From Fluent Text to Valid Science: Guideline-Grounded Refinement for Constrained Scientific Generation in CO_2RR Catalyst Design

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used for open-ended scientific design, but fluent proposals may still violate domain-specific scientific constraints. We study this problem as **constrained open-ended scientific generation**, where outputs must satisfy decomposable requirements rather than match a single reference answer. We propose a guideline-grounded verification-refinement method that represents domain knowledge as structured guidelines, diagnoses generated proposals through mechanistic diagnostic states over rule violations, mechanism consistency, and causal-chain completeness, and maps detected failure modes to targeted revision actions. We instantiate the method in CO_2 reduction reaction catalyst design, where proposals are evaluated for chemical-structure validity, mechanistic validity, and synthesis feasibility. We construct two complementary evaluation sets, including controlled synthetic tasks and real-world tasks derived from recent CO_2RR studies. Across multiple LLM backbones, guideline-grounded refinement consistently improves scientific validity over other baselines. Blinded expert evaluation, coverage-isolated testing, and prospective case studies further show that the proposed method produces more scientifically coherent and actionable catalyst proposals.

1 Introduction

Large language models (LLMs) have increasingly been used to support scientific research (Luo et al., 2025), hypothesis generation (Yang et al., 2024, 2025b) and material experimental design (Hu et al., 2025). In these settings, models can often produce fluent and plausible design rationales (Peters and Chin-Yee, 2025). However, scientific design is not a standard open-ended text generation task. A generated proposal may be well written while still violating known scientific principles. We view this setting as constrained open-ended scientific generation. This distinction is especially important in high-constraint scientific domains, where success depends on whether the proposal satisfies domain-specific constraints (Schmidgall et al., 2025).

We use CO_2 reduction reaction (CO_2RR) catalyst design as a representative testbed for constrained open-ended scientific generation (Gao et al., 2019). A valid catalyst design proposal must specify a material structure modification, a mechanistic rationale and reaction conditions (Yan et al., 2023). In this setting, LLM may (i) propose a material strategy that conflicts with established CO_2RR mechanisms, (ii) provide an explanation whose causal claims are internally inconsistent, or (iii) describe incorrect or incomplete synthesis steps. These errors are difficult to identify using general-purpose text evaluation methods (Zhu et al., 2024) or LLM-as-a-judge (Li et al., 2025).

Evaluating this type of generation is challenging for three reasons: 1) Catalyst design is open-ended. There is usually no single reference answer, and multiple distinct designs may be scientifically plausible; 2) Traditional text similarity metrics are poorly aligned with scientific validity, since a generated proposal can differ substantially from existing examples while remaining valid; 3) Generic self-refinement methods offer limited interpretability for scientific correction (Kamoi et al., 2024; Huang et al., 2024). Existing methods lack a validator that incorporates materials science knowledge.

We therefore formulate CO2RR catalyst design as an instance of constrained open-ended scientific generation, where the key criterion is satisfaction of verifiable scientific constraints (Han et al., 2026). As shown in Fig 1, the goal is to generate candidate catalyst designs that can be checked against decomposed scientific requirements. This perspective shifts the focus from holistic text quality to constraint-aware evaluation, making it possible to diagnose specific scientific failure modes and provide targeted feedback for revision. To evaluate different aspects of this formulation, we construct two

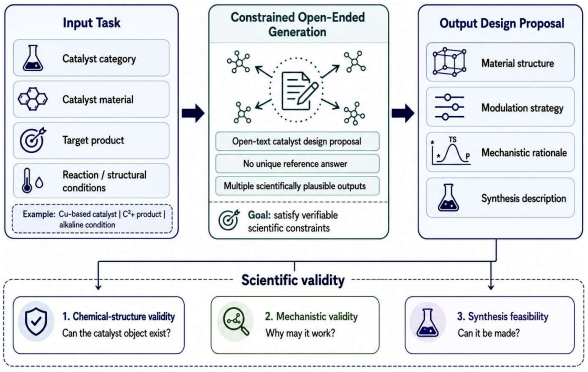


Figure 1: Task formulation for CO_2RR catalyst design.

complementary test sets. The **Synthetic Set** consists of 304 controlled CO_2RR design tasks with systematically specified material, product, and condition factors. The **Real-World Set** is derived from recent experimental CO_2RR studies consisting of 100 catalyst papers annotated by experts. For each paper, experts annotate a design target, key logic constraints, and the reported catalyst design. The goal is to generate scientifically plausible designs from the target alone.

In this work, we present a domain-adaptable guideline-grounded verification-refinement method for constrained scientific generation. We instantiate and evaluate it in CO_2RR catalyst design. The central role is to operationalize scientific validity as a set of decomposable and verifiable constraints. Specifically, generated design candidates are evaluated along three coupled dimensions: chemical-structure validity, **mechanistic validity**, and synthesis feasibility (Verboekend et al., 2016). Since mechanistic validity is central to catalyst design, we further decompose it into **rule violation**, **mechanism consistency**, and **causal-chain completeness**. Based on these diagnostic signals, the module uses a rule-driven feedback mapper to produce failure-mode-aware revision guidance. The feedback is not intended to provide a target answer or a new design directly. Instead, it identifies the type and approximate location of scientific weaknesses, such as explicit rule violations, logical inconsistencies, or missing causal links. This design preserves the open-ended nature of catalyst design while making refinement more interpretable and scientifically meaningful.

Overall, the contributions of this paper are:

1. We formulate constrained open-ended scientific generation and instantiate it in CO_2RR catalyst design with two complementary evaluation sets.

2. We propose a guideline-grounded verification-refinement method that operationalizes scientific validity through structured guidelines, mechanistic diagnostic states, causal-chain representations, and failure-mode-aware revision policies.

3. Experiments across multiple LLMs show consistent improvements over other baselines, with additional support from coverage-isolated evaluation, blinded expert assessment, and case studies.

2 Related Work

2.1 LLMs for Scientific and Catalyst Design

Large language models (LLMs) have increasingly been applied to scientific discovery and materials research (Guo et al., 2023; Boiko et al., 2023; Wang et al., 2025b). In chemistry and materials science, recent studies have shown that LLMs can help convert unstructured scientific literature into structured knowledge (Chen et al., 2024), build catalyst databases (Wang et al., 2023b), and support downstream reasoning over materials (Tang et al., 2025). Beyond knowledge extraction, LLMs and foundation models have also been explored for molecular generation (Zheng et al., 2025), materials design (Antunes et al., 2024; Sprueill et al., 2024), reaction prediction (Zhang et al., 2025), property estimation (Gao et al., 2023; Jablonka et al., 2024), catalyst filtering (Wang et al., 2025a), and scientific ideation (Marwitz et al., 2026). However, most existing work emphasizes retrieval quality, property prediction, and candidate screening. In high-constraint catalyst design, the central challenge is that a generated proposal may be fluent while still violating known mechanistic constraints. Our work addresses this gap by formulating LLM-based CO_2RR catalyst design as a constraint-aware scientific generation problem.

2.2 Evaluation for Open-Ended Generation

Evaluation and refinement of open-ended LLM generation have been widely studied. Traditional metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2019) mainly measure semantic similarity to reference texts. More recent LLM-as-a-judge methods provide flexible evaluation for open-ended outputs (Zheng et al., 2023; Wang et al., 2023a; Liu et al.,

2023; Chiang and Lee, 2023). However, these approaches typically focus on general factual consistency, rather than scientific constraints such as material structures, reaction mechanisms, and synthesis conditions. In parallel, self-refinement, reflection, and verifier-guided generation have shown that feedback can improve LLM outputs at training / inference time (Yao et al., 2022; Madaan et al., 2023; Shinn et al., 2023; Ke et al., 2024). Yet such feedback is often generic, making it difficult to distinguish scientific failure modes. Unlike these methods, we define task-specific mechanistic failure modes and connect them to structured CO_2RR guidelines. The resulting method is therefore a domain-grounded instantiation of constraint-aware generation for scientific design.

3 Methodology

We propose a guideline-grounded verification-refinement method for constrained scientific generation, and instantiate it in CO_2RR catalyst design, as shown in Fig. 2.

3.1 Task Formulation

We formulate CO_2RR catalyst design as an instance of *constrained open-ended scientific generation*. Formally, let X denote a design specification, D a generated proposal, and z its rationale. Given a guideline library $G = \{g_i\}_{i=1}^{N}$, we characterize scientific validity as a multi-dimensional constraint-satisfaction function, where each q_k measures one aspect of validity required by the task:

$$Q(D, z \mid X, G) = \{q_1, q_2, \ldots, q_K\}. \quad (1)$$

For CO_2RR catalyst design, the input is represented as $X = (c, m, p, r)$, where c denotes the catalyst category, m the catalyst material, p the target product, and r optional reaction conditions. The output proposal D is expected to describe the material structure, modulation strategy, mechanistic explanation, and feasible synthesis route. We instantiate the validity function as:

$$Q(D, z \mid X, G) = \{q_{\text{chem}}, q_{\text{mech}}, q_{\text{syn}}\}. \quad (2)$$

where q_{chem} measures chemical-structure validity, q_{mech} measures mechanistic validity, and q_{syn} measures synthesis feasibility.

3.2 Guideline-grounded Refinement

Fig. 2a illustrates the proposed guideline-grounded constraint refinement method. Given a catalyst design task X, the framework first retrieves factual catalyst examples as contextual evidence, generates an initial proposal, verifies the proposal against structured scientific guidelines, and then uses the diagnostic signals to guide iterative revision.

