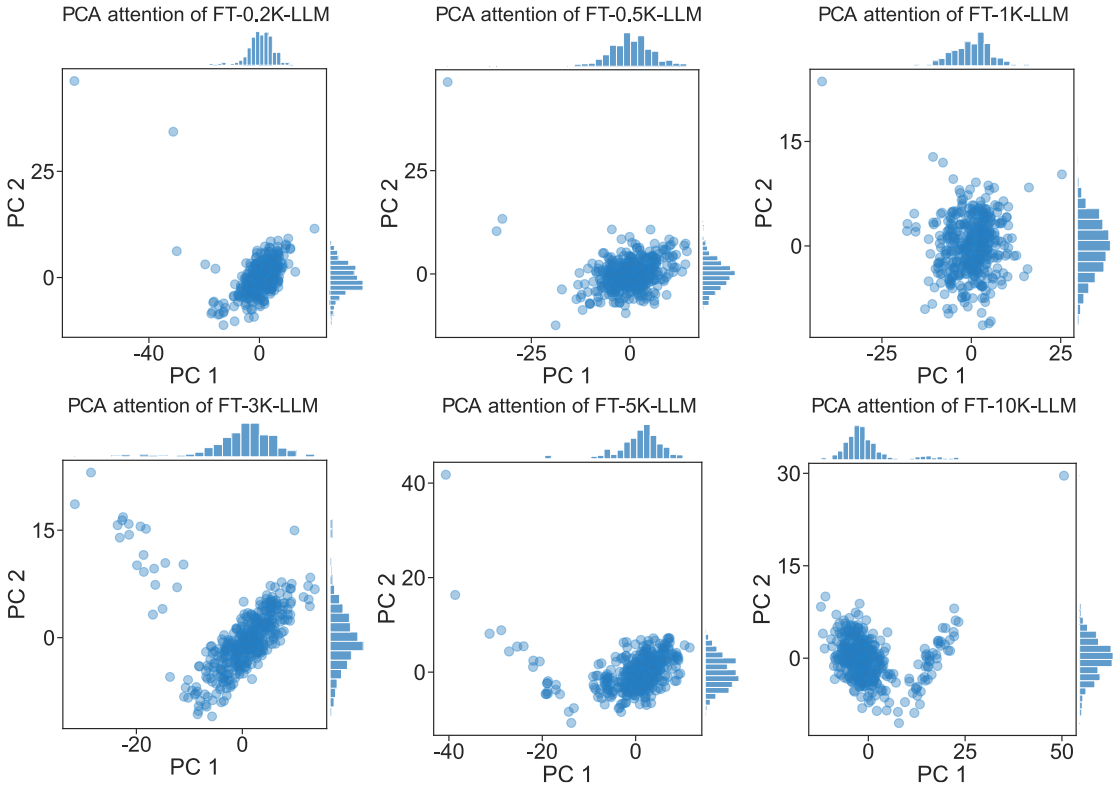

Figure S7 Principal component analysis of attention patterns for six models of different dataset size.

