

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 CATALYSTBENCH: A COMPREHENSIVE MULTI-TASK BENCHMARK FOR ADVANCING LANGUAGE MODELS IN CATALYSIS SCIENCE

Anonymous authors

Paper under double-blind review

## ABSTRACT

The discovery of novel catalytic materials is a cornerstone of chemical engineering and sustainable energy, yet it remains a complex, knowledge-intensive process. While Large Language Models (LLMs) have demonstrated remarkable potential in various scientific domains, their application to catalysis is hindered by the lack of specialized, multi-dimensional benchmarks to guide their development and evaluation. To bridge the critical gap, we introduce CatalystBench, a comprehensive and challenging benchmark meticulously constructed from scientific literature and public datasets, specifically designed to assess the capabilities of LLMs in the nuanced domain of catalyst design. The tasks covered by this benchmark dataset encompass the entire closed-loop process of catalyst development, including reading comprehension, experimental analysis and scheme reasoning. Based on this benchmark, we propose a Multi-head Full-task (MFT) domain-specific fine-tuning method that employs coupling task-specific output heads. We systematically compare with other three distinct fine-tuning strategies: Single-Task (ST), Full-Task (FT) and Multi-head Single-Task (MST). The extensive experiments demonstrate that the MFT strategy consistently achieves the most substantial performance improvements across all tasks, underscoring the effectiveness of explicit multi-task architectures in complex scientific reasoning. The resulting CatalystLLM significantly outperforms a wide array of state-of-the-art open-source and closed-source models on CatalystBench. We will publicly release both the CatalystBench benchmark and the CatalystLLM model, providing the community with a robust evaluation framework and a powerful new tool to accelerate AI-driven research in catalytic materials science.

## 1 INTRODUCTION

The advancement of catalysis is a cornerstone of modern science and industry, pivotal to achieving a sustainable future Swanson et al. (2025); Fu et al. (2025); Zhang et al. (2025b); Song et al. (2025). In general, the property of catalysts depends on the complex interplay of composition, crystal structure, surface active sites and regulation strategies Zhu et al. (2017); Chen et al. (2020). Designing new catalysts is therefore a formidable challenge: 1) **Vast candidate space**. The combination of multiple chemical elements, possible crystal phases and surface terminations leads to an astronomical search space, far beyond the reach of exhaustive experimental screening; 2) **Separated knowledge sources**. High-fidelity theoretical datasets derived from density functional theory (DFT) calculations capture key descriptors like adsorption energies and electronic properties Chanussot et al. (2021); Winther et al. (2019), while experimental literature documents synthesis conditions, stability and measured catalytic activities Chen et al. (2024). These two streams are often siloed, with no unified framework linking them for systematic analysis Zhang et al. (2025a). 3) **Lack of realistic evaluation frameworks**. Although AI methods have been applied to materials science, catalysis still lacks a benchmark that reflects the stepwise workflow of catalyst design. The model development remains fragmented and benchmarking across approaches is inconsistent.

Large Language Models (LLMs) have triggered a paradigm shift in “AI for Science” Abramson et al. (2024); Merchant et al. (2023); Szymanski et al. (2023), achieving notable success in domains such as bioengineering Luo et al. (2022); Edwards et al. (2021); Waisberg et al. (2024); Lamb et al.

(2024), materials discovery Zhang et al. (2024b); Kristiadi et al. (2024) and industrial process optimization Yang et al. (2023); Saka et al. (2024). However, their deployment in catalytic materials science reveals a critical gap Wang et al. (2025b): existing benchmarks rarely reflect the multi-modal, multi-stage workflows that characterize real catalyst R&D, where precise numerical regression, categorical decision making and open-ended mechanistic reasoning coexist within a single process. We compare the limitations of current benchmarks in the material field across multiple dimensions and conduct a detailed analysis in Appendix A.3. Models trained on homogeneous tasks or unified output formats struggle to preserve accuracy across this spectrum, often suffering loss-landscape interference between qualitatively different objectives.

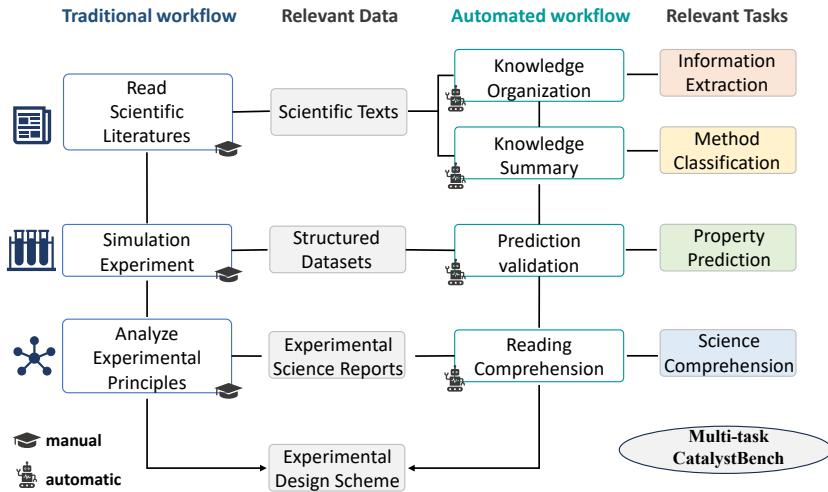


Figure 1: Data-driven model for scientific research processes in the materials field.

To bridge this knowledge gap, we present CatalystBench, the first multi-task benchmark tailored for catalysis science that explicitly unifies high-quality theoretical datasets and curated experimental literature into a structured Q&A format. Beyond aggregating and filtering data, we perform domain-specific curation and transformation so that each evaluation task directly corresponds to a stage in the catalyst design lifecycle, as shown in Fig 1. Given that these tasks span text-based reasoning and numerical property prediction, we propose a Multi-head Full-task Fine-tuning (MFT) strategy. It's an architecture adaptation where classification heads, regression heads and the language modeling head are trained in parallel but decoupled from one another. This architectural decoupling is motivated by the distinctive composition of catalyst design tasks, which uniquely combine high-precision numerical prediction, categorical judgement and open-ended scientific reasoning within a single workflow. While multi-head architectures have been explored in other areas of multi-task learning, this work constitutes the first systematic empirical validation and ablation in the catalysis science domain. We then evaluate our domain-adapted CatalystLLM against strong general and scientific-domain LLM baselines on CatalystBench, finding that it achieves state-of-the-art performance across most task categories. This not only validates the effectiveness of our benchmark and adaptation strategy, but also yields insights into which aspects of catalyst-related reasoning remain challenging for current LLMs. We highlight the primary contributions of this paper as follows:

- We introduce **CatalystBench**, the first multi-task benchmark for catalysis that integrates theoretical simulation data and experimental literature into a unified and task-oriented format reflecting the actual catalyst design workflow.
- We develop a **domain-specific adaptation strategy**. The Multi-head Full-task Fine-tuning (MFT) approach addresses the heterogeneous nature of catalyst-related tasks by decoupling numerical prediction and language generation within the model architecture. We systematically compare four different fine-tuning strategies and demonstrate the superiority of the MFT strategy.
- We conduct a comprehensive evaluation of CatalystLLM against **multiple strong baseline models**, demonstrating its SOTA performance on our proposed benchmark. Our experi-

108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161

ments and detailed analysis not only validate the effectiveness of our tuning strategies but also yield valuable insights into the current strengths and weaknesses of LLMs in catalysis.

## 2 RELATED WORKS

### 2.1 BENCHMARKS IN CHEMISTRY AND MATERIAL SCIENCE

The increasing application of LLMs in scientific discovery has led to the emergence of numerous benchmarks in chemistry and materials science. Initial efforts focused on assessing domain-specific knowledge. For instance, ChemBench evaluate the ability of LLMs on a wide range of chemical knowledge by constructing multiple-choice questions derived from textbooks and expert knowledgeZhang et al. (2024a); Mirza et al. (2024). Similarly, MaScQA Zaki et al. (2024) assesses the understanding of core materials science concepts using graduate-level exam questions, while SCIBENCH Wang et al. (2023b) tests college-level scientific problem-solving abilities.

More recently, benchmarks have evolved to probe more specialized and complex abilities. Chem-CoTBench Li et al. (2025) was specifically designed to evaluate the step-by-step chemical reasoning of LLMs, moving beyond simple factual recall. For predictive tasks, LLM4Mat-Bench Rubungo et al. (2025) provides a comprehensive suite for materials property prediction. Furthermore, Mat-Tools Liu et al. (2025) uniquely evaluates the ability of LLMs to interact with and utilize materials science software libraries. However, existing chemistry-focused LLM benchmarks either emphasize theoretical molecular-level understanding or constrained problem solving and rarely integrate the complementary knowledge from experimental scientific literature Guo et al. (2023); Li et al. (2024); Zhang et al. (2024a); Xie et al. (2024). This omission is critical for catalysis, where real catalyst surfaces often deviate substantially from the idealized models assumed in theoretical calculations Wang et al. (2023a) and design decisions depend heavily on synthesis conditions, stability data and structure-activity trends reported experimentally. Table 4 compares representative existing chemistry benchmarks with CatalystBench. CatalystBench addresses these gaps by combining high-fidelity theoretical datasets with curated experimental literature in a unified, task-oriented evaluation framework.

### 2.2 DOMAIN-SPECIFIC LLMs FOR CHEMISTRY AND MATERIALS SCIENCE

Beyond evaluating general models, a significant research direction involves creating specialized language models for chemistry and materials science through domain-specific fine-tuning or continued pre-training. Early efforts in this area often involved BERT-style encoder models Chithrananda et al. (2020); Ock et al. (2023); Trewartha et al. (2022); Zhao et al. (2024). More recent studies have shifted toward adaptive adjustments to large-scale generative LLMs. For example, ChemLLM Zhang et al. (2024a) is instruction-tuned on a large set of templated Q&A pairs to handle diverse chemical tasks conversationally and in the materials domain, DARWIN 1.5 Xie et al. (2024) adopts a multi-stage training strategy combining Q&A fine-tuning with multi-task learning to internalize complex materials knowledge. Other generative models include ChemFormer Irwin et al. (2022) for reaction prediction and CrystaLLM Antunes et al. (2024) for generating novel crystal structures.

Recently, several high-impact works further expand the landscape of chemistry LLMs. Llasmol Yu et al. (2024) leverages a high-quality instruction tuning dataset to significantly enhance chemical reasoning capabilities. Translation between Molecules and Natural Language Edwards et al. (2022) explores bidirectional mapping between molecular structures and textual descriptions, highlighting LLMs' potential for molecular communication. Instructmol Cao et al. (2023a) integrates multimodal data sources to build a versatile, reliable molecular assistant for drug discovery.

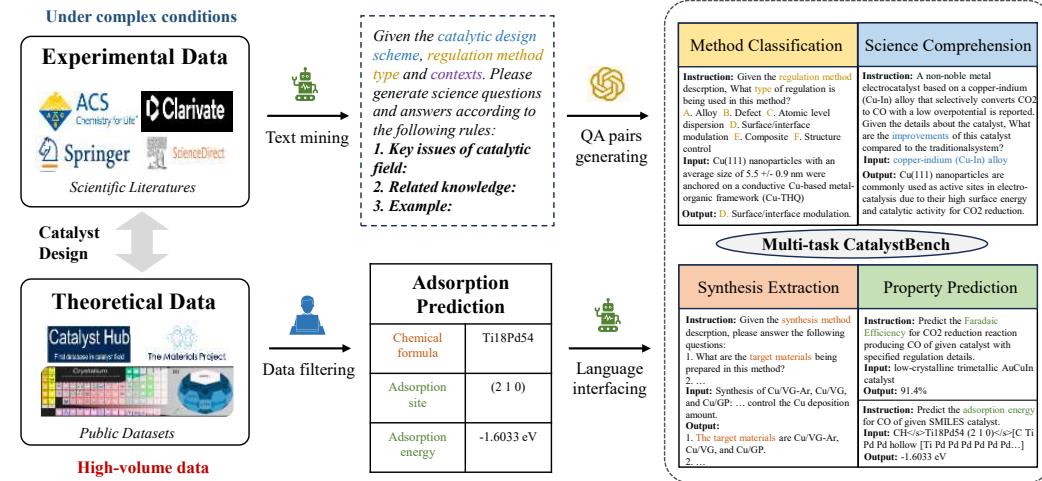
However, current domain-specific LLMs are typically optimized for single-task formats or unified output styles, focusing on either symbolic reasoning or purely structural prediction. In contrast, complex scientific scenarios such as catalyst design require the combined ability to comprehend domain literature, reason mechanistically and perform high-precision numerical prediction within a single workflow. The comparison results in Table 5 reveal that most domain-specific large models show limited improvement in numerical prediction tasks, further highlighting the challenges of current multi-task fine-tuning approaches. CatalystLLM tackles this challenge through a multi-head architecture explicitly adapted to the mixed-task nature of catalysis, enabling parallel handling of qualitative and quantitative tasks without mutual interference.

### 162 3 THE CATALYSTBENCH BENCHMARK

#### 164 3.1 OVERVIEW OF CATALYSTBENCH

166 In order to explore the abilities of LLMs in the field of materials science, we concentrate on three  
 167 fundamental capabilities: **Understanding**, **Reasoning** and **Explaining**. Fig 1 illustrates the data-  
 168 driven paradigm shift in the research process within the field of catalytic materials. In response to  
 169 this trend, the CatalystBench dataset combines theoretical simulations with scientific experimental  
 170 data to construct a series of tasks covering the entire process.

#### 172 3.2 CATALYSTBENCH CONSTRUCTION

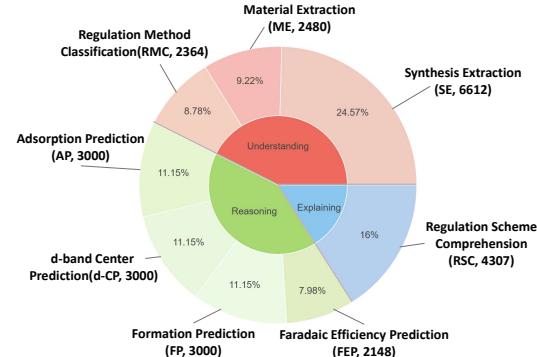


190 Figure 2: The process for CatalystBench construction and examples demonstration.

#### 192 3.2.1 DATA COLLECTION AND ORGANIZATION

194 Fig 2 shows the construction process and spe-  
 195 cific examples of CatalystBench. For theore-  
 196 tical simulation data, we select and integrate 8  
 197 publicly available catalytic datasets (Appendix  
 198 A) to quantitatively evaluate the **Understand-  
 199 ing** and **Reasoning** capabilities of LLMs. We  
 200 filter out key features from descriptors re-  
 201 lated to catalytic properties in public datasets,  
 202 such as SMILES strings. We design a set of  
 203 prompt templates to convert these features and  
 204 attributes into task-specified natural language  
 205 sentences. This leads to instructions that com-  
 206 bine inputs and expected outputs. For exam-  
 207 ple, the instruction of property prediction like:  
 208 'Predict the adsorption energy for CO of given  
 209 SMILES catalyst.' with input 'Ti<sub>18</sub>Pd<sub>54</sub>' and  
 our model should give the output '-1.6033 eV', which can be converted to a numeric value.

210 Beyond basic understanding and reasoning, we further explore the model's higher-order **Explain-  
 211 ing** capabilities, including deep understanding of catalytic regulation schemes documented in the  
 212 literature and accurate grasp of macro-level trends in specific catalytic fields. In this step, we ex-  
 213 tract catalyst regulation and synthesis schemes from scientific literature based on existing catalytic  
 214 knowledge frameworks. We use SciQAG Wan et al. (2024) to guide GPT-4o in interpreting a large  
 215 volume of regulation scheme texts, converting them into high-quality Q&A pairs. Then domain  
 experts perform annotation and filtering. The method details are described in Appendix A.4.



216 Figure 3: The statistics of CatalystBench Multi-  
 217 Task Data.

216 3.2.2 CATALYST TASKS  
217

218 The proportion of these three major categories is illustrated in the Fig 3. We examine these capa-  
219 bilities by constructing 8 diverse and broadly acknowledged practical catalytic material tasks. The  
220 dataset size for each task in CatalystBench is smaller than that of general benchmarks, reflecting  
221 a common constraint in catalytic research: high-fidelity data acquisition is domain-specific and re-  
222 quires substantial resources. To avoid data imbalance across tasks, we construct the entire dataset  
223 based on the quantity of experimental data, prioritizing task breadth to ensure comprehensive cov-  
224 erage of the entire process of the domain of catalyst design.