Factual retrieval and initial generation. For each input task X, the **Factual Example Retriever** retrieves a small set of relevant catalyst design cases $\mathscr{E}_X$ from a factual database. These cases contain reported CO_2RR catalyst designs and are used as few-shot demonstrations for the generator. The retrieved examples expose the model to plausible material modulation strategies, but are not used as final evaluation references. Given X and $\mathscr{E}_X$, the **Generator** produces an initial catalyst proposal:

$$D_0 = M_\theta(X, \mathscr{E}_X), \quad (3)$$

where M_θ denotes the backbone LLM. The proposal is expected to include the material structure, modulation strategy, mechanistic rationale, and synthesis route. Details of the factual database and retrieval procedure are provided in Appendix A, and prompts are provided in Appendix D.

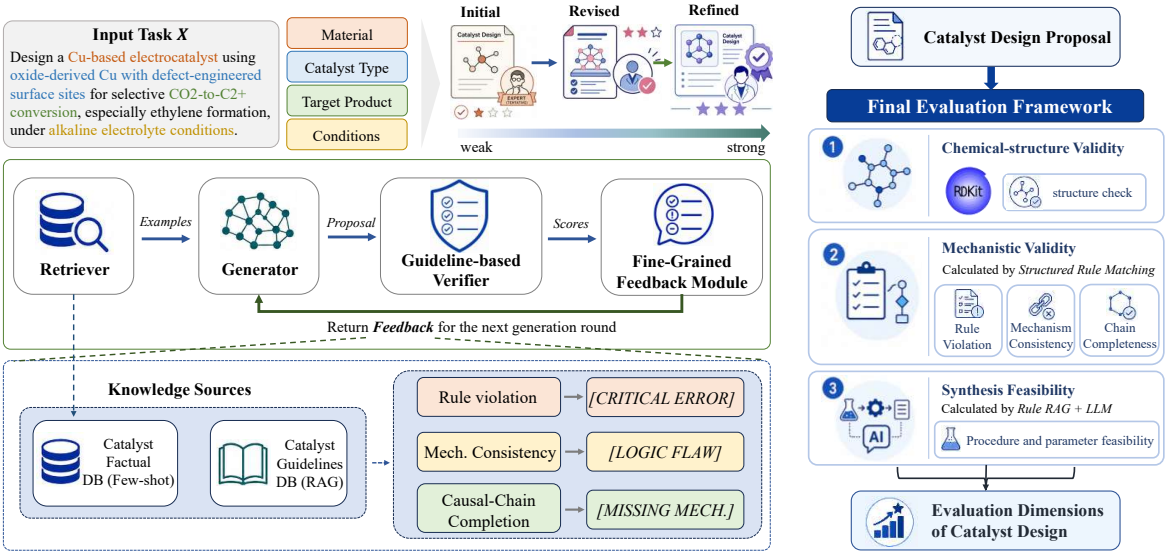
Guideline-based verification. The **Guideline-grounded Verifier** checks whether a generated proposal satisfies task-relevant scientific constraints. It operates on a structured guideline library constructed from textbooks, high-quality review articles, and CO_2RR domain literature. Each guideline records a textual constraint together with its material category, strategy category, catalytic property, condition constraint, source metadata, and guideline type. The guideline types include positive, negative, and conditional constraints, corresponding to supportive mechanisms, known mechanistic conflicts, and context-dependent principles.

Given a task X and proposal D_t, the verifier retrieves relevant guidelines $\mathscr{G}_X \subset \mathscr{G}$ and produces a mechanistic diagnostic state Δ_t. The diagnostic state summarizes rule violations, mechanism inconsistencies, missing causal links, and the corresponding evidence. Its detailed representation is introduced in Sec 3.3.

Iterative constraint refinement. The diagnostic state is converted into a revision action by a failure-mode-aware policy:

$$a_t = \pi(\Delta_t), \quad (4)$$

where a_t specifies the main type of scientific correction. The selected action is realized as



(a) Overview of the guideline-grounded verification-refinement method instantiated in CO_2RR catalyst design. The model generates an initial catalyst design proposal from the input task and retrieved factual examples. A guideline-grounded verifier diagnoses mechanistic weaknesses, and a fine-grained feedback module converts them into error-specific feedback signals for iterative refinement.

(b) The final evaluation framework. The decomposed scientific constraints include chemical-structure validity, mechanistic validity, and synthesis feasibility.

Figure 2: Overview of the guideline-grounded method and evaluation protocol for CO_2RR catalyst design.

a template-based feedback message grounded in $\mathscr{E}_{\text{diag}}$; the action space and policy are described in Sec 3.4. At refinement step t, the generator revises the previous proposal using the original task and the selected revision action:

$$D_{t+1} = M_\theta(X, \mathscr{E}_X, D_t, a_t). \tag{5}$$

The loop continues until the proposal satisfies the stopping criterion or the maximum number of iterations is reached.

3.3 Diagnostic State and Mechanistic Causal Chain

The verifier represents mechanistic validity as a diagnostic state. Given a proposal D_t, task X, and retrieved guidelines $\mathscr{G}_X$, it outputs:

$$\Delta_t = \left\{ v_{\text{rule}}, v_{\text{cons}}, v_{\text{chain}}, \mathscr{E}_{\text{diag}} \right\}, \tag{6}$$

where v_{rule} measures conflicts with scientific constraints, v_{cons} measures internal mechanism consistency, v_{chain} measures causal-chain completeness, and $\mathscr{E}_{\text{diag}}$ stores the supporting diagnostic evidence.

To make mechanistic diagnosis operational, we represent the rationale in a generated proposal as a set of causal links:

$$\hat{\mathscr{C}}(D_t) = \{u_1 \to u_2 \to \cdots \to u_n\}, \tag{7}$$

where each node u_i denotes a mechanistic state or effect, such as material modification or product selectivity. For each task X, the retrieved guidelines define a set of plausible mechanism chains:

$$\mathbb{C}_X = \{\mathscr{C}_X^{(1)}, \mathscr{C}_X^{(2)}, \ldots, \mathscr{C}_X^{(K)}\}, \tag{8}$$

where each $\mathscr{C}_X^{(k)}$ corresponds to one scientifically plausible route from material design to the target CO_2RR outcome. The verifier compares $\hat{\mathscr{C}}(D_t)$ with $\mathbb{C}_X$ and the retrieved guidelines to identify three types of mechanistic weaknesses:

- *contradicted edges*, where a causal link conflicts with a guideline;
- *unsupported edges*, where a claimed causal relation lacks mechanistic support;
- *missing edges*, where the proposal skips an intermediate link needed to justify the claimed catalytic effect.

These edge-level signals are used to compute the three diagnostic dimensions in Eq. 6. The detailed scoring definitions are provided in Sec 3.5.

3.4 Failure-Mode-Aware Revision Policy

Unlike generic self-refinement, the action specifies which scientific weakness should be corrected. We consider three feedback types. [CRITICAL ERROR] is triggered by explicit rule violations and asks the generator to remove or correct conflicting mechanistic claims. [LOGIC FLAW] is triggered by low mechanism

consistency and asks the generator to repair unsupported or internally inconsistent causal relations. [MISSING MECHANISM] is triggered by low causal-chain completeness and asks the generator to add missing intermediate mechanism links.

$$\pi(\Delta_t) = \begin{cases} a_{\text{conflict}}, & v_{\text{rule}} > \tau_{\text{rule}}, \\ a_{\text{repair}}, & v_{\text{cons}} < \tau_{\text{cons}}, \\ a_{\text{complete}}, & v_{\text{chain}} < \tau_{\text{chain}}, \\ a_{\text{stop}}, & \text{otherwise}. \end{cases} \tag{9}$$

The selected action is grounded in diagnostic evidence $\mathscr{E}_{\text{diag}}$ and instantiated as template-based feedback. When multiple conditions are satisfied, the policy generates composite feedback while preserving the priority order in Eq. 9. The feedback generation is detailed in Appendix F.

3.5 Evaluation Framework

In final assessment, we evaluate generated proposals using the same constraint-oriented view introduced in Sec 3.3. Specifically, we decompose scientific validity into three complementary dimensions: chemical-structure validity, mechanistic validity, and synthesis feasibility, as shown in Fig. 2b.

3.5.1 Chemical-structure validity.

Chemical-structure validity measures whether the proposed catalyst object is chemically meaningful and structurally interpretable. Because CO_2RR catalysts are often heterogeneous materials rather than small molecules, we use rule-based material consistency checks for extended catalysts and apply RDKit-based checks only to molecular-like components. Details are provided in Appendix E.1.

3.5.2 Mechanistic Validity

Mechanistic validity evaluates whether the proposal provides a coherent causal explanation from material design to CO_2RR outcome. Following the diagnostic representation in Sec 3.3, we assess three subdimensions: rule violation, mechanism consistency, and causal-chain completeness.

Rule Violation. Rule violation measures whether a generated proposal explicitly conflicts with known mechanistic constraints:

$$v_{\text{rule}}(D \mid X) = \frac{1}{|\mathscr{G}_X|} \sum_{g_i \in \mathscr{G}_X} \mathbb{I}_{\text{viol}}(R, g_i), \tag{10}$$

where $\mathbb{I}_{\text{viol}}(R, g_i)$ indicates whether the structured proposal representation R violates guideline g_i. A higher value indicates more severe conflict with known mechanistic constraints.