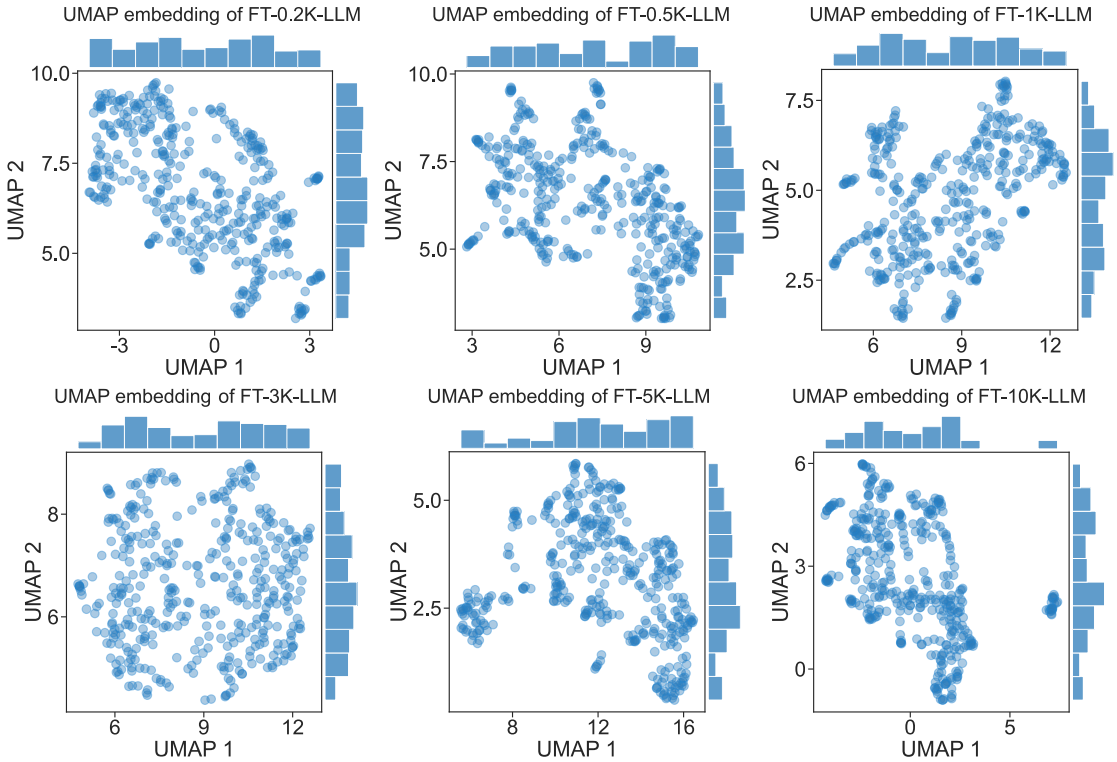

Figure S8 UMAP projections of embeddings for six models of different dataset size.