225 3.3 QUALITY ASSURANCE  
226

227 To ensure the reliability and reproducibility of CatalystBench, we adopt a multi-tiered quality assur-  
228 ance workflow across both theoretically simulated datasets and Q&A pairs generated from experi-  
229 mental literature. Automated validation identifies and removes invalid or duplicate entries, followed  
230 by expert review of a representative subset. For literature-derived Q&A, a rule-based filtering pro-  
231 cess eliminates context questions, while multi-stage generation improves semantic accuracy and  
232 diversity. The complete verification procedures and error samples are provided in Appendix A.5.

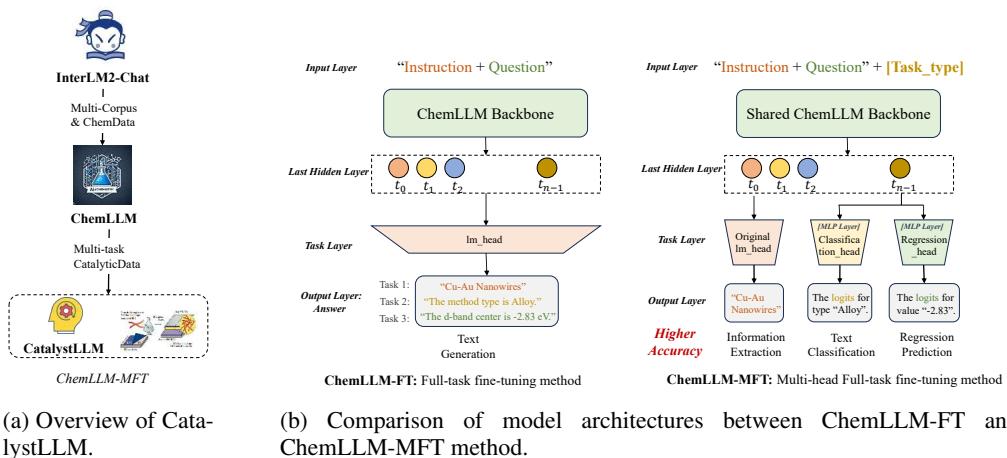
## 233 4 METHODOLOGY

234 4.1 BASE LANGUAGE MODEL  
235

236 We compare the performance of the current open-source models on 3 representative tasks of  
237 CatalystBench-full. From Table 5, ChemLLM Zhang et al. (2024a) has bridged this gap through  
238 domain-specific instruction fine-tuning on ChemData, infusing the model with a solid chemical  
239 foundation. Therefore, we choose ChemLLM-7B as the core base model for multi-task fine-tuning.

240 4.2 FINE-TUNING STRATEGIES  
241

242 Our goal is to design a fine-tuning architecture that simultaneously harnesses cross-task synergies  
243 and avoids performance degradation due to task heterogeneity. Fig 4a illustrates the 4 paradigms  
244 investigated, with the Multi-head Full-task(MFT) strategy as our proposed solution.



263 Figure 4: Overview of fine-tuning strategies and final CatalystLLM

264 1) **Single-task:** We fine-tune the LLMs using a training set of each specific task, which allows us to  
265 assess the models' ability to adapt to individual tasks and establish a baseline for a fair comparison.  
266 We obtain a fine-tuned model for each task in this setup, named 'ChemLLM-ST'.  
267

268 2) **Multi-head Single-task:** Considering the inherent differences in the attributes of different tasks,  
269 we adopt a shared LLM backbone and three task-specific prediction head structures. We obtain a  
270 fine-tuned model for each task in this setup, named 'ChemLLM-MST'.

270 3) **Full-task:** we fine-tune the LLMs using a mixture of all training datasets from 8 tasks, as shown  
 271 in Fig 4b. The single instruction-following generative head is trained jointly on all 8 CatalystBench  
 272 tasks. The single instruction follows the generation header (original *lm\_head*) during joint training  
 273 across all 8 CatalystBench tasks, with all task outputs routed to a unified generation decoder. We  
 274 obtain one fine-tuned model to perform all the tasks in this setup, named ‘ChemLLM-FT’.

275 4) **Multi-head Full-task:** The MFT method extends FT method by decoupling the output space. As  
 276 shown in Fig 4b, the task type token  $\tau$  is appended to the end of the input sequence, where  $\tau$  directly  
 277 controls which dedicated prediction head is activated after the final hidden representation  $\mathbf{h}[t_{n-1}]$ .  
 278 Here,  $t_{n-1}$  denotes the last non-padding token position in the encoded input, while  $\mathbf{h}[t_{n-1}]$  represents  
 279 the corresponding context embedding generated by the shared ChemLLM backbone Transformer  
 280 network. For classification tasks,  $\mathbf{h}[t_{n-1}]$  is fed into a task-specific MLP classification head, produc-  
 281 ing a probability distribution over predefined categories. For regression tasks, the same  $\mathbf{h}[t_{n-1}]$  input  
 282 is processed by a regression head optimized for mean squared error, generating a single continuous  
 283 value. In the original generation task, decoding proceeds directly via the original *lm\_head* without  
 284 architectural modifications, enabling the model to generate domain-specific explanatory text.

285 By training the shared backbone jointly across all tasks while isolating output modalities, MFT  
 286 method alleviates the interference between loss landscapes of qualitatively different tasks. In prac-  
 287 tice, we balance their contributions via a weighted composite loss and compare the impact of loss-  
 288 function in Appendix B.6. This approach leverages cross-domain feature learning that is critical in  
 289 catalytic material research, while preserving the precision of task-specific inference.

## 291 5 EXPERIMENTAL SETUP

### 293 5.1 BASELINE MODELS

295 For the sake of fairness, We select 15% of the data for each task in CatalystBench as the test  
 296 set, according to catalyst material type. Then we compare the performance of different LLMs on  
 297 CatalystBench-test, including open-source models of comparable scale, such as LLaMA-2 Touvron  
 298 et al. (2023), Mistral Jiang et al. (2024), ChatGLM GLM et al. (2024) and Qwen3-8B Bai et al.  
 299 (2023), as well as closed-source models with strong instruction-following capabilities, such as GPT-  
 300 3.5 Ye et al. (2023), GPT-4 Achiam et al. (2023) and deepseek. Additionally, for regression tasks  
 301 involving the prediction of material properties, we introduce comparisons with competitive ML  
 302 algorithms. We select some SOTA baseline models, such as CatBERTa Ock et al. (2023) and GAP-  
 303 CatBERTa Ock et al. (2024) from OC20 dataset and ML algorithm such as GPTchem Jablonka et al.  
 304 (2024). For certain tasks, we directly use the ML results from the original dataset papers Gao et al.  
 305 (2023). We provide a detailed introduction of these baseline models in Appendix B.

### 306 5.2 FINE-TUNING SETUP

308 To achieve efficient fine-tuning, we adopt Low-Rank Adaptation (LoRA) to reduce the computa-  
 309 tional cost with a rank of 8, a scale factor of 16.0 and a dropout rate of 0.1. We fine-tune all linear  
 310 modules. The training uses the AdamW optimizer with a learning rate of 5e-5, combined with a  
 311 linear decay scheduler with warm-up. Additionally, we further enhance training speed and model  
 312 robustness through techniques such as bf16 mixed precision and Flash Attention-2.

### 314 5.3 EVALUATION PROTOCOL

316 We employ a suite of tailored evaluation metrics. For the Text Classification and Information Ex-  
 317 traction tasks, we report both Accuracy and the balanced F1-score to account for potential class  
 318 imbalances. For the Regression tasks, we use MAE and  $R^2$  score to measure predictive accuracy.

319 Given the non-factual features of the semantic Q&A task, we develop a multi-dimensional evaluation  
 320 protocol. This protocol begins by calculating the STS score, which measures the sentence-level  
 321 semantic similarity between the generated answer and the reference answer. To assess the domain  
 322 expertise of the answers, we employ a model-based evaluation framework M. Bran et al. (2024)  
 323 where both gpt-4o and deepseek-r1 are prompted to score the generated answers on a scale from 1-10  
 across three criteria: reasonableness, accuracy and Usability. Finally, we conduct human evaluation

Model	ME		SE		RMC		FEP		AP		d-CP		FP	
	ACC	F1	ACC	F1	ACC	F1	R2	MAE	R2	MAE	R2	MAE	R2	MAE
<b>Closed-source LLMs</b>														
claude-3	0.93	0.95	0.84	0.90	0.58	0.72	0.30	6.73	0.37	3.58	0.33	4.21	0.39	5.24
Gemini-2.5	<b>0.97</b>	0.98	0.87	0.91	0.66	0.80	0.29	3.19	0.34	4.15	0.36	5.78	0.40	4.98
gpt-3.5	0.89	0.93	0.78	0.86	0.49	0.62	0.28	5.25	0.35	4.20	0.34	4.50	0.38	5.41
gpt-4.0	0.96	0.98	0.88	0.92	0.69	0.81	0.36	3.10	0.43	3.05	0.41	3.35	0.46	4.02
gpt-4.1	<b>0.98</b>	<b>0.99</b>	<b>0.91</b>	<b>0.95</b>	<u>0.75</u>	<u>0.85</u>	<u>0.56</u>	<u>2.51</u>	0.61	2.40	<u>0.59</u>	<u>2.15</u>	0.65	2.68
<b>Open-source LLMs</b>														
deepseek-v3	0.92	0.95	0.83	0.89	0.61	0.74	0.33	4.20	0.40	3.55	0.38	3.85	0.43	4.45
LLaMA2-7B	0.83	0.86	0.73	0.79	0.39	0.50	0.25	4.47	0.24	4.63	0.34	3.79	0.35	4.12
Qwen3-8B	0.86	0.91	0.74	0.83	0.42	0.54	0.26	4.59	0.31	3.85	0.32	3.49	0.36	4.26
Mistral-7B	0.88	0.92	0.76	0.84	0.44	0.57	0.29	3.09	0.37	3.49	0.36	4.14	0.39	5.15
ChatGLM3-6B	0.84	0.90	0.74	0.83	0.45	0.57	0.33	4.87	0.39	3.07	0.37	3.48	0.41	4.38
ChemLLM	0.93	0.95	0.79	0.88	0.52	0.66	0.45	2.80	<u>0.63</u>	<u>2.05</u>	0.54	2.36	0.64	2.75
Darwin1.5	0.91	0.94	0.79	0.89	0.50	0.64	0.44	2.81	0.59	3.13	0.54	2.42	0.68	2.01
<b>CatalystLLM</b>	<b>0.98</b>	<b>0.99</b>	0.89	0.94	<b>0.81</b>	<b>0.89</b>	<u>0.73</u>	<u>1.72</u>	<b>0.81</b>	<b>1.24</b>	<b>0.73</b>	<b>1.49</b>	<b>0.80</b>	<b>1.34</b>

Table 1: The performance of CatalystLLM with other LLMs on tasks with factual answers. The best model is in bold font and the second-best is underlined.

on a random subset of the test dataset, with domain experts providing the final judgment on the model’s practical applicability and domain-specific correctness(Details in Appendix C).

## 6 RESULTS AND ANALYSIS

### 6.1 COMPARISON OF FINE-TUNING STRATEGIES

After fine-tuning the ChemLLM-7B model, we evaluate their performance on CatalystBench-test for each task. In Fig 5, we compare task performance to identify how different fine-tuning strategies influence the results. The results show that the MFT strategy achieves the best performance, with the highest average performance improvement rate relative to the ST baseline, reaching 12.44%. This indicates that combining multi-task learning with task-specific output heads is more effective than applying either technique in isolation. Among these, MT fine-tuning resulted in an average improvement of 9.24%, while task-specific output head design achieve an average improvement of only 5.13%. This highlights that coupling architectures help tailor the model’s outputs for a specific task format, but the impact is limited without multi-task training. When all tasks share underlying semantic features, the model can leverage these features to learn shared and transferable representations. We name the final ChemLLM-MFT model as CatalystLLM, an open-source foundational LLM specifically designed for catalytic science. We provide a detailed analysis of the suitability of MFT architecture for catalytic process tasks in Appendix B.1.

### 6.2 MAIN BENCHMARK RESULTS

We evaluate the performance of LLMs on catalytic tasks on CatalystBench-test and report the results in Table 1 and Table 2. Table 1 compares the performance of CatalystLLM with other LLMs on tasks with factual answers. The results show that CatalystLLM achieve the best performance across all 7 tasks and 12 metrics, significantly outperforming general LLMs of similar scale. Compared to ChemLLM, CatalystLLM demonstrates a significant improvement in its capabilities to understand and apply catalytic knowledge, highlighting the effectiveness of multi-task catalytic data fine-tuning. ChemLLM and Darwin1.5 outperform top general models on most metrics. This indicates that

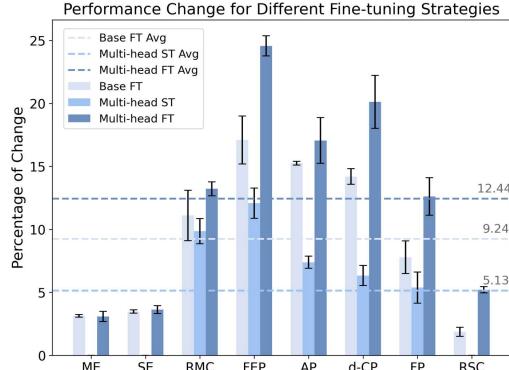


Figure 5: Comparison of the impact of different fine-tuning strategies on task performance. Base-ST serves as the baseline model and the bar chart shows the performance differences between different models on specific tasks.

378 specialized models can learn more precise chemical named entities, reaction conditions and critical  
 379 features in complex structure-activity relationships through fine-tuning in specific domains.  
 380

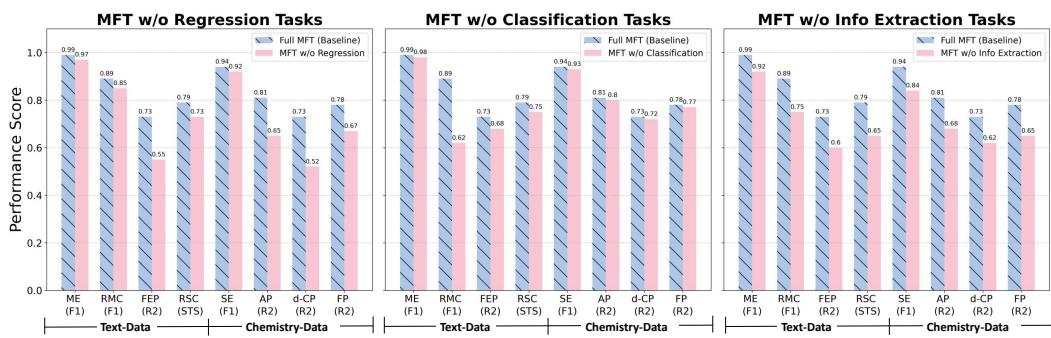
381 Notably, general LLMs perform exceptionally well in tasks such as ME, SE and RMC tasks. For  
 382 instance, gpt-4.1 achieves higher accuracy and F1-scores than CatalystLLM in the SE task. However,  
 383 in numerical regression prediction tasks, there is a significant gap in capability between general  
 384 models and specialized models, indicating that general LLMs currently can not accurately perform  
 385 prediction tasks which require deep scientific reasoning and quantitative calculations. For numerical  
 386 regression tasks, we also compare the results of CatalystLLM and traditional single-task competitive  
 387 ML algorithms in Appendix B.5.

388 Table 2 compares the performance of CatalystLLM with other LLMs in semantic un-  
 389 derstanding tasks, including multi-dimensional  
 390 evaluation metrics. In all three evaluation di-  
 391 mensions, CatalystLLM achieves the optimal  
 392 or suboptimal results, demonstrating the poten-  
 393 tial of domain-specific fine-tuning to enhance  
 394 answer quality. Additionally, the STS scores of  
 395 general LLMs are typically higher than those  
 396 of open-source models, reflecting the superior  
 397 answer generation capabilities. General LLMs  
 398 typically have higher LLM Scores, while spe-  
 399 cialized models including CatalystLLM have  
 400 relatively higher expert scores. This reflects  
 401 two phenomenons: 1) Top general models gen-  
 402 erate answers that are fluent, logically clear and  
 403 comprehensive, which aligns with the prefer-  
 404 ences of gpt-4o and deepseek-r1; 2) General models are prone to scientific hallucinations in spe-  
 405 cialized fields like catalytic science Xu et al. (2024). These answers may contain incorrect catalyst  
 406 performance analyses or explanations that conflict with physical chemistry principles and such er-  
 407 rors can only be identified by domain experts. We provide an error example of LLM-score and  
 408 expert scores in Appendix C.2. Additionally, we provide case studies to highlight the impact of  
 CatalystLLM’s domain expertise on catalytic materials scientists in Appendix C.3.