Mechanism Consistency. Mechanism consistency measures whether the material structure, modulation strategy, and claimed catalytic effect form an internally coherent rationale:

$$c_{\text{cons}} = f_{\text{cons}}(R_{\text{mech}}). \tag{11}$$

The corresponding score v_{cons} is derived from the predicted class and confidence. A higher score indicates that the proposal provides a more internally consistent mechanistic explanation.

Causal-Chain Completeness. Causal-chain completeness evaluates whether the generated mechanism chain covers the key intermediate links of at least one plausible route defined in Sec 3.3. Given the causal links extracted from the generated proposal, denoted as $\hat{\mathscr{C}}(D)$, and the task-specific candidate chains $\mathbb{C}_X$, we define:

$$v_{\text{chain}}(D \mid X) = \max_{\mathscr{C} \in \mathbb{C}_X} \frac{|\hat{\mathscr{C}}(D) \cap \mathscr{C}|}{|\mathscr{C}|}. \tag{12}$$

A higher score indicates that the proposal covers more necessary links of at least one plausible mechanistic route, while still allowing multiple scientifically valid mechanisms for the same design target.

Coverage-isolated verification and evaluation. To avoid evaluator-refiner circularity, we separate the guideline library into refinement-time guidelines and held-out evaluation constraints. The refinement-time verifier only accesses high-level mechanism guidelines, while the final evaluator uses disjoint fine-grained constraints over material, strategy, product, and condition. Final rule text, rule IDs, matched evidence, and reference designs are never exposed during generation or feedback. Details of the guideline split, overlap filtering, manual audit, and feedback masking are provided in Appendix E.5.

3.5.3 Synthesis Feasibility

Synthesis feasibility evaluates whether a proposed catalyst design is supported by an executable preparation route. Given the synthesis description in proposal D, we first parse it into a structured protocol:

$$P_{\text{syn}} = \phi_{\text{syn}}(D) = \{(o_i, m_i, c_i, \theta_i)\}_{i=1}^{L}, \tag{13}$$

where o_i, m_i, c_i, and θ_i denote the operation, material, condition, and process parameters. A Rule RAG + LLM assessor compares P_{syn} with retrieved synthesis rules $\mathscr{R}_{\text{syn}}(D)$ and assigns:

$$q_{\text{syn}}(D) = f_{\text{syn}}\left(D, P_{\text{syn}}, R_{\text{syn}}(D)\right). \tag{14}$$

Details of the synthesis rule database, error types, and evaluation is detailed in Appendix E.3.

4 Experiments

4.1 Datasets

Synthetic Set. The Synthetic Set contains 304 exploratory Cu-based CO_2RR catalyst design tasks. Each task is specified by a set of structured fields, including the catalyst category, catalyst material, target product, and optional reaction conditions. The tasks are manually combined to form controlled design settings (Saha et al., 2022). This set is designed to stress-test whether generated proposals satisfy known mechanistic constraints under controlled design conditions. Detailed dataset construction and distributions are provided in Appendix C.1 and C.2.

Real-World Set. The Real-World Set is constructed from 100 high-impact CO_2RR experimental papers published in 2024 and 2025. For each paper, we derive a realistic catalyst design task from the Introduction section, which typically describes the scientific motivation, design challenge, and target catalytic objective. Additional statistics are provided in Appendix C.3 and C.4.

4.2 Evaluation Metrics

We use the evaluation framework in Sec. 3.5. For the Real-World Set, we additionally report a post-hoc **soft matching rate** against expert-annotated key logic constraints, detailed in Appendix E.4.

4.3 Baselines and Hyperparameters

Baselines. We compare our method against a representative set of strong baselines under different generation strategies. Specifically, we consider ChemLLM (Zhang et al., 2024) and CatalystLLM (Chen et al., 2026) as domain-specific models; DeepSeek-v3.1 (Guo et al., 2025) and Qwen3-8B (Yang et al., 2025a) as open-source models; and Gemini-2.5-pro (Comanici et al., 2025) and GPT-4o (Achiam et al., 2023) as closed-source models. For each model, we evaluate four generation strategies: Generate-Only, RAG-Only, Best-of-N, and our guideline-grounded refinement procedure. We also consider several stronger refinement-style baselines that are commonly used in open-ended LLM generation in Sec. 5.2.

Hyperparameters. We finally use CatalystLLM-7B as the main backbone generator for primary experiments. For each input task, the retriever returns the top-$k_f = 3$ most relevant catalyst design cases from the factual database. For guideline-grounded verification, the verifier retrieves the top-$k_g = 5$ relevant guidelines. The refinement process is performed within a maximum of $T = 3$ refinement iterations. Early stopping is applied when the proposal satisfies the predefined mechanistic validity threshold, denoted by τ_{mech}. For synthesis-feasibility assessment, the Rule RAG module retrieves the top-$k_s = 8$ relevant synthesis rules, which are then provided to the GPT-5.2 assessor.

For the Best-of-N baseline, we generate $N = 10$ candidate proposals for each task and select the candidate with the highest verifier score. To reduce randomness, all reported results are averaged over 3 independent runs. For all baselines, we use the same prompts, retrieval settings, and coverage-isolated evaluation protocol.

5 Main Results

All automatic results are reported under the coverage-isolated protocol described in Sec.3.5.2. Our analysis focuses on three research questions:

- **RQ1.** Does verifier-guided refinement improve scientific validity of design proposals?
- **RQ2.** Which components contribute to the improvement?
- **RQ3.** How does performance change with inference cost and iteration budget?

5.1 RQ1. Guideline-grounded Refinement Improves Scientific Validity

Table 1 and Table 2 report the main results on the Synthetic Set and Real-World Set. Across all models and datasets, our guideline-grounded refinement method consistently improves scientific validity over generate-only, RAG-only, and Best-of-N settings. The gains are observed for both general models and domain-specialized models. Strong general models achieve competitive performance under RAG-only settings, but still benefit from guideline-grounded refinement. Among domain-specialized models, CatalystLLM-7B benefits substantially from refinement and achieves particularly strong performance on rule violation and causal-chain completeness.

To obtain an external final-output assessment independent of our automatic scores, we conduct a

Model	Chemical-structure Validity ↑				Rule Violation ↓				Mechanism Consistency ↑				Causal-chain Completeness ↑				Synthesis Feasibility ↑			
	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours
General Models																				
Qwen3-8B	0.649	0.785	0.906	**0.930**	0.464	0.412	0.214	**0.163**	0.593	0.857	0.871	**0.902**	0.101	0.401	0.462	**0.542**	0.652	0.756	0.804	**0.853**
DeepSeek-v3.1	0.703	0.899	0.929	**0.948**	0.431	0.279	0.181	**0.098**	0.636	0.866	0.912	**0.938**	0.110	0.451	0.517	**0.579**	0.722	0.792	0.834	**0.869**
GPT-4o	0.708	0.902	0.934	**0.940**	0.466	0.401	0.202	**0.152**	0.602	0.864	0.881	**0.922**	0.108	0.468	0.528	**0.582**	0.746	0.827	0.865	**0.898**
Gemini-2.5-pro	0.712	0.936	0.940	**0.942**	0.422	0.258	0.167	**0.116**	0.658	0.890	0.926	**0.954**	0.117	0.479	0.539	**0.591**	0.764	0.839	0.882	**0.912**
Domain-Specialized Models																				
ChemLLM-7B	0.724	0.923	0.949	**0.960**	0.398	0.142	0.102	**0.054**	0.542	0.848	0.865	**0.904**	0.100	0.412	0.473	**0.551**	0.492	0.706	0.758	**0.795**
CatalystLLM-7B	0.727	0.921	0.962	**0.969**	0.371	0.129	0.064	**0.014**	0.562	0.885	0.892	**0.951**	0.129	0.531	0.547	**0.639**	0.498	0.819	0.861	**0.907**

Table 1: Main results on the Synthetic Set. Each metric is reported under four generation settings: generate-only, RAG-only, Best-of-N, and our method. The standard deviations are provided in Table 14.

Model	Soft Match ↑				Rule Violation ↓				Mechanism Consistency ↑				Causal-chain Completeness ↑				Synthesis Feasibility ↑			
	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours
General Models																				
Qwen3-8B	0.06	0.40	0.52	**0.66**	0.499	0.436	0.241	**0.175**	0.514	0.813	0.846	**0.887**	0.088	0.372	0.436	**0.515**	0.618	0.712	0.775	**0.828**
DeepSeek-v3.1	0.11	0.57	0.69	**0.82**	0.451	0.298	0.196	**0.112**	0.553	0.836	0.881	**0.914**	0.097	0.426	0.490	**0.566**	0.728	0.804	0.852	**0.894**
GPT-4o	0.13	0.64	0.75	**0.85**	0.484	0.319	0.216	**0.165**	0.543	0.822	0.862	**0.892**	0.092	0.327	0.453	**0.525**	0.712	0.801	0.851	**0.892**
Gemini-2.5-pro	0.15	0.68	0.78	**0.88**	0.441	0.279	0.181	**0.132**	0.582	0.861	0.901	**0.928**	0.106	0.452	0.513	**0.586**	0.735	0.813	0.861	**0.901**
Domain-Specialized Models																				
ChemLLM-7B	0.10	0.55	0.66	**0.81**	0.417	0.163	0.121	**0.076**	0.462	0.802	0.838	**0.889**	0.101	0.413	0.474	**0.561**	0.445	0.758	0.812	**0.861**
CatalystLLM-7B	0.14	0.62	0.73	**0.87**	0.388	0.126	0.090	**0.042**	0.490	0.845	0.885	**0.927**	0.114	0.502	0.539	**0.646**	0.483	0.812	0.849	**0.889**

Table 2: Main results on the Real-World Set. Each metric is reported under four generation settings: generate-only, RAG-only, Best-of-N, and our method. The standard deviations are provided in Table 15.

blinded expert preference study. This preference study is conducted on all Synthetic Set tasks by five CO_2RR experts. For each selected task, experts compare outputs generated by Generate-only, Best-of-N, and our guideline-grounded refinement method. The outputs are anonymized and randomly shuffled, and experts are asked to rank them according to overall scientific validity. Fig. 3 shows both the overall ranking distribution and the dimension-level pairwise preferences. Experts most frequently rank our method as the best output among the compared methods. This suggests that the gains observed in automatic evaluation are reflected in human scientific assessment. Because final scoring uses held-out fine-grained constraints that are not available during refinement, these improvements cannot be explained by direct optimization against the same guideline entries.