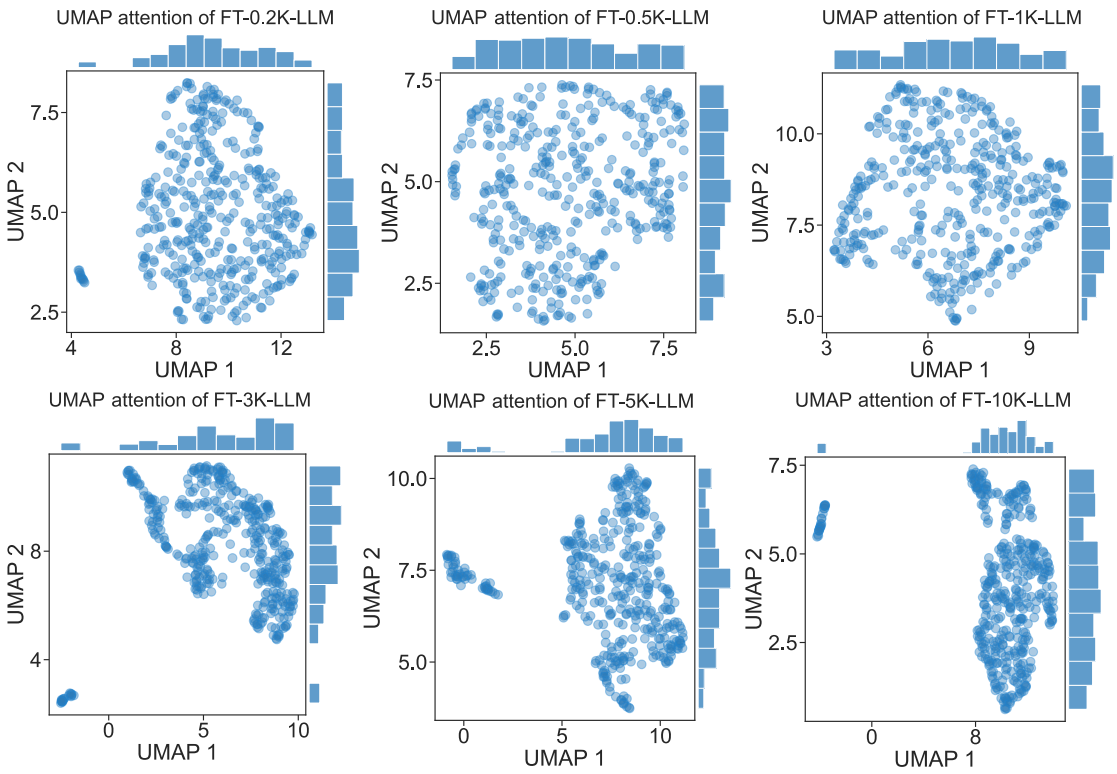

Figure S9 UMAP projections of attention patterns for six models of different dataset size.

5.3 Q&A categories evaluation

To deeply assess the performance distribution of the dual LLM across different knowledge tiers, we categorized the question answering tasks into general and synthesis-oriented questions. The evaluation was conducted using expert human scoring criteria, with domain specialists assigning scores across multiple qualitative dimensions. Figure S10 presents the probability density distributions of scores for general and synthesis Q&A across the evaluation metrics. Overall, performance on general questions consistently surpassed that on synthesis related questions across all metrics. For Accuracy and Reasonableness, score distributions for general questions are strongly concentrated in higher score regions, exhibiting narrower and more compact shapes. This indicates that the model provides stable, correct, and logically coherent responses when addressing conceptual definitions or generalized reaction mechanisms. In contrast, the score distributions for synthesis related questions are broader, with Accuracy in particular showing signs of increased dispersion. This reflects greater variability in model performance when handling tasks that require precise quantitative parameters and complex surface chemistry reasoning. For the Completeness metric, while general questions maintain a slight overall advantage, synthesis related questions display noticeably higher dispersion. This arises because synthesis-oriented questions require the model to explicitly identify all key experimental elements, including the substrate, specific temperature, and activation method. Any omission directly reduces the completeness score, whereas general questions allow a wider range of acceptable explanations.

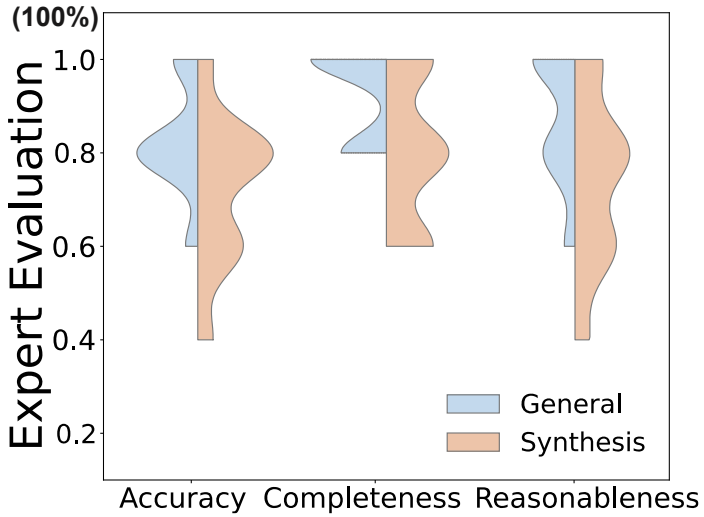

Figure S10 The differences in evaluation metrics and their distributions for two types of questions.