### 409 6.3 ABLATION STUDY

410 We compare the impact of different dataset settings and experimental settings on model capabilities  
 411 to explore the key factors that determine LLM domain capabilities. In the experimental setup, all  
 412 models are based on the optimal MFT architecture.

#### 413 6.3.1 COLLABORATIVE EFFECTS OF TASK COMBINATIONS



428 Figure 6: The experiments on collaborative effects of task combinations.  
 429

430 We investigate the potential collaborative relationship among different types of tasks under multi-  
 431 task learning framework. We remove regression tasks, classification tasks and information extraction

432 tasks from the training data separately and then evaluate the corresponding performance. As shown  
 433 in Fig 6, compared to the Full-MFT baseline, the performance of all ablation models decrease on  
 434 the corresponding tasks. Additionally, we divide these tasks into two groups based on the type of  
 435 input data: Text data and Chemical data and compare the collaborative effects of tasks within and  
 436 between groups.

437 **Intra-group tasks exhibit strong synergistic effects.** For example, when RMC data is removed,  
 438 the performance of other tasks within the same group decreases significantly. This indicates that  
 439 learning “how to regulate” (RMC) provides critical contextual knowledge for LLM to understand  
 440 “how effective the regulation is” (FEP) and “the significance of the regulation scheme” (RSC).

441 **Inter-group tasks exist certain degree of synergistic effects.** For example, in the “MFT w/o  
 442 Regression” setting, the performance of the RMC and RSC tasks is also affected to some extent. This  
 443 reveals a deeper phenomenon of domain-specific fine-tuning model: it does not merely memorize  
 444 input-output patterns from training data but construct an overall knowledge model of the catalytic  
 445 domain. CatalystLLM can span different data modalities, such as chemical formulas and natural  
 446 language, thereby aiding decision-making in other tasks.

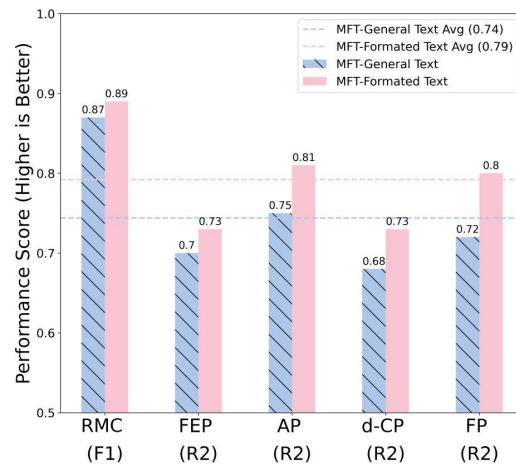
#### 448 6.3.2 THE IMPACT OF INPUT FORMAT ON MODEL PERFORMANCE

449 To investigate the advantages of Catalyst-  
 450 Bench, we investigate the impact of input data  
 451 format on the performance of LLMs. The ex-  
 452 periments aim to compare two input strate-  
 453 gies: one provided only unstructured core in-  
 454 formation, while the other provided structured  
 455 and complete input. The two input strategies  
 456 and related generated prompts are shown in  
 457 the Appendix E.2. As shown in Fig 7, given  
 458 the same amount of information, the quality and  
 459 format of input data are key factors de-  
 460 termining the upper limit of LLMs’ per-  
 461 formance in professional tasks, especially in the  
 462 chemical representation of catalysts. The sin-  
 463 gle SMILES string sequence can only provide  
 464 a macro-level representation, while structured  
 465 knowledge allows the model to learn the quan-  
 466 tity and connection methods of internal atoms,  
 467 thereby learning potential structure-activity re-  
 468 lationships.

## 469 7 CONCLUSION AND DISCUSSION

470 This work presents CatalystBench, a comprehensive benchmark dataset covering crucial processes  
 471 in catalytic science and systematically demonstrates the effectiveness of multi-head full-task fine-  
 472 tuning in adapting to complex scientific tasks. Building upon this foundation, we develop Cata-  
 473 lystLLM, a high-performance specialized language model tailored for the field of catalysis. Our  
 474 experiments demonstrate the superiority of CatalystLLM over existing LLM models and highlight  
 475 the pivotal role of CatalystBench in enhancing performance. Further analytical experiments also  
 476 provide crucial insights for developing LLM models for catalytic materials.

477 CatalystLLM attains SOTA performance, yet certain limitations remain. Its knowledge scope is con-  
 478 strained by the coverage of the CatalystBench dataset, potentially reducing predictive accuracy for  
 479 catalytic materials outside this range. While the 7B-parameter scale balances efficiency and capabil-  
 480 ity, larger models may exhibit superior reasoning and learning capacity. Furthermore, the inclusion  
 481 of data from existing literature introduces potential bias related to source selection. Future work will  
 482 aim to broaden CatalystBench to encompass a wider variety of catalytic systems and tasks, enhance  
 483 CatalystLLM’s downstream inference and embed the model within a closed-loop catalyst discov-  
 484 ery framework, enabling AI predictions to directly inform synthesis and characterization, thereby  
 485 accelerating the development of novel materials.



486 Figure 7: The impact of input format on model  
 487 performance.

486 ETHICS COMPLIANCE  
487

488 This paper adheres to the ICLR Code of Ethics. We acknowledge that our research follows ethical  
489 guidelines, ensuring compliance with all relevant regulations. No human subjects were involved in  
490 this research. All data used in the study are publicly available datasets or generated synthetically.  
491 We have taken care to ensure that all data used complies with relevant data protection laws. Datasets  
492 used in this research do not contain any personal, confidential, or sensitive information. We confirm  
493 that the research presented is original and has not been plagiarized. All sources and data are properly  
494 cited and documented. We confirm that the research complies with all applicable laws and ethical  
495 standards, including data usage and research practices. This statement is meant to address potential  
496 concerns in accordance with the ICLR Code of Ethics.

497 REPRODUCIBILITY STATEMENT  
498

500 We ensure reproducibility by providing all necessary resources for others to replicate our experi-  
501 ments. Specifically:

- 502 1) Benchmark Data and Experiment Setup: All datasets used in experiments and detailed data pro-  
503 cessing steps are publicly accessible. The experimental setup is also provided, including hyperpa-  
504 rameters, evaluation metrics and the benchmarking environment.
- 505 2) Source code: The code used for data processing and benchmarking experiments is provided in  
506 the supplementary materials, containing all scripts required to reproduce the experiments described  
507 in the paper.

509 By providing these resources, we aim to achieve full reproducibility of our research findings and  
510 facilitate subsequent studies based on this work. Detailed information required for reproduction can  
511 be found in the supplementary materials.

512 REFERENCES  
513

514 Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf  
515 Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure  
516 prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

517 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-  
518 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical  
519 report. *arXiv preprint arXiv:2303.08774*, 2023.

520 Luis M Antunes, Keith T Butler, Ricardo Grau-Crespo, Ricardo Grau-Crespo, Ricardo Grau-Crespo,  
521 and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language mod-  
522 eling. *Nature Communications*, 15(1):10570, 2024.

523 Viraj Bagal, Rishal Aggarwal, PK Vinod, U Deva Priyakumar, PK Vinod, U Deva Priyakumar,  
524 PK Vinod, U Deva Priyakumar, PK Vinod, and U Deva Priyakumar. Molgpt: molecular gener-  
525 ation using a transformer-decoder model. *Journal of chemical information and modeling*, 62(9):  
526 2064–2076, 2021.

527 Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge,  
528 Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

529 Suryanarayanan Balaji, Rishikesh Magar, Yayati Jadhav, and Amir Barati Farimani. Gpt-molberta:  
530 Gpt molecular features language model for molecular property prediction. *arXiv preprint  
arXiv:2310.03030*, 2023.

531 He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. Instructmol: Multi-modal integration  
532 for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint  
arXiv:2311.16208*, 2023a.

533 Zhonglin Cao, Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Moformer: self-  
534 supervised transformer model for metal–organic framework property prediction. *Journal of the  
American Chemical Society*, 145(5):2958–2967, 2023b.

540 Jinho Chang, Jong Chul Ye, Jong Chul Ye, Jong Chul Ye, and Jong Chul Ye. Bidirectional generation  
 541 of structure and properties through a single molecular foundation model. *Nature Communications*,  
 542 15(1):2323, 2024.

543 Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane  
 544 Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020  
 545 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

546 Hui Chen, Xiao Liang, Yipu Liu, Xuan Ai, Tewodros Asefa, and Xiaoxin Zou. Active site engineer-  
 547 ing in porous electrocatalysts. *Advanced Materials*, 32(44):2002435, 2020.

548 Xueqing Chen, Yang Gao, Ludi Wang, Wenjuan Cui, Jiamin Huang, Yi Du, and Bin Wang. Large  
 549 language model enhanced corpus of co2 reduction electrocatalysts and synthesis procedures. *Sci-  
 550 entific Data*, 11(1):347, 2024.

551 Seyone Chithrananda, Gabriel Grand, Bharath Ramsundar, Bharath Ramsundar, Bharath Ramsun-  
 552 dar, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular  
 553 property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

554 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit  
 555 Dhillon, Marcel Blstein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the  
 556 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-  
 557 bilities. *arXiv preprint arXiv:2507.06261*, 2025.

558 Yi Du, Ludi Wang, Mengyi Huang, Dongze Song, Wenjuan Cui, and Yuanchun Zhou. Autodive: An  
 559 integrated onsite scientific literature annotation tool. In *Proceedings of the 61st Annual Meeting  
 560 of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 76–85,  
 561 2023.

562 Carl Edwards, ChengXiang Zhai, Heng Ji, Heng Ji, and Heng Ji. Text2mol: Cross-modal molecule  
 563 retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical  
 564 Methods in Natural Language Processing*, pp. 595–607, 2021.

565 Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation  
 566 between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

567 Fei Fu, Qing-Qing Li, Fangrong Wang, Jie Hu, Tian-Tian Wang, Yun-Pei Liu, Weihong Xu, Zhili  
 568 Lin, Fu-Qiang Gong, Qi-Yuan Fan, et al. Synergizing a knowledge graph and large language  
 569 model for relay catalysis pathway recommendation. *National Science Review*, 12(8):nwaf271,  
 570 2025.

571 Yang Gao, Ludi Wang, Xueqing Chen, Yi Du, and Bin Wang. Revisiting electrocatalyst design by  
 572 a knowledge graph of cu-based catalysts for co2 reduction. *ACS Catalysis*, 13(13):8525–8534,  
 573 2023.

574 Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Gnnemann. Fast and  
 575 uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint  
 576 arXiv:2011.14115*, 2020a.

577 Johannes Gasteiger, Janek Groß, and Stephan Gnnemann. Directional message passing for molec-  
 578 ular graphs. *arXiv preprint arXiv:2003.03123*, 2020b.

579 Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Gnnemann, Zachary Ulissi,  
 580 C Lawrence Zitnick, and Abhishek Das. Gemnet-oc: developing graph neural networks for large  
 581 and diverse molecular simulation datasets. *arXiv preprint arXiv:2204.02782*, 2022.

582 Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas,  
 583 Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to  
 584 glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

585 Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang  
 586 Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on  
 587 eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.

594 Chaitanya B Hiragond, Niket S Powar, Junho Lee, and Su-II In. Single-atom catalysts (sacs) for photo-  
 595 catalytic co<sub>2</sub> reduction with h<sub>2</sub>o: Activity, product selectivity, stability, and surface chemistry.  
 596 *Small*, 18(29):2201428, 2022.

597

598 Ross Irwin, Spyridon Dimitriadis, Jiazen He, and Esben Jannik Bjerrum. Chemformer: a pre-  
 599 trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3  
 600 (1):015022, 2022.

601 Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leverag-  
 602 ing large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169,  
 603 2024.

604

605 Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen  
 606 Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The ma-  
 607 terials project: A materials genome approach to accelerating materials innovation. *APL materials*,  
 608 1(1), 2013.

609 Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bam-  
 610 ford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al.  
 611 Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

612

613 Gueyoung Jung, Matti A Hiltunen, Kaustubh R Joshi, Richard D Schlichting, and Calton Pu. Mis-  
 614 tral: Dynamically managing power, performance, and adaptation cost in cloud infrastructures. In  
 615 *2010 IEEE 30th International Conference on Distributed Computing Systems*, pp. 62–73. IEEE,  
 616 2010.

617 Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan,  
 618 and Gerbrand Ceder. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data*,  
 619 6(1):203, 2019.

620

621 Agustinus Kristiadi, Felix Strieth-Kalthoff, Marta Skreta, Pascal Poupart, Alán Aspuru-Guzik, and  
 622 Geoff Pleiss. A sober look at llms for material discovery: Are they actually good for bayesian  
 623 optimization over molecules? *arXiv preprint arXiv:2402.05015*, 2024.

624 Bernadette Lamb, Jonathan Herskovitz, Marta Jonson, Harlan Sayles, and Faruq Pradhan. S2185  
 625 use of large language model (llm) chatbots for the generation of specialized gastrointestinal diet  
 626 meal plans: A pilot study. *Official journal of the American College of Gastroenterology—ACG*,  
 627 119(10S):S1562, 2024.

628

629 Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian,  
 630 and Yu Li. Beyond chemical qa: Evaluating llm’s chemical reasoning with modular chemical  
 631 operations. *arXiv preprint arXiv:2505.21318*, 2025.

632 Jiatong Li, Junxian Li, Yunqing Liu, Dongzhan Zhou, and Qing Li. Tomg-bench: Evaluating llms  
 633 on text-based open molecule generation. *arXiv preprint arXiv:2412.14642*, 2024.

634

635 Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic  
 636 graphs. *arXiv preprint arXiv:2206.11990*, 2022.

637

638 Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved equivariant  
 639 transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023.

640 Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization  
 641 branches out*, pp. 74–81, 2004.

642

643 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,  
 644 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint  
 645 arXiv:2412.19437*, 2024a.

646 Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. Git-mol: A multi-modal large language  
 647 model for molecular science with graph, image, and text. *Computers in biology and medicine*,  
 171:108073, 2024b.

648 Siyu Liu, Jiamin Xu, Beilin Ye, Bo Hu, David J Srolovitz, and Tongqi Wen. Mattools: Bench-  
 649 marking large language models for materials science tools. *arXiv preprint arXiv:2505.10852*,  
 650 2025.

651

652 Renqian Luo, Lai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt:  
 653 generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioin-*  
 654 *formatics*, 23(6):bbac409, 2022.

655

656 Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe  
 657 Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelli-*  
 658 *gence*, 6(5):525–535, 2024.

659

660 Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and  
 661 Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85,  
 662 2023.

663

664 Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martíño Ríos-García, Benedict Emoek-  
 665 abu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke,  
 666 Anagha Aneesh, et al. Are large language models superhuman chemists? *arXiv preprint*  
 667 *arXiv:2404.01475*, 2024.

668

669 Janghoon Ock, Chakradhar Guntuboina, Amir Barati Farimani, Amir Barati Farimani, Amir  
 670 Barati Farimani, and Amir Barati Farimani. Catalyst energy prediction with catberta: unveil-  
 671 ing feature exploration strategies through large language models. *ACS Catalysis*, 13(24):16032–  
 672 16044, 2023.

673

674 Janghoon Ock, Srivathsan Badrinarayanan, Rishikesh Magar, Akshay Antony, and Amir Barati Fa-  
 675 rimani. Multimodal language and graph learning of adsorption configuration in catalysis. *Nature*  
 676 *Machine Intelligence*, 6(12):1501–1511, 2024.

677

678 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic  
 679 evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association*  
 680 *for Computational Linguistics*, pp. 311–318, 2002.

681

682 Andre Niyongabo Rubungo, Kangming Li, Jason Hattrick-Simpers, and Adji Bouso Dieng.  
 683 Llm4mat-bench: benchmarking large language models for materials property prediction. *Ma-*  
 684 *chine Learning: Science and Technology*, 6(2):020501, 2025.

685

686 Abdullahi Saka, Ridwan Taiwo, Nurudeen Saka, Babatunde Abiodun Salami, Saheed Ajayi, Kabiru  
 687 Akande, and Hadi Kazemi. Gpt models in construction industry: Opportunities, limitations, and  
 688 a use case validation. *Developments in the Built Environment*, 17:100300, 2024.

689

690 Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller.  
 691 Schnet—a deep learning architecture for molecules and materials. *The Journal of chemical physics*,  
 692 148(24), 2018.

693

694 Zhilong Song, Shuaihua Lu, Minggang Ju, Qionghua Zhou, and Jinlan Wang. Accurate prediction  
 695 of synthesizability and precursors of 3d crystal structures via large language models. *Nature*  
 696 *Communications*, 16(1):6530, 2025.

697

698 Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of ai  
 699 agents designs new sars-cov-2 nanobodies. *Nature*, pp. 1–3, 2025.

700

701 Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted,  
 702 Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. An autonomous  
 703 laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, 2023.