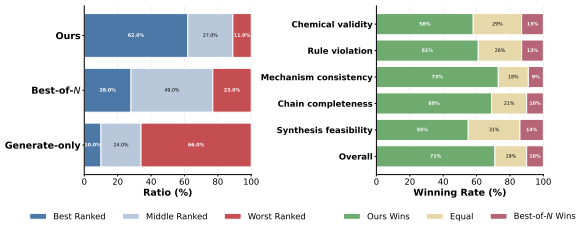


Figure 3: Blinded expert evaluation on Synthetic Set.

To demonstrate that the capabilities of our method extend beyond a single evaluation framework, we conduct additional experiments on CatalystBench (Chen et al., 2026), a Q&A catalyst comprehension benchmark, as shown in Table 19.

5.2 RQ2. Ablation Study

To understand which components of the guideline-grounded framework are responsible for the improvement, we ablate factual retrieval, guideline-based verification, decomposed scoring, and failure-mode-aware feedback.

Table 3 shows that each component contributes to the overall mechanism-validity score (The aggregation formula is provided in Appendix E.2). The retrieved examples help initial generation but cannot replace refinement. The largest drops occur when the verifier signals become less structured. Removing fine-grained feedback reduces the score to 0.831, and removing multi-dimensional scoring further reduces it to 0.782. Replacing the guideline-grounded verifier with a general LLM-as-a-judge also leads to a lower score of 0.794. These results indicate that mechanistic refinement benefits from decomposed guideline-grounded diagnostics.

5.3 RQ3. Inference Cost Analysis

We analyze how mechanistic validity changes with the number of refinement iterations. The initial Factual-only output is treated as iteration 0, and each subsequent round applies verification and revision before generating the next proposal. As shown in Fig. 4, our method improves rapidly in the first few rounds, increasing from 0.79 to over 0.88 within three rounds. Compared with the

Variant	Factual Retrieval	Verifier	Multi-dim. Scoring	Fine-grained Feedback	Mech. Score ↑
Full Method	✓	Guideline	✓	✓	**0.882**
w/o Factual Retrieval		Guideline	✓	✓	0.842
Generic Self-Refine (same budget)	✓	Guideline	✓		0.831
w/o Decomposed Scoring	✓	Guideline			0.782
Domain-Prompted LLM-as-a-Judge	✓	GPT-5.2	✓		0.794
Factual-only	✓				0.541
Generate-only					0.464

Table 3: Ablation study of the proposed method on the overall mechanism-validity score.

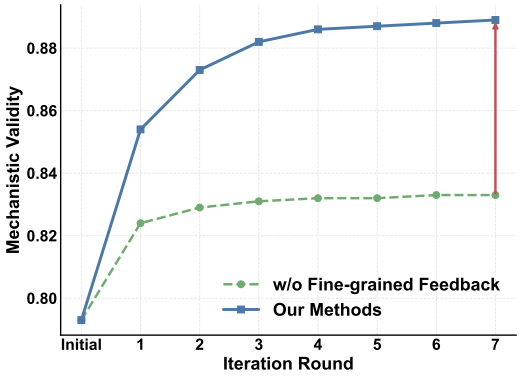


Figure 4: Mechanistic validity across iterations.

Method	Calls	Total Tokens	Time (s)	Relative Cost
Generate-only	1	3,950	50.4	1.0×
Factual-only	1	5,420	58.7	1.4×
B-o-N (N=10)	10	39,800	431.6	10.1×
Ours (1 iter)	2	9,180	109.5	2.3×
Ours (2 iter)	3	14,260	163.8	3.6×
Ours (3 iter)	4	19,740	221.4	5.0×

Table 4: Computational inference cost comparison under Gemini-2.5-pro as the backbone generator. Relative cost is normalized by Generate-only.

variant without fine-grained feedback, our method achieves consistently higher mechanistic validity and a larger improvement margin.

Table 4 reports the inference cost under Gemini-2.5-pro. As expected, refinement introduces additional cost, because each iteration requires an additional verification-and-revision step. However, the cost remains substantially lower than Best-of-N. This shows that iterative refinement achieves stronger mechanistic validity with a relatively lower cost. In the main experiments, we therefore use a limited iteration budget to balance performance and inference cost. Additional experiment results are provided in Appendix G.

5.4 Human and Experimental Validation

To ensure the reliability of our automatic verifier, we conduct a blinded expert agreement study. Five CO_2RR electrocatalysis experts evaluate 180 proposals sampled from Generate-only, Factual-only, Best-of-N, and our refinement method, together with expert-constructed plausible innovation and invalid cases. The verifier achieves **0.86** Top-1 ranking accuracy and **0.82** pairwise accuracy against expert preferences, with Pearson correlations of **0.86** on overall mechanistic validity. Expert annotations also show strong inter-rater reliability, with an **ICC of 0.82** for overall mechanistic validity. We provide the full annotation protocol and details in Appendix H.

Beyond text-level validation, we further conduct two prospective wet-lab case studies on catalyst designs that are, to the best of our knowledge, previously unreported under comparable CO_2RR settings, as detailed in Appendix I.1. Each case reports the generated catalyst structure, predicted mechanism chain, key experimental result, and supporting characterization evidence. Appendix I also includes iterative refinement examples and failure case analysis to illustrate both the correction process and the remaining limitations of the framework.

6 Conclusion

We formulate LLM-based CO_2RR catalyst design as a constrained open-ended generation problem, where the central challenge is satisfaction of verifiable scientific constraints. To this end, we propose a guideline-grounded method that combines structured scientific guidelines, mechanistic diagnostic states, causal-chain representations, and failure-mode-aware revision policies. We design a decomposed evaluation framework for open-ended scientific design proposals. Experiments show that our method consistently improves scientific validity over generation baselines. Coverage-isolated evaluation, expert assessment, and prospective wet-lab cases further support that the improvements reflect more than direct optimization against the verifier. To support reproducibility, we provide the guideline metadata and code at `https://anonymous.4open.science/r/guideline-C388/`.

Limitations

Although our method improves the scientific validity of LLM-generated CO_2RR catalyst design proposals, several limitations remain.

First, the method depends on the coverage and quality of the guideline library. Although we construct the library from textbooks, review articles, and domain literature with expert quality control, it may still miss emerging mechanisms and rare catalyst systems. Future work can expand the guideline library to more catalyst systems and continuously update it with new literature. To support reproducibility, we will also release the derived guideline library, factual example database, Synthetic Set, and Real-World Set annotations. Due to copyright restrictions, we will provide source identifiers, structured fields, and split metadata of source literature.

Second, our large-scale evaluation focuses on text-level catalyst design proposals rather than systematic experimental validation. Although we provide two prospective wet-lab validation cases, they are intended as preliminary evidence of experimental actionability rather than exhaustive validation of all generated proposals.

Third, the method is developed and evaluated primarily in the context of electrocatalyst design. The proposed method is domain-adaptable. Applying it to other scientific design tasks requires constructing domain-specific guideline schemas, validity dimensions, and failure-mode definitions. Future work can adapt the same verification-refinement protocol to other materials design and scientific discovery tasks, such as battery materials and photocatalysts.

Ethical Considerations

This work studies LLM-assisted scientific design for CO_2RR catalyst proposals. The proposed framework is intended to improve the scientific validity of generated hypotheses, not to replace expert judgment or experimental validation. All generated catalyst designs should be treated as preliminary hypotheses and require further expert review and experimental or computational verification before practical use.

A potential risk of LLM-based scientific generation is that fluent but invalid proposals may be mistaken for reliable scientific knowledge. We mitigate this risk by grounding verification in structured domain guidelines, decomposing scientific validity into interpretable dimensions, and reporting coverage-isolated evaluation and expert agreement. Nevertheless, the verifier remains limited by the coverage and quality of the guideline library and may miss emerging mechanisms or rare catalyst systems.

The factual examples, guideline library, and Real-World Set are constructed from scientific literature and expert annotation. To reduce data leakage, we separate factual retrieval, refinement-time guidelines, and held-out evaluation constraints. For anonymous review, we release only anonymized sample data, prompt templates, and reference code; fuller resources will be released after publication where copyright and anonymity constraints permit.