6. Retrieval-Augmented Generation

We developed a retrieval-augmented generation (RAG)-based question-answering framework for applications in surface chemistry. The response generation module is built on the OpenAI's ChatGPT (gpt-4.1), with the temperature set to 0.7 to balance generative diversity and response stability. In parallel, a dedicated router large language model (Router-LLM) is employed for semantic query routing. Its sole function is to classify each incoming query into one of two categories, namely Synthesis Q&A or General Q&A. This routing strategy directs domain-specific queries to the appropriate FAISS-indexed vector database and thereby reduces cross-domain interference during retrieval. To ensure stable and near-deterministic routing behavior, the Router-LLM operates at a low temperature of 0.1.

For text vectorization, we use the local Hugging Face embedding model all-MiniLM-L6-v2[11]. Separate vector databases are constructed for unstructured surface-chemistry texts and structured reaction-condition JSON documents according to task type. Dense vector indexing is implemented using FAISS, with incremental batch construction (batch size = 1000) and sharded persistence to support scalable document storage and retrieval. During inference, the Router-LLM first identifies the task category of the user query and activates the corresponding system prompt and vector database. A history-aware retriever is then invoked, in which an LLM reformulates the retrieval query on the basis of the dialogue context.

The reformulated query is subsequently processed by a hybrid retrieval module that combines a FAISS-based dense retriever with a BM25 keyword retriever, each configured with a top-k value of 10. The outputs of the two retrievers are integrated using an ensemble retriever with weights of 0.8 for vector similarity and 0.2 for lexical matching, thereby balancing semantic relevance against exact term overlap. Retrieved candidates are further refined using a local cross-encoder re-ranking model (BAAI/bge-reranker-large[12]), which reduces the candidate set to the top three documents. These documents, together with the system prompt and the user query, form the final RAG context used for answer generation.

The chat history is updated after each interaction, enabling history-aware retrieval and context-consistent multi-turn dialogue. This design supports context-sensitive and knowledge-grounded responses for both surface-chemistry and chemical-synthesis queries. In addition, for every

generated response, the system outputs a structured metadata record containing the source document ID, file path, and a content preview. These metadata enable manual verification of the evidentiary basis of individual claims and therefore enhance the transparency and trustworthiness of the system in scientific research settings.

7. Real-world experimental case

Debrominative coupling of 5,10,15,20-tetra(4-bromophenyl)porphine on Ag(111)

To provide prospective experimental support for the practical utility of the model, we examined the on-surface debrominative coupling of 5,10,15,20-tetra(4-bromophenyl)porphine (TPP-Br_4) on Ag(111). This system is a representative platform for studying surface-assisted Ullmann-type coupling, low-dimensional nanostructure formation, and the construction of covalently linked surface-confined architectures. A key challenge in this reaction is the identification of an appropriate thermal activation window that enables the transition from precursor deposition to debromination and subsequent intermolecular covalent coupling.

Importantly, the specific temperature predictions described below were not taken from prior literature reports on this system. Instead, they were generated prospectively by the fine-tuned LLM assistant before the experiment was performed. Based on its analysis of the TPP-Br_4/Ag(111) system, the model suggested that, owing to the catalytic activity of Ag(111), partial C–Br bond cleavage could occur between room temperature and 100 °C, accompanied by dissociation of Br atoms on the surface. The model further proposed that annealing in the range of approximately 150−200 °C would provide a favorable kinetic window for intermolecular coupling, in which thermally activated diffusion and radical recombination would become sufficiently accessible to produce an extended covalent network.

These model-generated predictions were then tested experimentally by high-resolution scanning tunneling microscopy (STM). After annealing at 100 °C, STM images showed that the TPP-Br_4 molecules remained largely dispersed on the surface. At the same time, bright protrusions attributed to dissociated Br atoms were observed around the molecular periphery and near step edges, consistent with the onset of C−Br bond cleavage (Figure S11). When the sample was further annealed to 200 °C, the surface morphology changed markedly. The STM images revealed the formation of an ordered two-dimensional covalent network, accompanied by the disappearance of the Br-related bright protrusions observed at lower temperature. Under the imaging conditions used here(V_{bias} = -1.5 V, I_t = 50 pA), isolated precursor molecules were no longer the dominant surface species after annealing at 200 °C, and the surface was instead characterized by the coupled network structure.