704

705 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko-  
 706 lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open founda-  
 707 tion and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

702 Amalie Trewartha, Nicholas Walker, Haoyan Huo, Sanghoon Lee, Kevin Cruse, John Dagdelen,  
 703 Alexander Dunn, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Quantifying the ad-  
 704 vantage of domain-specific pre-training on named entity recognition tasks in materials science.  
 705 *Patterns*, 3(4), 2022.

706 Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, and Andrew G Lee. Large language model (llm)-  
 707 driven chatbots for neuro-ophthalmic medical education. *Eye*, 38(4):639–641, 2024.

708

709 Yuwei Wan, Aswathy Ajith, Yixuan Liu, Ke Lu, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu  
 710 Kit, Tong Xie, and Ian T Foster. Sciqag: a framework for auto-generated scientific question  
 711 answering dataset with fine-grained evaluation. *CoRR*, 2024.

712 Ludi Wang, Yang Gao, Xueqing Chen, Wenjuan Cui, Yuanchun Zhou, Xinying Luo, Shuaishuai  
 713 Xu, Yi Du, and Bin Wang. A corpus of co2 electrocatalytic reduction process extracted from the  
 714 scientific literature. *Scientific Data*, 10(1):175, 2023a.

715 Ludi Wang, Dongze Song, Qiang Cui, Xueqing Chen, Yuanchun Zhou, Wenjuan Cui, and Yi Du.  
 716 Autodive+: An adaptive model enhanced multimodal online annotation tool. In *Companion Pro-  
 717 ceedings of the ACM on Web Conference 2025*, pp. 2919–2922, 2025a.

718

719 Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R  
 720 Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level sci-  
 721 entific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*,  
 722 2023b.

723

724 Xinning Wang, Jian Xu, Aslan H Feng, Yi Chen, Haiyang Guo, Fei Zhu, Yuanqi Shao, Minsi Ren,  
 725 Hongzhu Yi, Sheng Lian, et al. The hitchhiker’s guide to autonomous research: A survey of  
 726 scientific agents. *Authorea Preprints*, 2025b.

727 Kirsten T Winther, Max J Hoffmann, Jacob R Boes, Osman Mamun, Michal Bajdich, and Thomas  
 728 Bligaard. Catalysts-hub. org, an open electronic structure database for surface reactions. *Scientific  
 729 data*, 6(1):75, 2019.

730

731 Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and  
 732 interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

733

734 Tong Xie, Yuwei Wan, Yixuan Liu, Yuchen Zeng, Shaozhou Wang, Wenjie Zhang, Clara Grazian,  
 735 Chunyu Kit, Wanli Ouyang, Dongzhan Zhou, et al. Darwin 1.5: Large language models as  
 736 materials science adapted learners. *arXiv preprint arXiv:2412.11970*, 2024.

737

738 Changwen Xu, Yuyang Wang, and Amir Barati Farimani. Transpolymer: a transformer-based lan-  
 739 guage model for polymer property predictions. *npj Computational Materials*, 9(1):64, 2023.

740

741 Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of  
 742 large language models. *arXiv preprint arXiv:2401.11817*, 2024.

743

744 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu,  
 745 Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint  
 746 arXiv:2505.09388*, 2025.

747

748 Fangkai Yang, Pu Zhao, Zezhong Wang, Lu Wang, Jue Zhang, Mohit Garg, Qingwei Lin, Saravan  
 749 Rajmohan, and Dongmei Zhang. Empower large language model to perform better on industrial  
 750 domain-specific question answering. *arXiv preprint arXiv:2305.11541*, 2023.

751

752 Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou,  
 753 Chao Gong, Yang Shen, et al. A comprehensive capability analysis of gpt-3 and gpt-3.5 series  
 754 models. *arXiv preprint arXiv:2303.10420*, 2023.

755

756 Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. Llasmol: Advancing large language  
 757 models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset.  
 758 *arXiv preprint arXiv:2402.09391*, 2024.

759

760 Mohd Zaki, NM Anoop Krishnan, et al. Mascqa: investigating materials science knowledge of large  
 761 language models. *Digital Discovery*, 3(2):313–327, 2024.

756 Baicheng Zhang, Zhuoying Zhu, Huirong Li, Jiaqi Cao, and Jun Jiang. Revolutionizing chemistry  
757 and material innovation: An iterative theoretical-experimental paradigm leveraged by robotic ai  
758 chemists. *CCS Chemistry*, 7(2):345–360, 2025a.

759

760 Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang,  
761 Xiangyu Yue, Wanli Ouyang, et al. Chemllm: A chemical large language model. *arXiv preprint*  
762 *arXiv:2402.06852*, 2024a.

763

764 Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based  
765 agent system for materials science. *arXiv preprint arXiv:2409.00135*, 2024b.

766

767 Yu Zhang, Yang Han, Shuai Chen, Ruijie Yu, Xin Zhao, Xianbin Liu, Kaipeng Zeng, Mengdi Yu,  
768 Jidong Tian, Feng Zhu, et al. Large language models to accelerate organic chemistry synthesis.  
769 *Nature Machine Intelligence*, pp. 1–13, 2025b.

770

771 Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen  
772 Zhu, Su Zhu, et al. Chemdfm: a large language foundation model for chemistry. *arXiv preprint*  
773 *arXiv:2401.14818*, 2024.

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810 TABLE OF CONTENTS IN APPENDIX  
811

813	<b>A Data Construction</b>	<b>17</b>
814	A.1 Data Source . . . . .	17
815	A.2 Task Formulation . . . . .	18
816	A.3 The comparison of Chemical Benchmark . . . . .	18
817	A.4 Q&A Generation . . . . .	20
818	A.5 Quality Assurance Details . . . . .	20
819	A.5.1 Theoretical Q&A Data Calibration . . . . .	20
820	A.5.2 Open-ended Q&A Data Calibration . . . . .	20
821	A.6 Error Samples . . . . .	21
822	A.7 Data Distribution and Diversity Analysis . . . . .	22
823		
824	<b>B Experimental Details</b>	<b>24</b>
825	B.1 Domain-specific Challenges in CatalystBench Multi-task Learning . . . . .	24
826	B.2 Baseline LLMs . . . . .	24
827	B.3 Baseline traditional ML methods . . . . .	25
828	B.4 Evaluation among open-sourced models . . . . .	26
829	B.5 Evaluation with traditional ML methods . . . . .	27
830	B.6 Additional Ablation study . . . . .	28
831	B.6.1 The impact of weighted loss-function on model performance . . . . .	28
832	B.6.2 OOD Leave-One-Task-Out Evaluation . . . . .	29
833		
834	<b>C Evaluation Details</b>	<b>30</b>
835	C.1 LLM-based evaluation . . . . .	30
836	C.2 Expert evaluation . . . . .	30
837	C.3 Case Study: Unique Correct Reasoning in Complex Catalytic Context . . . . .	31
838		
839	<b>D Question Templates</b>	<b>32</b>
840	D.1 Label Definition . . . . .	32
841	D.2 Material Extraction Templates . . . . .	33
842	D.3 Synthesis Extraction Templates . . . . .	34
843	D.4 Regulation Method Classification Templates . . . . .	34
844	D.5 Property Prediction Templates-General text . . . . .	35
845	D.6 Property Prediction Templates-Chemical text . . . . .	35
846	D.7 Regulation Scheme Comprehension . . . . .	35
847		
848	<b>E Prompts</b>	<b>37</b>
849	E.1 Prompts for QA Generating . . . . .	37
850	E.2 Prompts for chemical-text generation . . . . .	37

864	E.3 Prompts for LLM-based evaluation . . . . .	39
865		

866	<b>F Statement on the Use of Large Language Models</b>	39
867		

## 868 A DATA CONSTRUCTION

870

871 We introduce the various sources of CatalystBench data, describe the specific tasks within each  
 872 category and clarify the complete process used to compile this dataset.

873

Ability	Task	Task Type	Datasets	Dataset Type	#count	Evaluation Metrics
Understanding	Material Extraction (ME)	Generation	ElectroCatalytic Reduction(1 & 2)	Experimental Data	2,480	Accuracy, F1 score
	Synthesis Extraction (SE)	Generation	ElectroCatalytic Reduction(1 & 2), Solution-based synthesis dataset	Experimental Data	6,612	
	Regulation Method Classification(RMC)	Classification	ElectroCatalytic Reduction(1 & 2)	Experimental Data	2,364	
Reasoning	Faradaic Efficiency Prediction(FEP)	Prediction	ElectroCatalytic Reduction(1 & 2)	Experimental Data	2,148	R <sup>2</sup> , MAE
	Adsorption Prediction (AP)	Prediction	OC20-Dense, SACs Dataset	Theoretical Simulation	3,000	
	d-band Center Prediction(d-CP)	Prediction	SACs Dataset, Catalysis-Hub, Catalytic Material Database(CMD)	Theoretical Simulation	3,000	
	Formation Prediction (FP)	Prediction	SACs Dataset, Material Project, Catalytic Material Database(CMD)	Theoretical Simulation	3,000	
Explaining	Regulation Scheme Comprehension(RSC)	Generation	ElectroCatalytic Reduction(1 & 2)	Experimental Data	4,307	STSscore, LLM Score

892

893 Table 3: The statistics of all tasks, datasets, dataset type and evaluation metrics.

894

## 895 A.1 DATA SOURCE

896

897 We extract data from 8 open-source databases. The name and description of each dataset are listed  
 898 below.

899

- 900 • **ElectroCatalytic Reduction 1 Wang et al. (2023a)<sup>1</sup>**: An open-source corpus of electro-  
 901 catalytic CO<sub>2</sub> reduction extracted from science literatures Du et al. (2023).
- 902 • **ElectroCatalytic Reduction 2 Chen et al. (2024)<sup>2</sup>**: A text-mining dataset describing the  
 903 CO<sub>2</sub> reduction process catalyzed by copper-based electrocatalysts Wang et al. (2025a),  
 904 which specifically includes material, regulation method, product, Faradaic efficiency and  
 905 relevant synthesis conditions.
- 906 • **Solution-based synthesis dataset Kononova et al. (2019)<sup>3</sup>**: A dataset of 35,675 solution-  
 907 based synthesis procedures extracted from the scientific literature. Each procedure contains  
 908 essential synthesis information including the precursors and target materials, their quanti-  
 909 ties and the synthesis actions and corresponding attributes.
- 910 • **OC20-Dense Chanusso et al. (2021)<sup>4</sup>**: The largest dataset of catalyst-adsorbate inter-  
 911 actions to date, designed to accelerate catalyst discovery through machine learning. The  
 912 dataset contains over a million DFT calculations covering an extremely wide range of cat-  
 913 alyst materials (including bulk, surfaces and nanoparticles) and adsorbates.
- 914 • **Catalytic Material Database(CMD)<sup>5</sup>**: CMD contains material composition, properties,  
 915 reactions, products and other information.
- 916 • **SACs Hiragond et al. (2022)<sup>6</sup>**: A dataset of catalyst samples constructed for electrocat-  
 917 alytic reactions such as OER or ORR. SACs contains dozens of transition metals, combined

918 with several different coordination environments and substrates, resulting in hundreds of  
 919 catalyst samples.  
 920

- 921 • **Catalyst Hub Winther et al. (2019)<sup>7</sup>**: A featured database for surface reactions contains  
 922 more than 100,000 chemisorption and reaction energies obtained from electronic structure  
 923 calculations and is continuously being updated with new datasets.
- 924 • **Materials Project Jain et al. (2013)<sup>8</sup>**: The Materials Project provides computed information  
 925 on known and predicted materials as well as powerful analysis tools to inspire and  
 926 design novel materials.

## 928 A.2 TASK FORMULATION

930 We evaluate the understanding, reasoning and explanation capabilities of LLMs in catalyst design by  
 931 constructing 8 widely recognized practical catalytic material tasks. Table 3 summarizes these tasks,  
 932 including their task types from a machine learning perspective, the datasets used for evaluation and  
 933 the evaluation metrics.

934 **Simple Instruction Data Synthesis Method:** Converting structured chemical data into reasoning-  
 935 adjusted data suitable for training LLMs presents two main challenges: 1) the creation of diverse  
 936 templates and 2) the integration of chemical logic and reasoning into QA pairs. The information  
 937 extraction tasks focus on identifying composition and operational parameters related to material  
 938 design from scientific literatures. The classification tasks predict discrete categories or labels, while  
 939 the regression tasks focus on continuous numerical property values. For template diversity, we  
 940 initially developed a foundational seed template to meet the requirements of specific tasks. Using  
 941 gpt-4o, We generate a series of prompt templates that express different meanings while maintaining  
 942 semantic consistency. These diverse templates enhance the model’s ability to interpret and respond  
 943 to different instruction formats. For each structured data entry, we randomly select one of these  
 944 templates to create a single-round dialogue sample.

945 The source datasets typically use various representations of catalytic materials to characterize the  
 946 corresponding catalytic properties Cao et al. (2023b); Balaji et al. (2023); Xu et al. (2023); Ock  
 947 et al. (2023). We propose a key hypothesis: When the model is only provided with the material  
 948 common name and its SMILES representation, it is difficult for LLM to learn the complex structure-  
 949 property relationships implicit in the data. To validate this hypothesis and explore the impact of  
 950 input information richness on model performance, we create prompts based on the source dataset  
 951 to generate different types of catalytic characterization input data. The prompts are included in  
 952 Appendix E.2.

953 **Complex Scenario Data Synthesis Method:** For information-rich complex texts such as synthesis  
 954 schemes, we enriched the data by constructing multi-round dialogues to provide contextual depth  
 955 and logical consistency. By simulating the deep thinking and step-by-step reasoning involved in  
 956 catalytic material design processes, we created a highly specialized multi-round dialogue dataset,  
 957 reducing the need for dialogue rounds and human intervention. The above template examples are  
 958 included in Appendix D.

## 959 A.3 THE COMPARISON OF CHEMICAL BENCHMARK

960 Catalysis research presents unique challenges that distinguish it from broader chemistry or mate-  
 961 rials science applications, particularly in the combination of heterogeneous knowledge modalities  
 962 and the multi-stage, interdependent workflow of real catalyst design. While prior LLM benchmarks  
 963 in chemistry and materials science have explored domain-specific reasoning or property prediction,  
 964

965 <sup>1</sup><https://doi.org/10.57760/sciedb.07106>

966 <sup>2</sup><https://doi.org/10.57760/sciedb.13290>

967 <sup>3</sup><https://doi.org/10.6084/m9.figshare.16583387.v4>

968 <sup>4</sup><https://opencatalystproject.org/>

969 <sup>5</sup><http://cmd.us.edu.pl/catalog/>

970 <sup>6</sup><https://catalysis-ncepu-hvkydg736ykqeq26d5gxrn.streamlit.app/>

971 <sup>7</sup><http://www.catalysthub.net/>

972 <sup>8</sup><https://next-gen.materialsproject.org/>

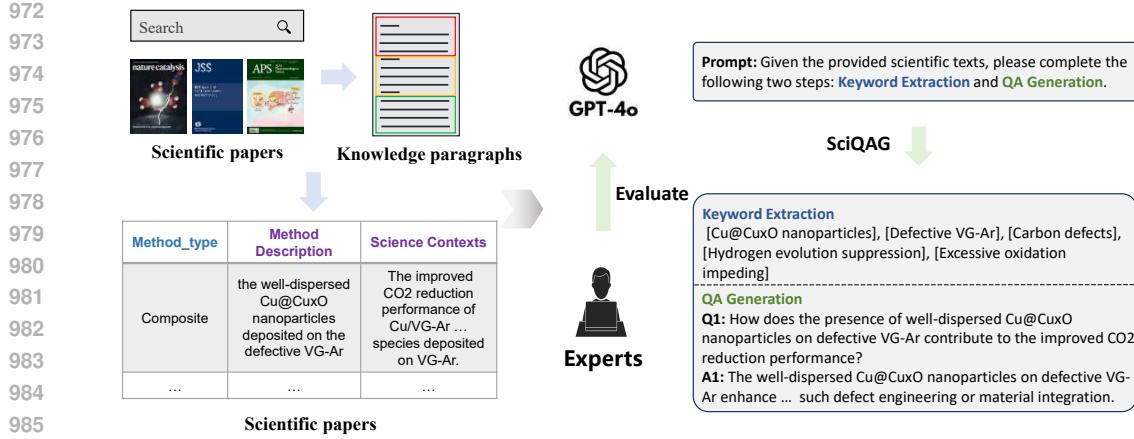


Figure 8: The process of converting regulation method text into high-quality Q&A pairs using SciQAG.

none have simultaneously aligned tasks to the full, practical catalyst R&D pipeline — from extraction of synthesis parameters in literature, through structure–property prediction, to the mechanistic interpretation of regulation strategies.