The expert evaluation involves annotators with CO_2RR background and does not involve personal or sensitive data. Results are reported only in aggregate. Finally, although CO_2RR catalyst design is relevant to sustainable energy, repeated LLM generation incurs computational cost. We therefore report inference-cost analysis and use a limited iteration budget in the main experiments.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

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

Xueqing Chen, Jian Xu, Ludi Wang, Yang Gao, Huihan Zhu, Yuanchun Zhou, Yi Du, and Cheng-Lin Liu. 2026. Catalystbench: A comprehensive multi-task benchmark for advancing language models in catalysis science. In *The Fourteenth International Conference on Learning Representations*, volume 2026.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into automatic evaluation using large language models. *arXiv preprint arXiv:2310.05657*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, and Evan Rosen. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Dunfeng Gao, Rosa M Arán-Ais, Hyo Sang Jeon, and Beatriz Roldan Cuenya. 2019. Rational catalyst and electrolyte design for co2 electroreduction towards multicarbon products. *Nature Catalysis*, 2(3):198–210.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, and Xiao Bi. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in neural information processing systems*, 36:59662–59688.

Taolin Han, Shuang Wu, Jinghang Wang, Yuhao Zhou, Renquan Lv, Bing Zhao, and Wei Hu. 2026. Molquest: A benchmark for agentic evaluation of abductive reasoning in chemical structure elucidation. *arXiv preprint arXiv:2603.25253*.

Zhaolin Hu, Yixiao Zhou, Zhongan Wang, Xin Li, Weimin Yang, Hehe Fan, and Yi Yang. 2025. Osda agent: Leveraging large language models for de novo design of organic structure directing agents. In *The thirteenth international conference on learning representations*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pages 32808–32824.

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

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, and Hongning Wang. 2024. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13034–13054.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, and Tianhao Wu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, and Yiming Yang. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Thomas Marwitz, Alexander Colsmann, Ben Breitung, Christoph Brabec, Christoph Kirchlechner, Eva Blasco, Gabriel Cadilha Marques, Horst Hahn, Michael Hirtz, and Pavel A Levkin. 2026. Predicting new research directions in materials science using large language models and concept graphs. *Nature Machine Intelligence*, pages 1–10.

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

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

Uwe Peters and Benjamin Chin-Yee. 2025. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4).

Paramita Saha, Sk Amanullah, and Abhishek Dey. 2022. Selectivity in electrochemical co2 reduction. *Accounts of chemical research*, 55(2):134–144.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. 2024. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *arXiv preprint arXiv:2402.10980*.

Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, and Yilun Zhao. 2025. Chemagent: Self-updating library in large language models improves chemical reasoning. *arXiv preprint arXiv:2501.06590*.

Danny Verboekend, Nicolas Nuttens, Roel Locus, Joost Van Aelst, Patrick Verolme, Johan C Groen, Javier Pérez-Ramírez, and Bert F Sels. 2016. Synthesis, characterisation, and catalytic evaluation of hierarchical faujasite zeolites: milestones, challenges, and future directions. *Chemical Society Reviews*, 45(12):3331–3352.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11.

Ludi Wang, Xueqing Chen, Yi Du, Yuanchun Zhou, Yang Gao, and Wenjuan Cui. 2025a. Catalm: empowering catalyst design through large language models. *International Journal of Machine Learning and Cybernetics*, 16(5):3681–3691.

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

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299.

Xinming Wang, Jian Xu, Aslan H Feng, Yi Chen, Haiyang Guo, Fei Zhu, Yuanqi Shao, Minsi Ren, Hongzhu Yi, and Sheng Lian. 2025b. The hitchhiker's guide to autonomous research: A survey of scientific agents. *Authorea Preprints*, pages 1–24.

Tianxiang Yan, Xiaoyi Chen, Lata Kumari, Jianlong Lin, Minglu Li, Qun Fan, Haoyuan Chi, Thomas J Meyer, Sheng Zhang, and Xinbin Ma. 2023. Multiscale co2 electrocatalysis to c2+ products: reaction mechanisms, catalyst design, and device fabrication. *Chemical Reviews*, 123(17):10530–10583.

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

Hengpan Yang, Qing Lin, Chao Zhang, Xinyao Yu, Zhong Cheng, Guodong Li, Qi Hu, Xiangzhong Ren, Qianling Zhang, Jianhong Liu, and 1 others. 2020. Carbon dioxide electroreduction on single-atom nickel decorated carbon membranes with industry compatible current densities. *Nature communications*, 11(1):593.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13545–13565.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. 2025b. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *International Conference on Learning Representations*, volume 2025, pages 33251–33277.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yu Zhang, Yang Han, Shuai Chen, Ruijie Yu, Xin Zhao, Xianbin Liu, Kaipeng Zeng, Mengdi Yu, Jidong Tian, and Feng Zhu. 2025. Large language models to accelerate organic chemistry synthesis. *Nature Machine Intelligence*, 7(7):1010–1022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, and Eric Xing. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui

Pan. 2025. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, 7(3):437–447.

Xianrui Zhong, Yufeng Du, Siru Ouyang, Ming Zhong, Tingfeng Luo, Qirong Ho, Hao Peng, Heng Ji, and Jiawei Han. 2024. Actionie: Action extraction from scientific literature with programming languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12656–12671.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7b: Improving helpfulness and harmlessness with rlaif. In *First Conference on Language Modeling*.

A Retriever Details

The retriever provides factual catalyst design examples for initial generation. Its role is to supply the generator with domain-relevant design context. In our method, the factual retriever is separated from the guideline-grounded verifier: the factual database stores concrete catalyst design cases, whereas the guideline library stores transferable scientific constraints used for verification and feedback.

A.1 Factual Example Database

We construct a factual example database from reported CO_2RR catalyst design cases in the literature. Specifically, we collect 1,208 high-quality journal articles published between 2020 and 2023 and extract concrete catalyst design instances from them. This time range is used to provide representative prior design knowledge while avoiding overlap with the Real-World Set. Each entry corresponds to a concrete catalyst design instance and is stored with structured fields, including catalyst category, catalyst material, material structure, modulation strategy, target product, reaction conditions, mechanistic rationale, and synthesis description. Table 5 represents a factual case example. These fields allow the retriever to identify examples that are relevant to the input task and to present them in a format suitable for prompting the generator. The final factual database contains 1,589 catalyst design cases.

The database is used only to support initial generation. Given a new design task, retrieved examples provide the model with plausible design patterns, such as common material modifications, product-oriented modulation strategies, and structure–property relationships. The retrieved examples serve as contextual demonstrations that help the generator produce a scientifically grounded initial proposal.

A.2 Retrieval Representation and Similarity Computation

Given an input task $X = (c, m, p, r)$, where c denotes the catalyst category, m denotes the catalyst material, p denotes the target product, and r denotes optional reaction or structural conditions, the retriever ranks factual examples according to their relevance to X. Each database entry is represented as $E_i = (c_i, m_i, a_i, p_i, r_i, h_i, s_i)$, where a_i denotes the modulation strategy, h_i denotes the mechanistic ra-

Field	Example Entry
Catalyst category	Cu-based CO_2RR electrocatalyst
Catalyst material	Oxide-derived Cu with Zn incorporation
Material structure	Porous Cu surface containing dispersed Zn species and abundant grain boundaries
Modulation strategy	Heteroatom incorporation and defect-rich surface engineering
Target product	C_2 products, mainly ethylene
Reaction conditions	Alkaline electrolyte; CO_2-saturated environment
Mechanistic rationale	Zn incorporation and oxide-derived surface reconstruction can modify the local electronic structure of Cu sites, tune *CO adsorption strength, and increase the local coverage of carbon-containing intermediates. The defect-rich Cu surface and grain boundaries may further promote *CO retention and C–C coupling, thereby improving selectivity toward C_2 products.
Synthesis description	Cu and Zn precursors are deposited on a conductive substrate, followed by thermal oxidation or electrochemical oxidation to form a mixed oxide surface. The catalyst is then electrochemically reduced under CO_2RR conditions to generate an oxide-derived Cu–Zn surface with porous morphology and defect-rich active sites.
Source doi	10.1002/adfm.202210938 (Yang et al., 2020)

Table 5: Example entry in the factual catalyst design database. Each entry is stored with structured fields and used as contextual evidence for initial generation.

tionale, and s_i denotes the synthesis description.

For retrieval, we convert both the input task and each factual example into text representations by concatenating their structured fields. We use CatBERT (Ock et al., 2023) as the encoder, a Transformer model based on the RoBERTa architecture, specifically designed for predicting adsorption energies in catalytic systems. The semantic relevance score of x and e_i is computed by cosine similarity:

$$S_{\text{sem}}(X, E_i) = \frac{\mathbf{x}^\top \mathbf{e}_i}{\|\mathbf{x}\|_2 \|\mathbf{e}_i\|_2}. \tag{15}$$

To better reflect the structured nature of catalyst design tasks, we combine semantic similarity with field-level matching signals:

$$S_{\text{field}}(X, E_i) = \sum_{j \in \mathscr{F}} w_j \cdot \text{sim}_j(X, E_i), \tag{16}$$

where $\mathscr{F} = \{c, m, p, r\}$ denotes the set of task fields, and w_j is the weight for field j with $\sum_{j \in \mathscr{F}} w_j = 1$. In our implementation, we assign larger weights to fields that more directly determine the design goal: $w_c = 0.25$, $w_m = 0.25$, $w_p = 0.35$, and $w_r = 0.15$. For categorical fields such as catalyst category and target product, $\text{sim}_j(\cdot)$ is computed as ontology-normalized matching. For material and condition fields, we use normalized token overlap with synonym normalization, allowing partial matches between related material forms or reaction conditions. The final retrieval score combines semantic relevance and field-level compatibility:

$$S(X, E_i) = \lambda S_{\text{sem}}(X, E_i) + (1 - \lambda) S_{\text{field}}(X, E_i). \tag{17}$$

We set $\lambda = 0.7$ in all experiments, giving higher priority to CatBERT-based semantic relevance while retaining explicit control over key task fields. The final retrieved set is defined as the top-k examples under $S(X, E_i)$. The retrieved examples $\mathscr{E}_X$ are then formatted as contextual demonstrations for the generator. Each example is presented with its task specification and corresponding design proposal, including material structure, modulation strategy, mechanistic rationale, and synthesis description.