Overall, the experimentally observed structural evolution, from dispersed, partially debrominated precursor species at 100 °C to an extended covalent network at 200 °C, was in good agreement with the temperature window predicted in advance by the model. This result therefore provides prospective, experimentally grounded support for the model's ability to suggest plausible activation conditions and qualitative reaction outcomes in a representative on-surface coupling system.

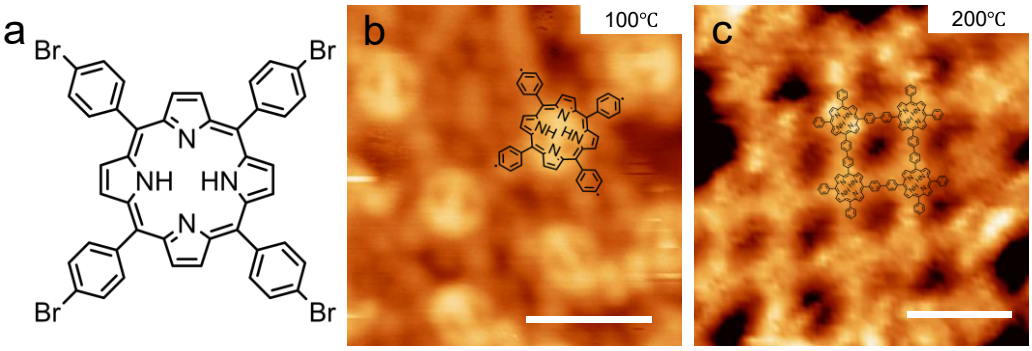

Figure S11. (a) Chemical Structure of TPP-Br_4. STM characterization of TPP-Br_4 on Ag(111) after annealing at (b) 100 °C and (c) 200 °C. At 100 °C, the molecules remain largely dispersed, and dissociated Br atoms are observed as bright protrusions, indicating the onset of debromination. At 200 °C, the precursor is converted into a two-dimensional covalent network. Imaging conditions: V_{bias} = -1.5 V, I_t = 50 pA. Scale bars: (b) 2 nm and (c) 3 nm.

8. References

1. 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, *Journal of machine learning research*, 2020, **21**, 1-67.
2. B. Cao, J. Xiong, J. Ma, Y. Tian, Y. Hu, M. He, L. Zhang, J. Wang, J. Hui and L. Liu, Bgolearn: a Unified Bayesian Optimization Framework for Accelerating Materials Discovery, *arXiv preprint arXiv:2601.06820*, 2026.
3. J. Opitz and S. Burst, Macro f1 and macro f1, *arXiv preprint arXiv:1911.03347*, 2019.
4. 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 preprint arXiv:2407.21783*, 2024.
5. S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul and B. Bossan, Peft: State-of-the-art parameter-efficient fine-tuning methods, 2022.
6. T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, *ArXiv*, 2023, **abs/2305.14314**.
7. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, Bertscore: Evaluating text generation with bert, *arXiv preprint arXiv:1904.09675*, 2019.
8. J. Castro‐Esteban, F. Albrecht, S. Fatayer, D. Pérez, L. Gross and D. Peña, An on‐surface Diels–Alder reaction, *Angewandte Chemie*, 2021, **133**, 26550-26554.
9. J. Li, P. Brandimarte, M. Vilas-Varela, N. Merino-Diez, C. Moreno, A. Mugarza, J. S. Mollejo, D. Sanchez-Portal, D. Garcia de Oteyza and M. Corso, Band depopulation of graphene nanoribbons induced by chemical gating with amino groups, *ACS nano*, 2020, **14**, 1895-1901.
10. J. R. Landis and G. G. Koch, The measurement of observer agreement for categorical data, *biometrics*, 1977, 159-174.
11. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2.
12. S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian and J.-Y. Nie, *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*. 2024, 641-649.