Benchmark	Domain Focus	Task Type	Data Modalities	Workflow Alignment to Real Catalyst Design
ChemBench	General chemistry knowledge	Multiple-choice factual recall	Text only	✗ Fragmented
MaScQA	Materials science concepts	Q&A from graduate-level examinations	Text only	✗ Fragmented
SCIBENCH	College-level scientific problem solving	Multi-domain Q&A	Text only	✗ Fragmented
ChemCoTBench	Chemical reasoning	Step-wise chain-of-thought	Text symbolic expressions	✗ Fragmented
LLM4Mat-Bench	Material property prediction	Regression/classification on numerical properties	Structured material descriptors	✗ Partial
CatalystBench (ours)	Catalysis-specific, combining theory experiment	8 tasks: information extraction, method classification, numerical property prediction, regulation scheme comprehension	Text structured chemical descriptors numerical values	✓ Full workflow coverage

Table 4: The comparison of CatalystBench and the current Benchmark across dimensions such as domain categories, data modalities and task types.

Table 4 compares representative existing chemical benchmarks with CatalystBench in terms of task categories, data modalities and alignment to real-world catalyst design workflows. The comparison illustrates that CatalystBench uniquely integrates both high-fidelity theoretical simulations and curated experimental literature into a task suite directly reflecting the sequential steps of modern catalytic materials discovery.

1026  
1027

## A.4 Q&amp;A GENERATION

1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039

We use SciQAG Wan et al. (2024) to guide GPT-4o in interpreting a large volume of regulation scheme texts, converting them into high-quality comprehension Q&A pairs. The pre-defined task is defined as follows: Given a seed input catalytic design scheme text  $T$ , for each input text  $t$ , the generator should first generate 5 keywords  $k$  that capture the most important terms and concepts in the text and then generate a set  $S = \{(q_i, a_i)\}_{i=1}^n$  focusing on the generated keywords  $k$ , where  $\forall i \in \{1, 2, \dots, 5\}$ ,  $q_i$  is the question and  $a_i$  is the answer to  $q_i$ . We use the 1,198 catalytic design scheme texts from the ElectroCatalytic Reduction dataset as input to generate 5,990 seed Q&A pairs by prompting GPT-4o. The generating prompts are included in Appendix E.1. To reduce the occurrence of non-knowledge-based questions that can only be answered using information from the given article, we develop a rule-based method to remove all pairs containing “this paper” or “this study.” As shown in Fig 8, we ultimately generate 4,307 semantic understanding text Q&A pairs in this step.

1040

## A.5 QUALITY ASSURANCE DETAILS

1041  
1042  
1043  
1044  
1045

In order to guarantee the scientific credibility of CatalystBench and to ensure reproducibility of results, a multi-layered quality assurance framework was implemented during dataset development. This framework covers both theoretically simulated catalytic property data and open-ended Q&A content extracted and generated from experimental literature.

1046  
1047

## A.5.1 THEORETICAL Q&amp;A DATA CALIBRATION

1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057

For task data derived from theoretical simulations, we first perform foundational consistency checks before feeding it into the conversion interface and generating corresponding QA-formatted content. This process removes samples with obvious errors, such as invalid SMILES representations, missing chemical formulas, or attribute values falling outside the original data’s statistical range. Simultaneously, samples with duplicate structures or identical text are cleaned using hash fingerprint comparison and field uniqueness checks. To maintain scientific rigor while preserving large-scale data volume, we randomly selected 20% of samples from the cleaned theoretical dataset and invited two independent experts in catalytic materials for manual verification. The review process employs a double-blind cross-evaluation approach to ensure conclusions are not influenced by individual reviewer preferences and to achieve explicit consensus on answer consistency assessments.

1058  
1059

## A.5.2 OPEN-ENDED Q&amp;A DATA CALIBRATION

1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070

For open-ended comprehension Q&A data sourced from experimental literature and automatically generated by SciQAG-guided GPT-4o, we implement an additional rule-based filtering system to remove potentially noisy article-internal reference questions. This includes questions relying solely on non-generic knowledge cues like “this paper” or “this study.” This filtering effectively mitigates the risk of model over-reliance on specific textual contexts during training. Furthermore, during the generation phase, we avoid a single-round direct instruction-to-output approach. Instead, we design a multi-round generation process: first extracting keywords and key mechanisms, then guiding the model to generate Q&A focused on core scientific concepts, followed by structured rewriting and semantic consistency review. This multi-round strategy significantly mitigates GPT-4o’s tendency toward fixed stylistic expressions and non-scientific rhetoric in lengthy responses. Consequently, the retained data exhibits greater expressive diversity, enhanced scientific rigor and high consistency with actual catalytic mechanisms.

1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

We further introduce a verifiable knowledge attribution mechanism to reduce and identify potential hallucination content. During the extraction of contextual keywords, we concurrently build an internal retrieval database comprising three elements: 1) We retain the unique DOI identifiers of corresponding literature to ensure each Q&A pair can be traced back to its original publication; 2) We preserve the complete contextual paragraphs underlying generated questions for offline review and revalidation; 3) We associate extracted core scientific keywords with their corresponding text fragments, establishing a keyword-context-DOI mapping. This structured database not only provides an additional quality control measure beyond expert review for open-ended Q&A data but also enables rapid programmatic verification of answer consistency with original sources. This maintains the high standards for scientific verifiability and reproducibility of our datasets.

1080 We also randomly sample 50% of the dataset and conduct manual verification by experts in the field  
 1081 of catalysis. We invite several experts from the National Nanotechnology Center’s catalysis field  
 1082 to participate, including 2 associate researcher and 6 doctoral student. The review process covers  
 1083 three central aspects: **1) The reasonableness and scientific value of the open-ended Q&A task**  
 1084 **questions**, whether they reflect the development trends and focal points of the catalysis field; **2) The**  
 1085 **accuracy and completeness of the answers**, ensuring they include all key conditions; and **3) The**  
 1086 **relevance and contextual fidelity of the Q&A tasks**, ensuring there are no hallucinations  
 1087 generated by the model itself. After an initial filtering, approximately 90.8% of the Q&A tasks are  
 1088 identified accurate and scientifically meaningful. To further validate the rigor of our filtering process,  
 1089 we conduct IAA analysis on 500 randomly selected samples. Two domain experts independently  
 1090 annotate each sample as accept or reject achieving a Cohen’s Kappa coefficient of 0.75. This further  
 1091 confirms that the dataset reflects domain knowledge rather than model bias.  
 1092  
 1093  
 1094  
 1095  
 1096  
 1097  
 1098  
 1099

#### A.6 ERROR SAMPLES

1100  
 1101  
 1102  
 1103  
 1104 The errors in automatically generated open-ended Q&A responses primarily stem from models in-  
 1105 troducing inferences inconsistent with mainstream consensus in the catalysis field, mismatched ex-  
 1106 perimental conditions, or unfounded explanations of reaction mechanisms during answer generation.  
 1107 To clarify the feature of this issue, we present several representative error samples, all annotated and  
 1108 verified by two independent domain experts.  
 1109  
 1110  
 1111  
 1112  
 1113  
 1114  
 1115

##### Error Sample 1

###### Question:

1116 What is the specific role of Zn sites in the CO electro-reduction process within this Cu-Zn  
 1117 bimetallic catalytic system?  
 1118

###### Reference Answer:

1119 The Zn site serves as an efficient active center for CO generation, selectively reducing CO  
 1120 to the \*CO intermediate and synergistically promoting subsequent C-C coupling with the  
 1121 Cu site to form ethylene.  
 1122

###### The answer generated by LLM:

1123 Zn sites directly participate in C-O bond cleavage, promoting the reduction of CO to CO  
 1124 while enhancing the selectivity of Cu sites toward ethylene.  
 1125

###### The error causes:

1126 The model erroneously interprets the Zn site’s action as directly participating in bond  
 1127 cleavage, which contradicts the explicitly documented mechanism in the literature of  
 1128 “indirect action via intermediate generation.” This constitutes a mechanistic error.  
 1129

1134  
1135**Error Sample 2**

1136

**Question:**

1137

What is the primary contribution of nitrogen doping when employing an electrocatalytic system supported by nitrogen-doped graphene for Fe single-atom active sites?

1139

**Reference Answer:**

1140

Nitrogen doping alters the local electronic structure, increases the electron density at active sites and improves reaction kinetics by enhancing adsorption capacity.

1143

**The answer generated by LLM:**

1144

Nitrogen doping enhances the carrier conductivity, enabling it to maintain structural stability under high current densities.

1145

**The error causes:**

1146

The generated answer unduly emphasizes enhanced conductivity while overlooking the dominant effects verified in the literature: regulation of active site electronic structures and improvement in adsorption capacity. This error constitutes a critical omission of key factors.

1147

1148

1149

1150

1151

1152

1153

1154

**Error Sample 3**

1155

**Question:**

1156

What is the primary purpose of introducing nanoporous structures in Au-Cu alloys for CO electrolytic reduction?

1157

1158

**Reference Answer:**

1159

The porous structure increases the specific surface area and the number of exposed active sites while improving the mass transfer efficiency of reactants.

1160

1161

**The answer generated by LLM:**

1162

Naphthalene-based structures are primarily employed to suppress the formation of methane byproducts, enhancing selectivity through altered electron cloud distribution.

1163

1164

**The error causes:**

1165

The response focuses solely on selective enhancement and methane suppression as its sole objectives, lacking descriptions of surface structure and mass transfer effects. This results in insufficient scientific rationale and inconsistency with the original context keywords.

1166

1167

1168

1169

1170

1171

1172

1173

**A.7 DATA DISTRIBUTION AND DIVERSITY ANALYSIS**

1174

1175

To comprehensively demonstrate the composition of CatalystBench, this section conducts detailed statistical and diversity analysis of the dataset. We aim to reveal its coverage breadth and data distribution characteristics across dimensions such as catalyst types, regulation method types and reaction types through quantitative data, thereby validating its rationale and challenge as a comprehensive benchmark. Fig 9 illustrates the distribution characteristics of the catalysts and reaction types involved in CatalystBench.

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

**Distribution of Key Metal Elements:** We compile the key metal elements in catalysts across the dataset, particularly in tasks derived from experimental literature such as ME, SE and RMC tasks. As shown in Fig 9a, *Cu* is the most frequently occurring element, accounting for 43.7%. This aligns closely with our data sources' focus on the current state of research in *CO<sub>2</sub>* electrocatalytic reduction(*CO<sub>2</sub>RR*), where copper-based catalysts represent a major research focus. Simultaneously, the dataset extensively includes transition metals such as *Fe*, *Co* and *Ni*, which play crucial roles in electrocatalysis, along with various other metallic elements. This ensures the model can learn diverse elemental knowledge.

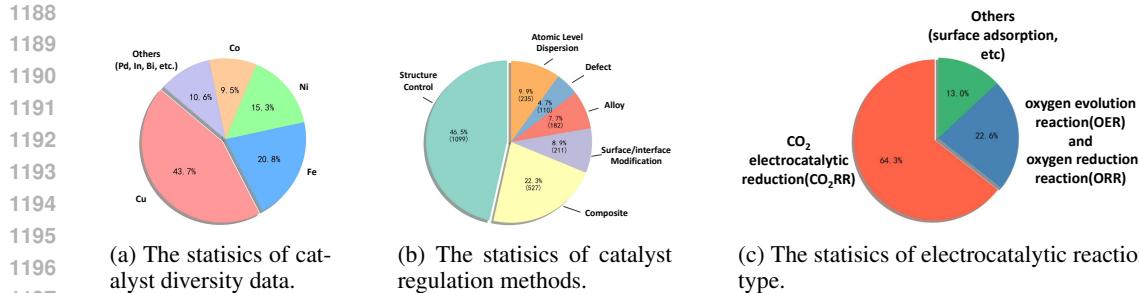


Figure 9: The distribution characteristics of the catalysts and reaction types involved in CatalystBench.

**Distribution of Catalyst Regulation Methods:** CatalystBench incorporates a variety of advanced catalyst design and control strategies. We statistically analyze 2,364 samples from the RMC task. As shown in Fig 9b, Structure control and Composite materials are the two most prevalent approaches, accounting for 46.5% and 22.3% respectively. This reflects the current mainstream trend of enhancing catalytic performance through constructing heterojunctions and multi-component synergies. Additionally, Surface modification, Alloying, Defect engineering and Atomic-level dispersion also account for significant proportions, comprehensively covering key technologies from macroscopic morphology to atomic-scale regulation.

**Distribution of Reaction Type:** CatalystBench encompasses several core reaction types in the electrocatalysis field. As shown in Fig 9c,  $CO_2$  electrocatalytic reduction( $CO_2RR$ ) constitute the largest proportion of the dataset, accounting for 64.3%, directly attributable to our integration of multiple literature sources and datasets focused on  $CO_2RR$ . Simultaneously, by incorporating datasets like SACs and OC20, we have also included a substantial amount of theoretical calculation data related to oxygen evolution reaction(OER) and oxygen reduction reaction(ORR), combined as 22.6%. Additionally, a small number of other surface adsorption and catalytic reactions are included, collectively forming a diverse set of reaction scenarios.

1242 **B EXPERIMENTAL DETAILS**  
12431244 **B.1 DOMAIN-SPECIFIC CHALLENGES IN CATALYSTBENCH MULTI-TASK LEARNING**  
12451246 CatalystBench presents a unique configuration of tasks that diverges substantially from typical  
1247 multi-task learning settings, primarily due to the heterogeneous nature of inputs, outputs and the  
1248 rich domain-specific dependencies intrinsic to catalytic science. Unlike general-purpose multi-task  
1249 benchmarks where tasks may share a uniform modality and output format, the catalytic design work-  
1250 flow necessitates the simultaneous handling of mixed input modalities comprising structured chem-  
1251 ical descriptors, unstructured scientific text and numerical parameters. These inputs range from  
1252 SMILES strings, atomic coordination tables, or facet specifications to free-form descriptions of syn-  
1253 thesis methods and regulation schemes extracted from literature, creating a complex representational  
1254 space in which feature alignment across modalities is non-trivial.1255 The heterogeneity further extends to output modalities. Regression tasks such as Adsorption Prediction  
1256 and d-band Center Prediction and Faradaic Efficiency Prediction demand high-precision numerical  
1257 values. Classification tasks, exemplified by Regulation Method Classification, require discrete cate-  
1258 gorical decisions grounded in chemical taxonomy. Generative tasks, including Material Extraction  
1259 and Regulation Scheme Comprehension, produce long-form textual responses that must balance flu-  
1260 ency and factual correctness while integrating domain knowledge. This diverse set of output spaces  
1261 imposes fundamentally different optimization landscapes; mean squared error (MSE) objectives for  
1262 regression tasks often operate on a much larger numerical scale than cross-entropy losses used in  
1263 classification, or token-level negative log-likelihoods in generation. Without explicit control mech-  
1264 anisms, the disproportionately high gradients from regression losses can dominate joint training,  
1265 suppressing learning in classification and generative objectives, a phenomenon that is particularly  
1266 detrimental in scientific settings where each capability is equally critical.1267 Moreover, tasks within CatalystBench are not independent in their knowledge requirements. The  
1268 Regulation Method Classification task, for instance, provides crucial mechanistic context that di-  
1269 rectly influences downstream predictive accuracy in Faradaic Efficiency estimation. Similarly, un-  
1270 derstanding textual regulation schemes is often prerequisite to correctly interpreting property trends.  
1271 These upstream-downstream dependencies mean that multi-task learning in this domain must not  
1272 only avoid interference between tasks, but also preserve and exploit beneficial cross-task syner-  
1273 gies. Naively unifying outputs in a shared generative head, as in the FT setting, disregards both the  
1274 modality-specific optimization needs and the asymmetric informational flow between tasks, often  
1275 leading to diluted performance gains. Conversely, the MST approach, while introducing modality-  
1276 specific heads, treats each task in isolation during training, forfeiting the representational transfer  
1277 that arises from shared backbone exposure to the full diversity of inputs.1278 Empirical evidence from Fig 5 underscores these points. The FT configuration suffers notable degra-  
1279 dation in numerical prediction tasks, reflecting loss-scale interference and insufficient modality dif-  
1280 ferentiation, whereas MST exhibits limited average improvement due to the absence of multi-task  
1281 semantic alignment. By contrast, the MFT strategy achieves the highest mean improvement rate of  
1282 12.44% over the ST baseline, leveraging joint backbone training to capture shared domain features,  
1283 while decoupled output heads mitigate loss conflict across heterogeneous modalities. These findings  
1284 substantiate that the challenges in CatalystBench, such as input heterogeneity, output space diver-  
1285 gence, loss imbalance and domain-specific task coupling necessitate an architecture that is simulta-  
1286 neously modality-aware and synergy-preserving, a requirement uniquely fulfilled by the proposed  
1287 MFT fine-tuning framework.1288 **B.2 BASELINE LLMs**  
12891290 The following sections will introduce the specific open-source LLMs and general closed-source  
1291 LLMs we evaluate in the main text. The comparison results are presented in the main text.1292 **LLaMA-2** Touvron et al. (2023) is a set of large language models with parameter ranges from 7  
1293 billion to 70 billion. The model architecture remains largely unchanged from LLaMA-1, but over  
1294 40% of the data is used to train the base model. Specifically, LLaMA-2 includes pre-trained and fine-  
1295 tuned models optimized for conversational applications and we have chosen to use LLaMA-2-7B as  
1296 the model for comparative evaluation.