B Guideline Library Construction

We construct a scientific guideline library to support mechanistic verification, error diagnosis, and feedback generation in CO_2RR catalyst design. Unlike the factual example database, the guideline library stores transferable scientific constraints. These constraints describe general relationships among catalyst materials, structural modulation strategies, reaction intermediates, product selectivity, and synthesis-related conditions. Fig. 5 illustrates this process.

B.1 Source Selection

The guideline library is built from textbooks, high-quality review articles, and domain literature. We select sources that provide reliable and transferable knowledge for CO_2RR catalyst design. The source pool includes textbooks on electrochemistry and

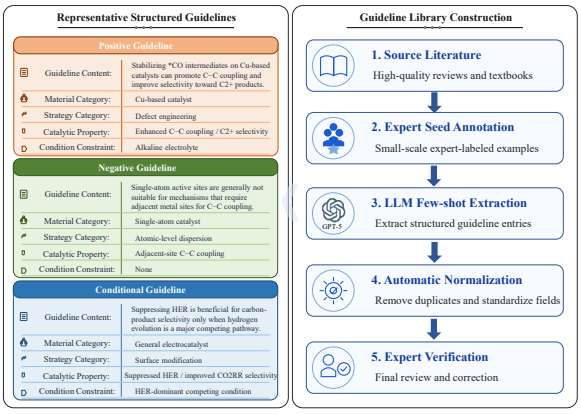


Figure 5: Example guideline entry and construction pipeline of the scientific guideline library.

catalysis, high-quality review articles on CO_2 reduction, and relevant domain literature. To avoid leakage into the Real-World Set, we ensure that the guideline source pool is isolated from the 2024–2025 papers used to construct the Real-World Set. Specifically, none of the Real-World Set source papers is included in the guideline extraction corpus. In addition, review articles and domain papers used for guideline extraction are screened to avoid extracting design-specific principles from the same catalyst studies that appear in the Real-World Set. Thus, the guideline library provides general mechanistic knowledge rather than direct access to the target designs or annotated logic constraints in the Real-World Set.

The source selection follows four criteria. First, the source should be scientifically reliable, such as a textbook, a high-impact review, or a well-cited domain paper. Second, the content should contain generalizable knowledge that can be applied beyond a single catalyst instance. Third, the described principle should be relevant to at least one component of the design task, such as material category, modulation strategy, catalytic property, reaction condition, or target product. Fourth, the source should not overlap with the 2024–2025 Real-World Set or expose the reported catalyst design or experimental evidence used for Real-World Set annotation. Table 6 summarizes the source composition used for guideline extraction.

B.2 Semi-Automatic Extraction Pipeline

We use a semi-automatic pipeline to extract structured guidelines from the selected sources. As shown in Fig. 5, the extraction process consists of four steps. First, domain experts annotate 60 seed examples from representative sources. Each annotated example identifies a candidate guideline and maps it to the required structured fields, including guideline content, material category, strategy category, catalytic property, condition constraint, and guideline type. These seed examples are then used to guide GPT-5.2 in extracting candidate guidelines from additional source passages.

Second, GPT-5.2 extracts candidate guidelines and converts them into structured entries. During this step, the model is instructed to preserve the scientific meaning of the source text while avoiding overly broad or unsupported statements. Each extracted entry is also assigned a guideline type: positive, negative, or conditional. Positive guidelines describe strategies or mechanisms that support desirable catalytic behavior, negative guidelines describe conflict conditions, and conditional guidelines specify principles that hold only under certain material or reaction settings.

Third, the extracted candidates are normalized and filtered. We standardize material names, target products, modulation strategies, and condition descriptions to reduce duplicates and improve retrieval consistency. We also remove entries that are redundant, too generic, or unsupported by the source passage. When multiple extracted entries describe the same principle, they are merged into a single normalized guideline.

Finally, a subset of the extracted guidelines is reviewed by domain experts for quality control. Experts check whether the guideline is scientifically valid, whether the assigned fields are appropriate, and whether the guideline can be used to support mechanistic verification. This process results in a structured guideline library that can be used for guideline-grounded verification and fine-grained feedback generation.

Source Type	Number	Role in Guideline Extraction
Textbooks	6	Foundational catalysis principles
Review articles	18	General CO_2RR mechanisms
Domain papers	256	Strategy-level evidence
Total	280	Source pool

Table 6: Statistics of sources used for guideline library construction.

B.3 Guideline Schema

As shown in Fig. 5, each guideline is stored as a structured entry so that it can be used for verification. Table 7 summarizes the main fields used

in the guideline library. During verification, the task specification and the parsed design proposal are matched against these fields to retrieve relevant constraints. For example, a proposal involving Cu-based catalysts and C_2 product formation can be matched to guidelines related to Cu surface structure, *CO adsorption, C–C coupling, and electrolyte-dependent selectivity. The structured fields also support error diagnosis by identifying whether a detected issue is associated with modulation strategy.

B.4 Guideline Types

As shown in Fig. 5, we categorize guidelines into three types: positive, negative, and conditional. This categorization is used to distinguish different roles of scientific constraints during verification.

Positive guidelines. Positive guidelines describe mechanisms, structural features, or design strategies that are expected to support desirable catalytic behavior. For example, a guideline may state that increasing local *CO coverage on suitable Cu sites can promote C–C coupling and improve C_2 product selectivity. Such guidelines provide supportive evidence for generated mechanism claims.

Negative guidelines. Negative guidelines describe known conflicts, unfavorable design choices, or mechanistic contradictions. They are mainly used to identify explicit rule violations. For example, a generated proposal may be flagged if it claims a product-selectivity effect that contradicts known intermediate adsorption behavior. Negative guidelines therefore provide the main basis for detecting [CRITICAL ERROR] feedback.

Conditional guidelines. Conditional guidelines describe principles that hold only under specific material, reaction, or structural conditions. For example, a modulation strategy may improve selectivity only for a certain catalyst family. These guidelines are important because many CO_2RR design principles are context-dependent. During verification, conditional guidelines are applied only when their condition constraints are satisfied.

B.5 Guideline Library Statistics

The final guideline library contains approximately 4,327 structured guidelines. Figure 6 summarizes the distribution of guidelines by guideline type, material category, modulation strategy, and task purpose. In terms of guideline type, positive and conditional guidelines account for the majority, while negative guidelines form a smaller but important subset. For material coverage, Cu-based catalysts occupy the largest portion, reflecting the central role of Cu in CO_2RR, especially for C_{2+} product formation. The library also includes other metal catalysts, single-atom catalysts, oxide or hydroxide catalysts, carbon-supported catalysts, and general mechanism guidelines, which helps the verifier handle diverse material settings beyond Cu-based systems.

For modulation strategy, the distribution is relatively balanced across different strategies. The fourth distribution shows the task-purpose split for guideline library construction. This partition is directly related to the verifier and final evaluation design. High-level mechanism guidelines are mainly used during the feedback-based verification stage to provide general mechanistic diagnostic signals, while fine-grained guidelines are reserved for more detailed final assessment.

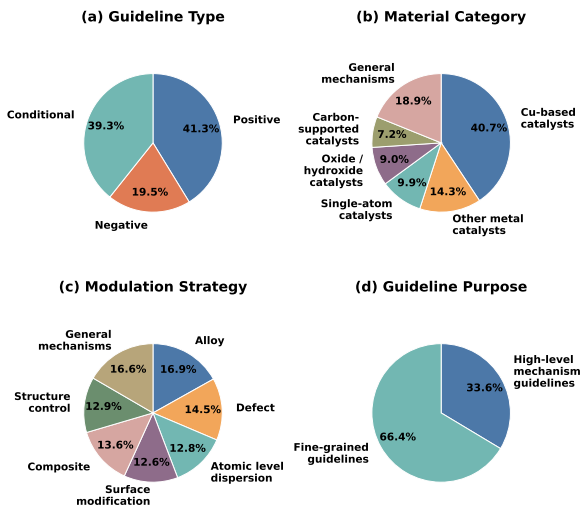


Figure 6: Statistics of the guideline library. We report the distribution of guidelines by (a) guideline type; (b) material category; (c) modulation strategy; and (d) task purpose, distinguishing high-level mechanism guidelines from fine-grained guidelines.

B.6 Expert Review and Quality Control

We perform expert review and quality control at two stages of guideline library construction: seed annotation and post-extraction validation. During the seed annotation phase, we invite two domain experts to manually annotate a small sample of representative source text paragraphs. These expert-labeled examples are used to calibrate the extrac-

Field	Description
Guideline content	The natural-language scientific principle, mechanism constraint, or design rule.
Material category	The catalyst family to which the guideline applies, such as Cu-based catalysts, oxide-derived catalysts, or carbon-supported catalysts.
Strategy category	The modulation strategy involved in the guideline, such as alloy, defect, atomic level dispersion, composite, surface modification, or structure control.
Catalytic property	The catalytic property described by the guideline, such as intermediate adsorption, C-C coupling, product selectivity, or faradaic efficiency.
Condition constraint	The reaction condition under which the guideline applies, such as electrolyte, pH, applied potential, or catalyst morphology.
Source metadata	The source document used for traceability and quality control.