1296 **Mistral** Jung et al. (2010) is an open-source large language model with 7 billion parameters, de-  
 1297 signed to offer high efficiency and performance on a wide range of downstream tasks. The model  
 1298 features a transformer architecture with various optimizations on tokenization and training data di-  
 1299 versity. In this study, we select Mistral-7B for comparative evaluation.

1300 **ChatGLM** GLM et al. (2024) is a bilingual large language model containing 6 billion parame-  
 1301 ters. It adopts the General Language Model (GLM) architecture and is optimized for dialogue and  
 1302 question-answering scenarios. ChatGLM-6B is pre-trained on extensive Chinese and English cor-  
 1303 pora, enabling strong cross-lingual generation capabilities.

1304 **Qwen3** Yang et al. (2025) are trained on diverse multilingual datasets to improve performance across  
 1305 languages and domains. The model is optimized for multilingual and multitask settings, benefiting  
 1306 from large-scale pre-training. For evaluation purposes, Qwen3-8B is selected as a representative of  
 1307 mid-sized models in the Qwen3 series.

1308 **ChemLLM** Zhang et al. (2024a) is an open-source language model deeply optimized for the chemi-  
 1309 cal field, with 7B parameters. ChemLLM has converted vast amounts of structured chemical knowl-  
 1310 edge sources into over 40 million high-quality single-round Q&A pairs, known as ChemData. By  
 1311 fine-tuning on this large-scale instruction dataset, ChemLLM can learn a wealth of chemical facts  
 1312 and understand and execute diverse chemical tasks through seamless conversational interactions.

1313 **Darwin1.5** Xie et al. (2024) is an open-source foundational LLM tailored for material science and  
 1314 chemistry. The model is designed with a two-stage training strategy, namely question-answering  
 1315 (QA) fine-tuning and multi-task learning, to enable LLM to proficiently perform chemical and ma-  
 1316 terials tasks.

1317 **Claude-3-haiku** is one member of the Claude-3 series of commercial large language models devel-  
 1318 oped by Anthropic. It is designed to provide fast, safe and helpful conversational abilities, partic-  
 1319 ularly for business and enterprise applications. Being a closed-source model, our interactions with  
 1320 Claude-3-haiku are conducted via API, employing few-shot inference to ensure fair and efficient  
 1321 evaluation.

1322 **Gemini-2.5-flash-preview** Comanici et al. (2025) is a pre-release version of Google’s Gemini-2.5  
 1323 series, offering improvements in dialogue understanding and response speed. As a closed-source  
 1324 model, we access Gemini-2.5-flash-preview through its API, enabling batch evaluation for compar-  
 1325 ative studies while maintaining low overhead.

1326 **Deepseek-v3** Liu et al. (2024a) is the third-generation model from the Deepseek series, developed  
 1327 to enhance code understanding, generation and natural language interactions. With advanced pre-  
 1328 training strategies on code and text data, Deepseek-v3 demonstrates strong capabilities across mul-  
 1329 tiple tasks.

1330 **GPT-3.5-turbo and GPT-4.** For closed-source models such as OpenAI GPT Family GPT-3.5-turbo  
 1331 Ye et al. (2023) and GPT-4 Achiam et al. (2023), we employ batch inference via APIs for conducting  
 1332 few-shot prompt inference. This approach significantly enhances evaluation efficiency and reduces  
 1333 overhead.

### 1336 B.3 BASELINE TRADITIONAL ML METHODS

1337 **CatBERT** Ock et al. (2023) is a Transformer model based on the RoBERTa architecture, specif-  
 1338 ically designed for predicting adsorption energies in catalytic systems. It can replace traditional  
 1339 graph neural network(GNN) methods, which rely on precise three-dimensional atomic coordinates,  
 1340 by using human-readable text inputs that describe the “adsorbate-catalyst” system. By processing  
 1341 text containing key information such as adsorbate types, catalyst host material compositions and  
 1342 interacting atoms, CatBERTa achieves accurate predictions of catalyst adsorption energies.

1343 **GAP-CatBERT** Ock et al. (2024) improves the accuracy of catalyst property prediction through a  
 1344 graph-assisted pre-training strategy. During the pre-training phase, the model utilizes a powerful  
 1345 graph neural network, EquiformerV2 Liao et al. (2023); Liao & Smidt (2022), to generate high-  
 1346 quality graph embeddings from precise atomic structures. While the CatBERTa model generates  
 1347 text embeddings from corresponding textual descriptions. Subsequently, this method aligns the  
 1348 embeddings from these two different modalities through contrastive learning, thereby transferring  
 1349 and injecting the rich structural knowledge contained in the graph model into the text model. The

1350 resulting GAP-CatBERTa achieves respectable accuracy in adsorption energy prediction tasks using  
 1351 only text input.

1352 **GPTchem** Jablonka et al. (2024) is an effective prediction method which combines the power-  
 1353 ful representation capabilities of LLM with Gaussian Process Regression(GPR) strategies. In this  
 1354 framework, the model uses LLM fine-tuned by massive amounts of chemical text to convert discrete  
 1355 chemical entity information into feature embedding vectors, which are then fed into a GPR model.  
 1356 GPR can not only predict specific chemical properties of molecules such as formation energy, but  
 1357 also provide a quantified confidence interval for the prediction results through Bayesian inference.

1358 **CGCNN** Xie & Grossman (2018) (**Crystal Graph Convolutional Neural Network**) is a graph  
 1359 convolutional neural network (GCN) model specifically designed for crystalline materials. Within  
 1360 CGCNN, the crystal structure is abstracted as a crystal graph, wherein atoms serve as the graph’s  
 1361 nodes and chemical bonds between atoms represent the graph’s edges. Node features typically rep-  
 1362 resent invariant properties characterising atomic species. By stacking graph convolutional layers  
 1363 and pooling layers, this model learns both local and global features within the crystal structure, ulti-  
 1364 mately enabling accurate prediction of various physicochemical properties such as formation energy  
 1365 and bandgap.

1366 **SchNet** Schütt et al. (2018) is a deep learning model specifically designed for molecular and pe-  
 1367 riodic structures, aiming to learn potential energy surfaces and quantum chemical properties. At  
 1368 its core, SchNet employs continuous filtering convolutional layers to process atomic environments,  
 1369 with these convolutional operations being rotationally and translationally invariant. It progressively  
 1370 refines atomic features through interaction blocks that utilise interatomic distance information. Cru-  
 1371 cially, SchNet introduces Gaussian Radial Basis Functions (RBFs) to represent interatomic dis-  
 1372 tances, enabling the model to capture long-range interactions that vary continuously with distance.  
 1373 The model efficiently and accurately predicts atomic energies, forces and other quantum chemical  
 1374 properties.

1375 **DimeNet++** Gasteiger et al. (2020a) is a geometric deep learning model that builds upon the  
 1376 DimeNet Gasteiger et al. (2020b) framework, focusing on capturing angular information and tripar-  
 1377 tite interactions within molecules. Unlike traditional GNNs that solely consider atoms and bonds,  
 1378 DimeNet++ introduces a Directional Message Passing mechanism. It accounts not only for the  
 1379 connection from atom  $i$  to  $j$ , but also for the message transmitted from atom  $k$  to  $i$  via  $j$ , thereby  
 1380 explicitly encoding the angular information formed by atoms  $i$ ,  $j$  and  $k$ . This explicit handling of  
 1381 the three-body term enables DimeNet++ to achieve higher accuracy and rotation invariance when  
 1382 predicting molecular properties, while simultaneously enhancing computational efficiency through  
 1383 optimised architecture.

1384 **GemNet-OC** Gasteiger et al. (2022) is a high-performance graph neural network model, specifically  
 1385 designed as a variant for catalyst adsorption system architectures. Building upon DimeNet and  
 1386 SchNet, it further emphasises the utilisation of geometric information. By integrating explicit three-  
 1387 body terms and higher-order angular information, this model accurately characterises the intricate  
 1388 interactions between adsorbates and surface atoms on catalysts, thereby enabling high-precision  
 1389 predictions of key catalytic properties such as adsorption energies. It stands as one of the state-of-  
 1390 the-art models in the field of catalysis that relies on atomic coordinate inputs.

1391 **Faradaic efficiency prediction** is a relatively new property prediction task in the field of catalytic  
 1392 materials. Gao et al. Gao et al. (2023) construct a knowledge graph of electrocatalysts based on sci-  
 1393 entific literature and propose a deep learning-based prediction model, which integrates the semantic  
 1394 information from the scientific literature (word embedding) with the correlation of knowledge triples  
 1395 (graph embedding) and realizes the prediction of the Faradaic efficiency for a targeted case.

1396

#### 1397 B.4 EVALUATION AMONG OPEN-SOURCED MODELS

1398

1399 Previous studies have shown that general LLMs often perform poorly when handling tasks requiring  
 1400 deep understanding of chemical structures, such as SMILES strings and complex chemical reason-  
 1401 ing, as they lack internalized, domain-specific knowledge systems Bagal et al. (2021); Liu et al.  
 1402 (2024b); Chang et al. (2024). To select an appropriate base model, we systematically evaluated  
 1403 multiple existing open-source LLMs with 7B/8B parameters on representative tasks from Catalyst-  
 Bench. The evaluation tasks encompass 3 core capabilities: structured numerical prediction, text

classification and text generation. This approach aims to comprehensively characterize the models' reasoning and generalization abilities across theoretical and experimental data conditions. The Results are shown in Table 5, which demonstrate that ChemLLM-7B maintains stable and outstanding performance across most tasks, exhibiting significant advantages over other candidate models particularly in reaction mechanism analysis and catalyst performance prediction. Therefore, we designate ChemLLM as the base model for fine-tuning to ensure subsequent experiments build upon a robust foundation of comprehensive performance.

The decision to limit the model size to small parameters stems from a balanced consideration of computational resources and application scenarios. On one hand, a 7B/8B-scale model can undergo multi-task fine-tuning on typical GPU clusters with relatively low training overhead, ensuring reproducibility and generalizability throughout the research process. On the other hand, this parameter scale already achieves satisfactory task performance in chemistry and materials domains while maintaining high operational feasibility for practical deployment.

## B.5 EVALUATION WITH TRADITIONAL ML METHODS

We compare CatalystLLM with traditional machine learning methods on 4 property prediction regression tasks in CatalystBench. The ML results for each task are obtained using specific single-task data. Taking the adsorption energy prediction task as an example, both CatBERTa and GAP-CatBERTa use the atomic text information of the catalyst, which is also the generic text in the template as Appendix D.5, as input to predict the adsorption energy value of the catalyst. These ML algorithms typically only accept one input format and do not have results for all tasks. For the AP, d-CP and FP tasks, we introduce mainstream graph neural network benchmark models currently prevalent in materials science, including CGCNN, SchNet, DimeNet++ and GemNet-OC. These models are widely regarded as standard benchmarks in adsorption energy prediction benchmarks such as OC20. We compare CatalystLLM against publicly available checkpoints of these algorithms, all of which are trained on the same data set size as CatalystBench. It is particularly noteworthy that these graph neural network models rely on complete atomic 3D geometric coordinates and lattice parameters as inputs. They explicitly encode 3D spatial structural information through message-passing mechanisms to achieve high-precision predictions. In contrast, CatalystLLM is limited to textual descriptions extracted from experimental literature or databases, along with simplified structured text sequences such as chemical formulas, crystal plane information and adsorption site descriptions. Additionally, due to differences in input representation formats, not all tasks have ML baselines, such as text classification and semantic understanding tasks. For information extraction tasks, LLMs like the GPT series are more accurate than traditional entity extraction methods. The detail results of evaluation is shown in Table 6 and Table 7.

As can be seen from Table 6, the advantage of CatalystLLM over traditional ML methods lies in its exceptional versatility. It can provide effective predictions across a wide range of task types, whereas other models typically focus on one or two specific tasks. In terms of specific performance, CatalystLLM achieve the best score of 0.73 in the d-CP task, outperforming GPTchem. In other tasks, CatalystLLM also demonstrate strong competitiveness. For example, its adsorption energy

Model	IE	RMC	AP
<b>ChemLLM-7B</b>	<b>0.86</b>	<b>0.52</b>	<b>0.63</b>
Darwin1.5	0.84	0.48	0.59
LLaMA2	0.76	0.38	0.24
Qwen3-8B	0.80	0.41	0.31

Table 5: The performance of four open-source LLMs of similar scale on representative tasks in CatalystBench.

Model	AP	d-CP	FP
CGCNN	0.79	0.66	0.78
SchNet	0.72	0.65	0.77
DimeNet++	0.84	0.69	0.80
GemNet-OC	<u>0.85</u>	<u>0.72</u>	<u>0.84</u>
CatBERTa	0.82	/	/
GAP-CatBERTa	<b>0.86</b>	/	/
GPTchem	/	0.69	<b>0.85</b>
CatalystLLM (ours)	0.81	<b>0.73</b>	0.80

Table 6: The results of CatalystLLM compared with machine learning baselines for 3 prediction tasks. The best model is in bold font and the second-best is underlined.

As can be seen from Table 6, the advantage of CatalystLLM over traditional ML methods lies in its exceptional versatility. It can provide effective predictions across a wide range of task types, whereas other models typically focus on one or two specific tasks. In terms of specific performance, CatalystLLM achieve the best score of 0.73 in the d-CP task, outperforming GPTchem. In other tasks, CatalystLLM also demonstrate strong competitiveness. For example, its adsorption energy

1458 prediction score is 0.81 and close to SOTA model GAP-CatBERTa. In summary, while highly optimized specialized models may perform better in certain single tasks, CatalystLLM demonstrates 1459 its immense value as an efficient catalyst design LLM through its balanced and competitive performance 1460 across a wide range of catalytic prediction tasks. 1461

1462 Notably, when compared against graph- 1463 structured prediction models, CatalystLLM’s 1464 predictive accuracy falls slightly below that 1465 of top-tier graph models like GemNet-OC and 1466 DimeNet++, which rely on precise geometric 1467 structures. Moreover, CatalystLLM outper- 1468 forms earlier SchNet models. This indicates 1469 CatalystLLM’s ability to effectively capture 1470 chemical knowledge and structural features 1471 implicit in textual descriptions through multi- 1472 task fine-tuning. Crucially, widely-used GNNs 1473 require 3D coordinates derived from costly 1474 DFT relaxations. CatalystLLM achieves com- 1475 parable accuracy using only text information. 1476 This enables instantaneous inference on vast 1477 chemical spaces where 3D structures are unknown. 1478 For the FEP prediction task, we introduce tradi- 1479 tional text prediction methods from the current 1480 scientific domain. As shown in Table 7, although 1481 it falls short of specially trained state-of-the-art 1482 models for this task, CatalystLLM significantly 1483 outperforms other ML approaches, demon- 1484 strating considerable potential in this domain. 1485

## 1481 B.6 ADDITIONAL ABLATION STUDY

1483 In addition to the ablation experiments described in the main text, we also compare the impact of 1484 different experimental settings on model capabilities to explore the crucial elements of LLM fine- 1485 tuning experiments. 1486

### 1487 B.6.1 THE IMPACT OF WEIGHTED LOSS-FUNCTION ON MODEL PERFORMANCE

1489 This experiment aims to verify the critical role 1490 of weighted loss-function calculation in multi- 1491 head Full-task fine-tuning setting. In the ex- 1492 perimental setup described in the main text, 1493 we manually adjust the weights based on the 1494 performance of ChemLLM-MST model on the 1495 validation set for each task to balance the 1496 contribution of different tasks to the model gra- 1497 dient update. In addition, we use an unweighted 1498 loss-function on the Multi-head architecture, in 1499 which the weight coefficients  $\lambda_{task,i}$  for all tasks 1500 are set to 1.