Table 7: Structured schema of a guideline entry.

tion prompt and provide demonstrations for GPT-5.2-assisted guideline extraction.

After GPT-5.2 extracts candidate guidelines from the selected sources, we conduct sampling-based expert validation. The team of experts consists of the two experts mentioned earlier and four doctoral students. Experts evaluate each sampled guideline according to three criteria: scientific correctness, support from the source passage, and correctness of structured field assignments. A guideline is considered valid only if it preserves the meaning of the original source and can be used as a verification constraint for catalyst design. On the expert-reviewed subset with 1,000 samples, the average acceptance rate of GPT-5.2-extracted guidelines is 89.6%. The average accuracy is 91.8% for guideline content correctness, 90.4% for source support consistency, and 89.7% for guideline type assignment. Most errors arise from over-generalized statements, unsupported causal claims, and ambiguity between positive and conditional guidelines.

Based on the review results, we apply three quality-control operations to the full library. Unsupported or scientifically incorrect guidelines are removed, ambiguous field assignments are corrected using normalized material, strategy, and condition vocabularies, and near-duplicate guidelines are merged to reduce redundancy. After quality control, the remaining guidelines form the structured guideline library used for verification-feedback generation and final evaluation.

C Dataset Construction

This section provides additional details on the construction of the two evaluation datasets used in this work: the **Synthetic Set** and the **Real-World Set**. The two datasets are designed for complementary purposes. The Synthetic Set provides controlled catalyst design tasks for testing whether generated proposals satisfy known mechanistic constraints, while the Real-World Set provides realistic design goals derived from recent experimental papers for evaluating open-ended catalyst design quality.

C.1 Synthetic Set Construction

The Synthetic Set contains 304 exploratory Cu-based CO_2RR catalyst design tasks. We construct the Synthetic Set by combining catalyst categories, material systems, target products, and reaction conditions that commonly appear in CO_2RR catalyst design. Specifically, we manually define a set of valid design factors, such as Cu-based catalysts, oxide-derived Cu, Cu alloys, defect-rich Cu surfaces, Cu-based tandem catalysts, and supported Cu structures. These factors are then combined with target products such as CO, CH_4, C_2H_4, ethanol, and broader C_{2+} products under relevant conditions including alkaline electrolyte, neutral electrolyte, local pH regulation, surface oxidation state control, and morphology constraints.

After generating candidate task specifications, we filter them according to two criteria. First, each task should be scientifically meaningful and correspond to a plausible CO_2RR design scenario. Second, each task should involve identifiable mechanistic constraints, so that generated proposals can be evaluated. The Synthetic Set is not intended as a reference-matching benchmark. For a given task, multiple catalyst designs may be scientifically plausible. Therefore, the evaluation focuses on our designed evaluation metrics: chemical-structure validity, mechanistic validity, and synthesis feasibility.

C.2 Dataset Distribution of Synthetic Set

To characterize the coverage of the Synthetic Set, we analyze its distribution across catalyst cate-

gories and target products. As shown in Fig. 7, the horizontal axis corresponds to Cu-based catalyst design categories, including oxide-derived Cu, Cu alloys, defect-rich Cu, Cu-based tandem catalysts, supported Cu catalysts, interface-engineered Cu catalysts, and other Cu-based designs. The vertical axis corresponds to target CO_2RR products, including CO, CH_4, C_2H_4, ethanol, and broader C_{2+} products. Each cell reports the number of design tasks associated with a specific catalyst-category and product pair.

C.3 Real-World Set Construction

The Real-World Set is constructed to evaluate whether generated catalyst design proposals remain scientifically reasonable under realistic and open-ended CO_2RR design scenarios. It is derived from recent experimental studies and reflects research questions that appear in actual catalyst design papers. We collect 100 high-impact CO_2RR experimental papers published in 2024–2025. For each paper, we construct one design task based on the scientific motivation and design objective described in the introduction section. This setting is intended to simulate an early-stage catalyst ideation scenario, where the model is asked to propose a scientifically plausible design given a realistic research objective.

For each instance, we further annotate three types of reference information from the source paper. First, we record the actual modulation strategy adopted in the paper, such as alloying, defect engineering, or local environment regulation. Second, we summarize the core mechanistic rationale reported by the authors, including the intended active-site effect, intermediate regulation, reaction pathway modulation, or product-selectivity mechanism. Third, domain experts annotate a set of key logic constraints that a reasonable proposal should satisfy. These constraints are used for soft matching and mechanistic evaluation.

In addition to the soft match rate, we also focus on the validity of the evaluation scheme itself. Multiple designs may be scientifically plausible for the same research goal. Therefore, evaluation also focuses on whether the generated proposal aligns with the expert-annotated design logic, avoids mechanistic contradictions, provides a coherent causal explanation, and remains synthetically feasible. This makes the Real-World Set a realistic test bed for open-ended CO_2RR catalyst design.

C.4 Dataset Distribution of Real-World Set

To characterize the coverage of the Real-World Set, we analyze the 100 design tasks from six perspectives: target product, catalyst material category, modulation strategy, mechanistic focus, experimental configuration, and publication year. The resulting statistics are shown in Fig. 8. These distributions provide a more detailed view of the diversity of realistic CO_2RR catalyst design scenarios covered by the benchmark.

D Prompt Templates

This section provides the prompt templates used for LLM generation, baseline evaluation and synthesis feasibility evaluation. In our method, retrieval, guideline matching, and feedback construction are implemented as modules separated from the generator. Therefore, we report the prompts used for initial generation, revision, and LLM-based baselines.

D.1 Initial Generation Prompt

Initial Generation Prompt

You are an expert in CO_2 reduction reaction (CO_2RR) electrocatalyst design.
Given the catalyst design task and retrieved factual catalyst design examples, generate a scientifically plausible catalyst design proposal. Use the retrieved examples as design context, but do not simply copy them.
Input Task:

- Catalyst category: `<CATEGORY>`
- Catalyst material: `<MATERIAL>`
- Target product: `<TARGET_PRODUCT>`
- Reaction conditions: `<CONDITIONS>`

Retrieved Factual Examples:

- `<RETRIEVED_EXAMPLE_1>`
- `<RETRIEVED_EXAMPLE_2>`
- `<RETRIEVED_EXAMPLE_3>`

Generate a catalyst design proposal that includes:

- the proposed material structure;
- the modulation strategy;
- the mechanistic rationale;
- a feasible synthesis route when applicable.

The proposal should be chemically meaningful, mechanistically coherent, and experimentally plausible.
Output Format:

1. Material structure:
2. Modulation strategy:
3. Mechanistic rationale:
4. Synthesis route:

D.2 Revision Prompt

Revision Prompt

You are an expert in CO_2 reduction reaction (CO_2RR) electrocatalyst design.
- Revise the previous catalyst design proposal according to the diagnostic feedback. - The feedback is generated by a rule-based verifier and identifies the main scientific weakness of the proposal. - Preserve scientifically reasonable parts of the original proposal and revise only the parts that need correction.
Input Task:

- Catalyst category: `<CATEGORY>`
- Catalyst material: `<MATERIAL>`
- Target product: `<PRODUCT>`
- Reaction conditions: `<CONDITIONS>`

Previous Proposal: `<PREVIOUS_PROPOSAL>`
Diagnostic Feedback:
`<RULE_BASED_FEEDBACK>`
Please revise the proposal by:

- removing claims that conflict with known CO_2RR mechanisms;
- improving consistency between material structure, modulation strategy, and target product;
- adding missing intermediate mechanistic links when needed;
- keeping the design experimentally plausible.

Do not merely improve wording. Focus on scientific correction.
Output Format:

1. Material structure:
2. Modulation strategy:
3. Mechanistic rationale:
4. Synthesis route:

D.3 Baseline Prompts

Generate-Only Baseline Prompt

You are an expert in CO_2 reduction reaction (CO_2RR) electrocatalyst design.
Given the following catalyst design task, generate a scientifically plausible catalyst design proposal.
Input Task:

- Catalyst category: <CATEGORY>
- Catalyst material: <MATERIAL>
- Target product: <PRODUCT>
- Reaction conditions: <CONDITIONS>

The proposal should include:

- material structure;
- modulation strategy;
- mechanistic rationale;
- synthesis route.