1501 Fig 10 shows the comparison of the results of 1502 the two sets of experiments. MFT-weighted 1503 demonstrate better performance in most tasks, 1504 proving that loss weighting is a necessary step 1505 to achieve high-performance multi-task learn- 1506 ing. Among them, MFT-Unweighted shows a 1507 significant decline in performance in tasks such 1508 as classification and information extraction. This 1509 may be because in multi-task learning, the num- 1510 ical scales of the loss functions of different tasks 1511 often differ by orders of magnitude. For example, 1512 the mean squared error in regression tasks has a larger scale than the cross-entropy loss in classi- 1513 fication tasks, leading the model to prioritize regression tasks during back-propagation. Therefore, 1514 we need to set different weight parameters to balance the contributions of different tasks to the to- 1515 tal gradient, thereby maintaining the multi-task capability of CatalystLLM. In addition, a possible

Model	MSE	MAE	$R^2$
multi-layer perceptron	0.03	0.15	0.48
support vector regression	0.06	0.20	0.11
linear regression	0.05	0.17	0.20
BRR	0.05	0.18	0.23
GPR	0.04	0.15	0.34
original paper	<b>0.01</b>	<b>0.08</b>	<b>0.84</b>
CatalystLLM (ours)	<u>0.02</u>	<u>0.13</u>	<u>0.73</u>

Table 7: The results of CatalystLLM compared with machine learning baselines for FEP task. The best model is in bold font and the second-best is underlined.

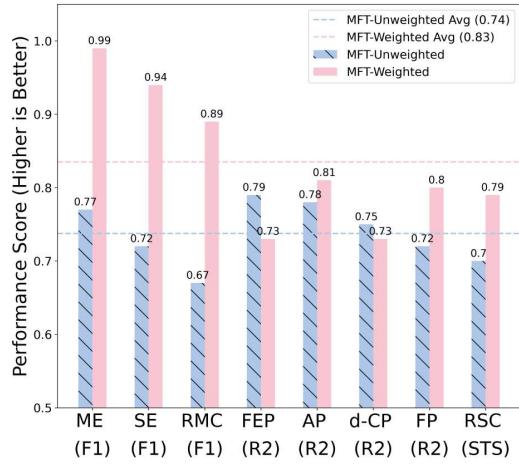


Figure 10: The impact of weighted loss-function of MFT on model performance.

28

1512 future optimization direction is to set the weight parameters as hyperparameters and dynamically  
 1513 adjust them during model training.  
 1514

1515 **B.6.2 OOD LEAVE-ONE-TASK-OUT EVALUATION**  
 1516

1517 To systematically evaluate CatalystLLM’s generalization capabilities on unseen tasks, we design  
 1518 a “Leave-One-Task-Out” experimental strategy. Building upon Section 6.3 of the main text, we  
 1519 further conduct out-of-domain performance analysis for specific tasks. Considering the prior  
 1520 relationships among the 8 CatalystBench tasks—particularly the strong correlations in input modalities  
 1521 and knowledge dimensions for certain tasks, we select the Faradaic Efficiency Prediction (FEP)  
 1522 task as the holdout set. This task is excluded from model training and only evaluated during the  
 1523 inference phase. This task belongs to the numerical regression category, with inputs derived from  
 1524 structured catalyst characterization and performance data in experimental literature. It exhibits po-  
 1525 tential upstream-downstream logical connections with tasks like Regulation Method Classification  
 1526 (RMC) and Regulation Scheme Comprehension (RSC). FEP prediction requires comprehensive un-  
 1527 derstanding and inference regarding the regulation methods, structural features and corresponding  
 1528 intermediate formation processes of the catalyst system.

1529 This experiment aims to validate whether CatalystLLM can retain a certain level of performance  
 1530 through cross-task transfer learning and shared representation modeling, even under training condi-  
 1531 tions with complete absence of task-specific data. We compare CatalystLLM with ChemLLM-FT  
 1532 and representative general-purpose open-source models LLaMA2-7B and Mistral-7B under iden-  
 1533 tical data partitioning and inference settings. All models are fine-tuned without FEP task training  
 1534 data, using the remaining seven tasks as training sets. The results are shown in Table 8.

1535 Experimental results indicate that under the  
 1536 FEP-withheld setting, CatalystLLM still signif-  
 1537 icantly outperforms baseline models on both  
 1538  $R^2$  and MAE metrics, representing an improve-  
 1539 ment of approximately 0.11 over ChemLLM-  
 1540 FT and exceeding 0.20 compared to general-  
 1541 purpose models. The outcome indicates that  
 1542 FEP task performance largely benefits from  
 1543 knowledge transfer and sharing within the  
 1544 multi-task architecture, particularly contribu-  
 1545 tions from structural regulation tasks in terms  
 1546 of semantics and patterns. Furthermore, de-  
 1547 spite performance declines compared to the original setting with FEP training data, CatalystLLM  
 1548 still generates predictions with physicochemical plausibility. This demonstrates that its generaliza-  
 1549 tion capability relies not only on task matching but also on the internal organization and expression  
 1550 of domain knowledge.

Model	$R^2$	MAE
ChemLLM-FT	0.58	3.21
LLaMA2-7B	0.49	3.87
Mistral-7B	0.51	3.75
<b>CatalystLLM</b>	<b>0.69</b>	<b>2.84</b>

Table 8: The experiments on OOD Leave-One-  
 Task-Out evaluation on FEP task

1566  
1567

## C EVALUATION DETAILS

1568  
1569  
1570

We conduct a multi-dimensional evaluation of open-ended Q&A tasks focused on understanding and explaining catalytic regulation schemes.

1571  
1572

### C.1 LLM-BASED EVALUATION

1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586

The RSC task aims to evaluate the model’s deep understanding of catalytic regulation schemes and its text generation capabilities. The output is a segment of natural language, rather than fixed labels or numerical values. Traditional automated evaluation metrics, such as BLEU Papineni et al. (2002) or ROUGE Lin (2004) based on phrase overlap rates, cannot effectively measure the factual accuracy of answers. For example, an answer may use different words from the standard answer but still be semantically correct and conversely. Therefore, we use gpt-4o and deepseek-r1 as evaluation models M. Bran et al. (2024). They are prompted to assume the role of a catalytic field expert, but has no access to external tools such as internet browsing. We collaborate with domain experts to design a set of evaluation prompts so that LLM would imitate scientists in their thinking patterns and give scores in different dimensions: **1) Reasonableness** assesses whether the content generated by the model follows basic scientific principles and causal relationships; **2) Accuracy** assesses whether the technical terms used in the model output are consistent with recognized scientific knowledge; **3) Usability** assesses whether the answer generated by the model is practical to the question and has a certain degree of reality and operability. The LLM-based evaluation prompt we designed is shown in Appendix E.3.

1587  
1588

### C.2 EXPERT EVALUATION

1589

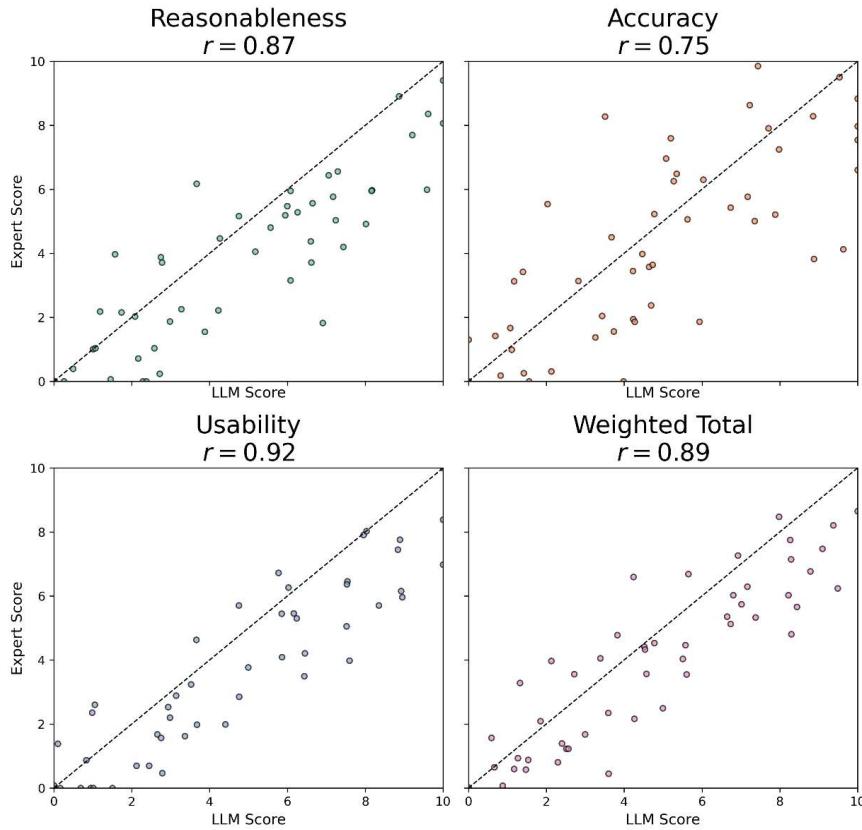
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619

Figure 11: The 3 sub-metrics and the overall Pearson score between LLM Score and Expert Score. Although general LLMs have vast knowledge reserves and powerful reasoning capabilities, they have a prevalent issue: the generation of hallucinations such as incorrect or fabricated information

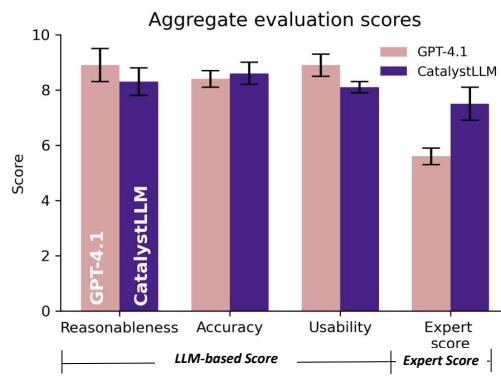
1620 presented as factual Xu et al. (2024). Therefore, we invite domain experts to evaluate the model’s  
 1621 performance on a test set of 100 samples. We observe a clear trend of consistency between LLM  
 1622 scores and expert ratings across the overall distribution of these samples. The Pearson correlation  
 1623 coefficient for the average scores of the 100 samples is 0.89, indicating a high degree of linear cor-  
 1624 relation in the numerical variations between the two types of scoring. We further analyze the 3  
 1625 sub-metrics and the overall score. As shown in Fig 11, samples rate highly by the LLM also tended  
 1626 to receive relatively high scores from experts, with a similarly pronounced consistency observed for  
 1627 low-scoring samples. This outcome demonstrates that despite divergences in assessing the accuracy  
 1628 of specific procedural details, both evaluation systems exhibit close alignment in judging overall  
 1629 answer quality trends. This consistency provides practical support for leveraging prompt-optimized  
 1630 LLM scoring during initial data screening and offers statistical justification for subsequent integra-  
 1631 tion with expert review to achieve efficient quality control.

1632 Regarding the variability in human scoring, Fig  
 1633 12 shows the differences between LLM eval-  
 1634 uation and human expert evaluation. Although  
 1635 human experts prefer CatalystLLM’s responses  
 1636 based on the accuracy of catalytic materials and  
 1637 corresponding mechanisms, LLMs prefer GPT-  
 1638 4.1’s answers, typically basing their evalua-  
 1639 tion on the fluency and apparent completeness of  
 1640 GPT-4.1 responses. Therefore, we focus on ex-  
 1641 cluding GPT 4.1’s answers that are complete  
 1642 and fluent enough to cover the question but  
 1643 contain incorrect information. Fig 13 shows an  
 1644 example of an illusion in LLM-based score and  
 1645 expert score. GPT4.1’s answer achieves higher  
 1646 LLM-based score, but it makes inferences that  
 1647 violated the principles of catalysis, which is  
 1648 “K<sup>+</sup> directly participates in C-O bond cleav-  
 1649 age”. In the field of electrocatalysis, there is  
 1650 a non-covalent interaction mechanism whereby  
 1651 alkali metal ions stabilize intermediates through  
 1652 electrostatic field effects rather than directly par-  
 1653 ticipating in bond formation or bond breaking. In  
 1654 general, LLMs tend to favor answers that are formally  
 1655 rigorous and logically coherent. However, in scientific evalua-  
 1656 tion, the correctness of the answer is  
 1657 more valuable than a complete and detailed explanation. For scientific tasks requiring real-world  
 1658 knowledge, LLM-based evaluation methods cannot replace expert human assessment for now.

### 1659 C.3 CASE STUDY: UNIQUE CORRECT REASONING IN COMPLEX CATALYTIC CONTEXT

1660 To more intuitively demonstrate CatalystLLM’s domain adaptation capability, we select a test case  
 1661 that only CatalystLLM could answer completely and correctly among all comparison models, which  
 1662 is shown in Appendix D.7. This task belongs to the Regulation Scheme Comprehension(RSC)  
 1663 category, involving a high-order coupled multi-metal catalytic system whose underlying mechanism  
 1664 was not directly present in the training data.

1665 Among all comparative model responses, CatalystLLM uniquely link the inferred product shift to-  
 1666 ward methane and hydrogen to the synergistic mechanism involving <sup>\*</sup>CO production at the Zn site.  
 1667 It accurately identifies that losing the Zn site reduces intermediate concentration and alters the reac-  
 1668 tion pathway, thereby disrupting the essential conditions for ethylene formation. This response not  
 1669 only aligns with established physicochemical principles but also resonates with mainstream under-  
 1670 standing of electrocatalytic mechanisms at the mechanistic level, demonstrating the model’s capacity  
 1671 for abstract reasoning when inferring complex sequential mechanisms. This case demonstrates that  
 1672 CatalystLLM can leverage cross-task learned catalytic reaction knowledge to make domain-logical  
 1673 inferences even without direct case training. This holds significant reference value for chemical  
 1674 engineers and catalytic scientists exploring novel multifunctional catalyst designs. More importantly,  
 1675 this case reflects the unique potential of domain-adapted LLMs in supporting high-level scientific  
 1676 problem-solving. Their accuracy stems not only from data coverage but also from the learned rep-  
 1677 resentation’s ability to capture and structurally express catalytic principles during fine-tuning.



1678 Figure 12: Average scores for gpt-4o and  
 1679 deepseek-r1 across three metrics and aggregated  
 1680 results from expert evaluators (n=100) across all  
 1681 tasks. Error bars indicate confidence intervals  
 1682 (95%).

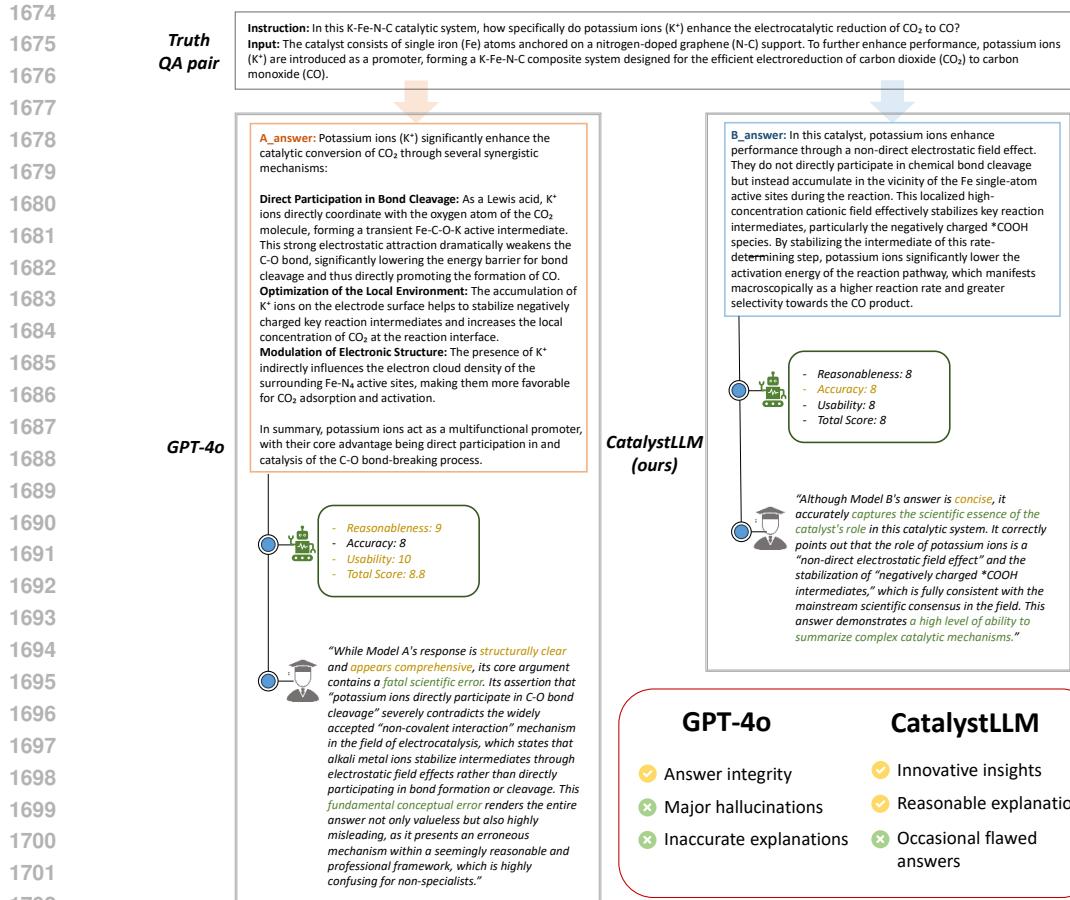


Figure 13: For the example truth Q&A pair of the same regulatory scheme semantic understanding task, the LLM-based scores and expert scores of answers given by GPT-4.1 and CatalystLLM. The yellow highlights indicate positive evaluations and the green highlights indicate negative evaluations.

## D QUESTION TEMPLATES

In recent years, there has been a surge in the application of machine learning to chemistry, resulting in a wealth of datasets and benchmarks in chemistry and material field. However, few of these benchmarks focus on assessing LLMs for tasks specific to Catalytic Science and a standardized evaluation technique has not yet been established. This section provides concrete examples for each catalytic science task category, demonstrating how to formulate inputs and showcasing the model’s expected outputs for representative problems.

### D.1 LABEL DEFINITION

**<Contexts>:** The text describing catalytic material control schemes or synthesis schemes in scientific literature. e.g. "The synthesized  $Cu$  catalysts show relatively high crystallinity dominated by  $\{111\}$  facets, promoting the generation of  $CH_4$ ."

**<Material\_representation>:** The actual representation of catalytic material typically shows its elemental composition or structural details, including SMILES strings, chemical formulas, CIF and other forms of existence. e.g.  $Cu2Pd$ .

**<Operations>:** The common operations in catalytic material synthesis schemes include ADDING, HEATING, CURING, ELECTROCHEMICAL ANODIZATION, FILTERING,

1728 DRYING, DIPPING and REACITON. e.g. The reaction mixture was then heated at 85 °C  
 1729 with constant stirring for 3 h.  
 1730

1731

1732

1733

1734

1735

1736

1737

1738

1739 <Operation conditions>: The relevant attributes of catalytic material synthesis operations,  
 1740 including heating temperature and duration. e.g. 85 °C and 3 h

1741

1742

1743

1744

1745

1746

1747

1748

1749

1750

1751

<Property>: The characteristics or properties of catalytic material describe its chemical  
 characteristics and crucial performance, including adsorption energy, d-band center,  
 formation energy and actual catalytic Faradaic efficiency. e.g. 90%.

1752

1753

1754

1755

1756

1757

1758

1759

1760

1761

1762

## D.2 MATERIAL EXTRACTION TEMPLATES

1763

1764

1765

1766

1767

### Material Extraction Templates

#### Instructions:

- What specific materials are mentioned as catalysts in the text?
- List all keywords that refer to catalyst materials or their types discussed in the article.
- Which terms in the document are associated with or describe catalyst materials?

Input: <Contexts>

Output: <Material\_representation>

1782 D.3 SYNTHESIS EXTRACTION TEMPLATES  
1783

1784

1785

1786 **Synthesis Extraction Templates**

1787

1788

**Instructions:**

- **Round 1: What is the target material synthesized in the given text?**

1790

1791

**Input:** <Contexts>

1792

1793

**Output:** The target material synthesized in the text is <Material\_representation>.

1794

1795

1796

1797

1798

1799

**Input:** <Contexts>

1800

1801

1802

1803

1804

1805

**Output:** The experimental operations mentioned are <Operations>

1806

1807

- **Round 3: For <Operations>, specify the relevant <Operation conditions>.**

1808

**Input:** The reaction mixture was then heated at 85 °C with constant stirring for 3 h.

1809

1810

**Output:** The <Operation conditions> is 800°C.

1811

1812

1813

1814

1815

1816

## D.4 REGULATION METHOD CLASSIFICATION TEMPLATES

1817

1818

**Regulation Method Classification Templates**

1819

**Instructions:**

1820

1821

- **What is the type of regulation method applied to the catalytic material in the described scheme?**
- **Which regulation approach or technique is used to modify the catalytic material as detailed in the text?**
- **Identify the specific method of regulation implemented in this catalytic material scheme.**

1822

1823

1824

1825

1826

1827

1828

1829

1830

1831

1832

**The options:** A. Alloy B. Defect C. Atomic level dispersion D. Surface/interface modification E. Composite F. Structure control

1833

1834

1835

**Input:** Cu/Bi bi-metal compounds derived from MOFs.**Output:** E. Composite.

1836 D.5 PROPERTY PREDICTION TEMPLATES-GENERAL TEXT  
18371838 **Property Prediction Templates**  
18391840 **Instructions:**

- What is the <Property> of the catalytic material represented by this <Material\_representation>?
- Can you provide the <Property> associated with this catalytic material <Material\_representation> for reduction of CO<sub>2</sub> to produce CO?
- What is the <Property> for the catalytic material described by <Material\_representation>?

1844  
1845 **Input:** CHTi<sub>18</sub>Pd<sub>54</sub> (2 1 0) [C Ti Pd Pd hollow [Ti Pd Pd Pd Pd Pd Pd ...]].  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853**Output:** -1.6033 eV.1854  
1855  
1856 D.6 PROPERTY PREDICTION TEMPLATES-CHEMICAL TEXT  
18571858 **Property Prediction Templates**  
18591860 **Instructions:**

- Calculate the adsorption energy of CH species on the *Ti<sub>18</sub>Pd<sub>54</sub>(210)* surface using the <Material\_representation>:

1864 **input:**

- **Adsorbate:** CH
- **Catalyst:** *Ti<sub>18</sub>Pd<sub>54</sub>*
- **Facet:** (2 1 0)
- **Active site:** hollow site coordinated by C, Ti, Pd, Pd

1865  
1866 **Output:** -1.6033 eV.  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
18741875 D.7 REGULATION SCHEME COMPREHENSION  
18761877 **User**  
18781879 **Instruction:**1880 How does the core-shell architecture of CoP@NiFe-LDH specifically address the common  
1881 limitations of using NiFe-LDH alone for the Oxygen Evolution Reaction (OER)?1882 **Method Description:**

1883 To construct a highly efficient catalyst for the Oxygen Evolution Reaction (OER), a hierarchical  
1884 core-shell structure was designed. The core consists of conductive cobalt phosphide  
1885 (CoP) nanowires, which serve as a scaffold. Vertically aligned nickel-iron layered double  
1886 hydroxide (NiFe-LDH) nanosheets were then grown directly on the surface of the CoP  
1887 nanowires. The resulting composite material, denoted as CoP@NiFe-LDH, leverages the  
1888 high intrinsic activity of the NiFe-LDH shell and the metallic properties of the CoP core.

1890

**CatalystLLM**

1891

1892

**[Answer]**

1893

1894

1895

1896

1897

1898

1899

**User**

1900

**Instruction:**

1901

1902

1903

1904

1905

1906

1907

1908

1909

1910

1911

1912

1913

Based on the described tandem catalysis mechanism, if the zinc (Zn) sites in this catalyst were to be completely removed or deactivated, leaving only the functional copper (Cu) sites, what change is expected for the catalyst's major product? Please explain the reasoning.

**Method Description:**

To efficiently convert carbon dioxide ( $CO_2$ ) into high-value ethylene ( $C_2H_4$ ), researchers designed a dual-function catalyst. This catalyst consists of atomically adjacent copper (Cu) and zinc (Zn) sites dispersed on a nitrogen-doped carbon (NC) support, denoted as Zn-Cu/NC. The core mechanism is a "tandem catalysis" process: first, the Zn sites act as efficient "CO factories," selectively reducing  $CO_2$  to the carbon monoxide (\*CO) intermediate. Subsequently, due to spatial proximity, the generated \*CO intermediate rapidly migrates to the adjacent Cu sites, where the critical C-C coupling reaction occurs to ultimately produce ethylene ( $C_2H_4$ ).

**CatalystLLM****[Answer]**

Removing the zinc (Zn) sites would shift the catalyst's selectivity from ethylene ( $C_2H_4$ ) to methane ( $CH_4$ ) and hydrogen ( $H_2$ ). This is because eliminating the Zn "\*CO factory" breaks the synergistic tandem mechanism, depriving the copper (Cu) sites of the high concentration of \*CO intermediates needed for C-C coupling. Consequently, the catalyst would behave as a less efficient, single-function copper system, where the reaction favors deep reduction to methane or hydrogen evolution.

1923

1924

1925

1926

1927

1928

1929

1930

1931

1932

1933

1934

1935

1936

1937

1938

1939

1940

1941

1942

1943

1944  
1945

## E PROMPTS

1946  
1947

Here are the prompt instructions we build during the benchmark construction and evaluation process.

1948

### E.1 PROMPTS FOR QA GENERATING

1949  
1950

#### <Q&A Generation>Prompt

1951

**Please complete the following tasks based on the text of the catalytic regulation scheme provided below and its context:**

1952  
1953**1. Task requirements:**1954  
1955

- Extract core keywords related to the catalytic scheme [Method description] from the text and context of the regulation scheme (e.g., reaction type, material system, innovative mechanisms, performance enhancement points, etc.).
- Based on the extracted keywords, identify 3–5 key questions regarding the characteristics, advantages and improvements of this regulation scheme compared to traditional catalytic schemes and provide answers based on the scheme and its context.
- Questions should focus on: regulation schemes, material innovation, performance advantages and practical application prospects. Avoid asking simple or definitional questions.
- Assume that the user does not have access to the original paper or any external sources, so ensure that the questions and answers are self-contained. - Answers should be concise and specific, based on textual and contextual [Contexts] facts.

1956  
1957**2. Example:**1958  
1959**Input:**1960  
1961

- Method description: {example\_regulation\_method\_description}
- Contexts: {example\_contexts}

1962  
1963**3. output format:**1964  
1965

- Please present the generated question-answer pairs in the following format:

1966  
1967**Q1: [Question 1]**1968  
1969**A1: [Answer 1]**1970  
1971**...****Q5: [Question 5]**1972  
1973**A5: [Answer 5]**1974  
1975

- Use of multiple sentence structures. Questions need to be phrased in a way that is easy to understand.

1976  
1977**Input:**1978  
1979

- Method description: {regulation\_method\_description}

1980  
1981

- Contexts: {contexts}

1982

### E.2 PROMPTS FOR CHEMICAL-TEXT GENERATION

1983  
1984

In cheminformatics, SMILES (Simplified Molecular Input Line Entry System) is a widely used method for representing molecules as linear strings. It encodes a molecule’s topological structure, atom types, bond types and partial stereochemical information using a sequence of ASCII characters. This compact format facilitates storage and transmission. However, this encoding relies on a character sequence governed by syntactic rules, with structural information implicit within the string pattern. For language models lacking specialized chemical parsing components, extracting relationships between atoms and bonds requires symbolic parsing, resulting in relatively low information explicitness.

Task Type	General Text	Formatted Text
RMC	"Nano-porous $Au_3Cu$ alloy	"Material: $Cu$ , Scheme: $Au_3Cu$ , Structure: Nano-porous"
FEP	"Nano-porous $Au_3Cu$ alloy material catalyzes $CO_2$ reduction to produce $CO$ "	"Material: $Cu$ , Scheme: $Au_3Cu$ , Structure: Nano-porous, Reactant: $CO_2$ , Product: $CO$ "
AP	" $FeN_4$ –graphene-*OH"	"Active_center: $Fe$ , Coordination: $N_4$ , Support: graphene, Adsorbate: *OH"

Table 9: The examples of different input format for CatalystBench tasks.

1998 In the comparative experiments of Section 6.3, we employ not only plain SMILES strings but also  
 1999 introduced structured molecular representations, such as explicit bond lists or atomic property ta-  
 2000 bles. Table 9 shows information on two input strategies from the same original catalyst dataset.  
 2001 These representations are theoretically equivalent to SMILES in information content, covering all  
 2002 connectivity information of the molecular structure. However, the structured format directly presents  
 2003 atomic indices, bond types and connectivity relationships, encoding the molecular graph topology as  
 2004 explicit data structures in the input. Since this representation aligns with LLMs’ parsing patterns for  
 2005 tabular and graph-structured data, it reduces the model’s need to infer relationships from character  
 2006 sequences, making it easier for the model to correctly understand and utilize the information.

2007  
 2008  
 2009  
 2010  
 2011  
 2012  
 2013  
 2014  
 2015  
 2016  
 2017  
 2018  
 2019  
 2020  
 2021  
 2022  
 2023

2024

<Catalytic chemical-text generation>Prompt

2025 You are an expert in materials science and chemistry. Your goal is to parse the un-  
 2026 structured, natural language text describing a catalytic material and convert it into a  
 2027 structured JSON object.

2028 1. Task requirements:

2029 - Analyze the [Material general text]: Carefully read the description of the catalytic  
 2030 material.

2031 - Identify Components: Extract the distinct chemical and structural components of the  
 2032 catalyst system, including:

2033 The primary catalytically Active Component (the core material driving the reaction).

2034 The Support or Substrate on which the active component is loaded.

2035 Any Dopants, Promoters, or Modifiers introduced to alter the properties.

2036 - Extract Key Features: Identify crucial characteristics for each component, such as  
 2037 material composition, morphology (e.g., nanoparticle, nanosheet, single-atom), crystal  
 2038 facets, size and specific properties (e.g., conductive, porous).

2039 - Generate Structured Formula: Create a concise, standardized formula-like string  
 2040 that represents the overall catalyst architecture (e.g., Active@Support, Dopant-  
 Active/Support).

2041 - Format as JSON: Organize all extracted information into a clean, hierarchical JSON  
 2042 object according to the specified Output Format. If a category is not mentioned in the  
 2043 text, use an empty list [].

2044 2. Example:

2045 Input:

2046 - Material general text: {example\_material\_general\_text}

2047 3. output format:

2048 - Material chemical text: {material.json}.

2049 - The output must be a single, valid JSON object with the following structure. Do not  
 2050 include any explanatory text outside of the JSON block.

2051 Input:

2052 - Material general text: {material\_general\_text}

2052  
2053

## E.3 PROMPTS FOR LLM-BASED EVALUATION

2054  
2055

## &lt;LLM-based evaluation&gt;Prompt

2056  
2057

Please complete the following tasks based on the truth QA pair and two answers by model A and model B provided below and its context:

2058  
2059  
2060

**1. Task requirements:**

- Extract core keywords related to the context of the regulation scheme from [Truth answer](e.g., reaction type, material system, innovative mechanisms, performance enhancement points, etc.).

- Extract the core content of the [Answer by model A] and [Answer by model B] without changing the original content.

- Based on the extracted keywords, Score the core content of the two answers according to the following scoring criteria. Provide specific details and explanations for any deductions.

**2. Scoring criteria:**

- Reasonableness: whether the answer is consistent with the description of the regulation method in the question. If the answer is consistent, 10 points will be given and if it is not, 0 points will be given.

- Accuracy: whether the answer is consistent with the key elements of the correct answer. 3 points if one key element is included, 6 points if two key elements are included and 10 points if all are consistent.

- Usability: whether the answer actually answers the question. If the response is unreliable which means it does not have any correct facts, 0 points will be given, If the response is reliable in its own right but does not match the question, 5 points will be awarded and if the response actually answers the question, 10 points will be awarded.

**2. Example:**

**Input:**

- Truth question: {example\_truth\_question}
- Truth answer: {example\_truth\_answer}
- Answer by model A: {example\_A\_answer}
- Answer by model B: {example\_B\_answer}

**3. output format:**

- Please rate the [Answer by model A] and [Answer by model B] out of 10 in terms of reasonableness, accuracy and usability. Reasonableness is 20%, accuracy is 50% and usability is 30%.

Reasonableness: [Score\_reasonableness]

Explanation for reasonableness: [Explanation\_reasonableness]

Accuracy: [Score\_accuracy]

Explanation for accuracy: [Explanation\_accuracy]

Usability: [Score\_usability]

Explanation for usability: [Explanation\_usability]

Total score: [Total\_score]

Analysis: [Analysis]

2094  
2095

## F STATEMENT ON THE USE OF LARGE LANGUAGE MODELS

2096  
2097

During the preparation of this manuscript, large language models (LLMs) were employed in a limited capacity to assist with linguistic refinement, grammar correction and improvement of clarity in the English prose. The scientific content, experimental design, data analysis and interpretation of results are entirely the work of the authors and no LLM was used for the generation of novel scientific claims or data fabrication. All factual statements, numerical results and interpretations have been manually verified by the authors to ensure accuracy and integrity. The assistance of LLMs was restricted to improving readability and cohesion and its use complies with prevailing academic publishing policies concerning AI-assisted writing.

2104

2105