Output Format:

1. Material structure:
2. Modulation strategy:
3. Mechanistic rationale:
4. Synthesis route:

RAG-only Baseline Prompt

You are an expert in CO_2 reduction reaction (CO_2RR) electrocatalyst design.
Given the catalyst design task, retrieved factual examples, and retrieved scientific guidelines, generate a catalyst design proposal. Use the retrieved content as background knowledge, but do not copy it verbatim.
Input Task:

- Catalyst category: <CATEGORY>
- Catalyst material: <MATERIAL>
- Target product: <PRODUCT>
- Reaction conditions: <CONDITIONS>

Retrieved Factual Examples:

- <RETRIEVED_EXAMPLE_1>
- <RETRIEVED_EXAMPLE_2>
- <RETRIEVED_EXAMPLE_3>

Retrieved Scientific Guidelines:

- <GUIDELINE_1>
- <GUIDELINE_2>
- <GUIDELINE_3>
- <GUIDELINE_4>
- <GUIDELINE_5>

Generate a catalyst design proposal that is chemically meaningful, mechanistically coherent, and experimentally plausible.
Output Format:

1. Material structure:
2. Modulation strategy:
3. Mechanistic rationale:
4. Synthesis route:

Best-of-*N* Baseline Prompt

You are an expert in CO_2 reduction reaction (CO_2RR) electrocatalyst design.
Given the catalyst design task, retrieved factual examples, and retrieved scientific guidelines, generate one candidate catalyst design proposal. The proposal should be scientifically plausible and should include material structure, modulation strategy, mechanistic rationale, and synthesis route.
Input Task:

- Catalyst category: `<CATEGORY>`
- Catalyst material: `<MATERIAL>`
- Target product: `<PRODUCT>`
- Reaction conditions: `<CONDITIONS>`

Retrieved Factual Examples:

- `<RETRIEVED_EXAMPLE_1>`
- `<RETRIEVED_EXAMPLE_2>`
- `<RETRIEVED_EXAMPLE_3>`

Retrieved Scientific Guidelines:

- `<GUIDELINE_1>`
- `<GUIDELINE_2>`
- `<GUIDELINE_3>`
- `<GUIDELINE_4>`
- `<GUIDELINE_5>`

Generate one candidate proposal. This prompt is sampled N times independently. The final candidate is selected by the verifier according to the mechanistic validity score.
Output Format:

1. Material structure:
2. Modulation strategy:
3. Mechanistic rationale:
4. Synthesis route:

LLM-as-a-judge Prompt

You are a general scientific evaluator for CO_2RR catalyst design.
Evaluate the following catalyst design proposal and assign an overall mechanistic validity score from 0 to 1. Focus on whether the proposal is scientifically reasonable, internally consistent, and sufficiently explains the connection between material design and CO_2RR performance.
Input Task:

- Catalyst category: `<CATEGORY>`
- Catalyst material: `<MATERIAL>`
- Target product: `<PRODUCT>`
- Reaction conditions: `<CONDITIONS>`

Generated Proposal: `<PROPOSAL>`
Please evaluate the proposal according to:

- whether it violates known CO_2RR mechanistic principles;
- whether the mechanism explanation is internally consistent;
- whether it provides a complete causal chain from material modification to catalytic outcome.

Output Format:

```
Rule violation score: <0-1>
Mechanism consistency score: <0-1>
Causal-chain completeness score: <0-1>
Overall score: <0-1>
Main issue: <brief description>
Explanation: <brief explanation>
```

D.4 Synthesis Feasibility Evaluation Prompts

Synthesis Feasibility Assessment Prompt

You are an expert evaluator for CO_2RR catalyst synthesis protocols.
Your task is to assess whether the generated synthesis route is experimentally feasible. You are given:

- the catalyst design task;
- the generated synthesis description;
- the structured synthesis protocol parsed from the description;
- retrieved synthesis rules.

Catalyst Design Task: `<INPUT_TASK>`
Generated Synthesis Description:
`<SYNTHESIS_DESCRIPTION>`
Parsed Synthesis Protocol:

```
Step 1: <operation, material, condition>
Step 2: <operation, material, condition>
...
```

Retrieved Synthesis Rules or Templates:

- `<RULE_1>`
- `<RULE_2>` ...
- `<RULE_8>`

Evaluate the synthesis route according to the following error types:

- **Missing operation**: required synthesis steps are omitted.
- **Incorrect operation**: an operation is chemically or experimentally inappropriate.
- **Missing parameter**: key parameters such as temperature, time, concentration, potential, atmosphere, or pH are absent.
- **Invalid parameter**: a provided parameter is unrealistic or outside a reasonable range.
- **Hallucinated step**: the synthesis contains fictitious or unsupported procedures that do not align with established catalyst preparation practices.

Use the retrieved rules as the primary evidence. Do not judge only by textual fluency. If the route is partially feasible, identify the main weakness and explain how it should be corrected.
Output Format:
Judgement: <Feasible / Partially feasible / Infeasible>
Synthesis feasibility score: <0-1>
Detected error types: <Missing operation / Incorrect operation / Missing parameter / Invalid parameter / Hallucinated step / None>
Rule evidence: <briefly cite the matched rule or template>
Correction suggestion: <concise correction-oriented suggestion>
Explanation: <brief explanation grounded in the retrieved rules>

E Evaluation Details

This section provides additional details for metrics that require task-specific evaluation procedures. The definitions of three submetrics are provided in Sec 3.5. Here, we describe the detailed chemical-structure validity, the overall mechanistic validity score, the synthesis feasibility assessment process, and the soft match evaluation used for the Real-World Set.

E.1 Chemical-Structure Validity

Chemical-structure validity evaluates whether the proposed catalyst object is chemically meaningful and structurally interpretable. In CO_2RR catalyst design, the proposed object is often not a standard small molecule, but a heterogeneous material involving surfaces, defects, dopants, interfaces, supports, oxide-derived structures, or reconstructed active sites. Therefore, this metric serves as a necessary material-object validity check that identifies proposals with invalid compositions, incompatible material categories, uninterpretable structural descriptions, or chemically implausible modification operations.

Given a generated proposal D, we first extract a structured material representation:

$$R_{\text{chem}} = \{m, c, a, u, i, r\}, \tag{18}$$

where m denotes the base material, c denotes the catalyst category, a denotes the modification operation, u denotes defect units, i denotes interface or support components, and r denotes relevant reaction conditions. The chemical-structure validity score is then computed as:

$$q_{\text{chem}}(D) = f_{\text{chem}}(R_{\text{chem}}), \tag{19}$$

where $f_{\text{chem}}(\cdot)$ aggregates material-level consistency checks rather than relying on a single molecular validity tool. Specifically, the validator checks whether the proposed material components are recognizable chemical entities, whether the stated catalyst category is compatible with the material composition, whether dopants, defects, supports, and interfaces are chemically plausible, and whether the described structural modification is meaningful for the given material class.

For catalyst proposals, RDKit-based checks are only applied to molecular or molecular-like components, such as organic modifiers, molecular-like precursors, ligands, or small molecular fragments involved in synthesis. In these cases, RDKit helps

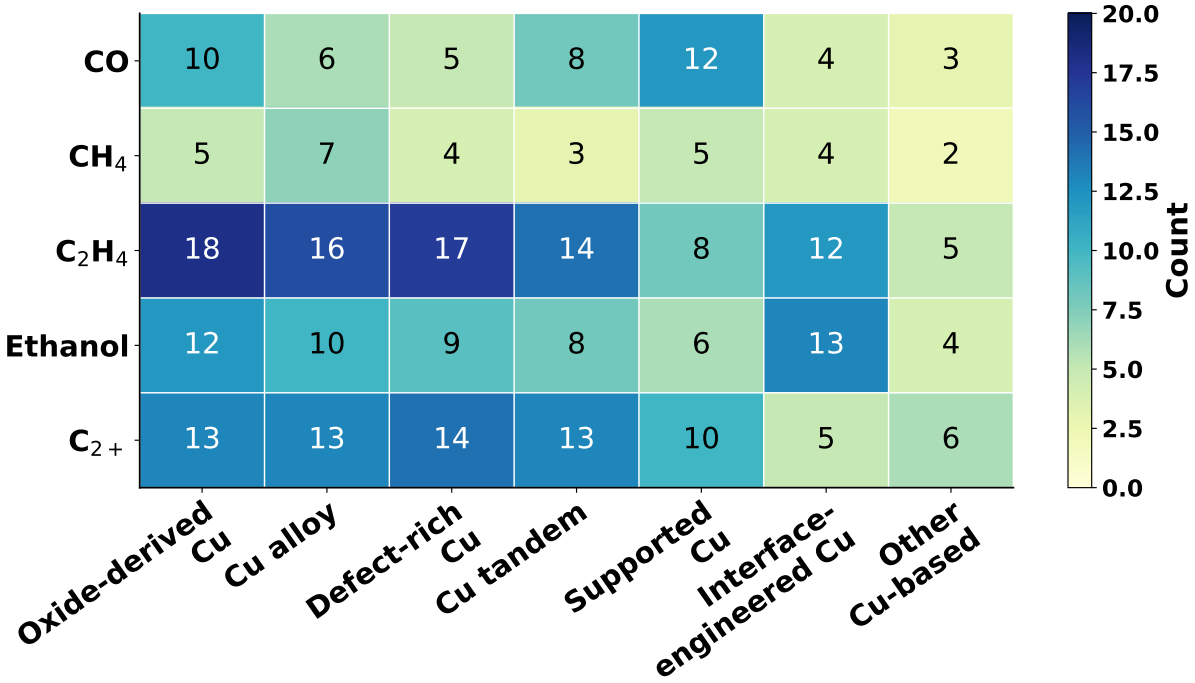


Figure 7: Distribution of the Synthetic Set across target products and catalyst categories. Rows correspond to target CO_2RR products, and columns correspond to Cu-based catalyst design categories. Each cell reports the number of design tasks in the corresponding product-category group.

verify molecular parseability, valence validity, connectivity, and basic descriptor consistency. For extended catalysts, surfaces, alloys, oxides, and defect-rich materials, the score is instead determined by rule-based material consistency checks over normalized catalyst entities and structural descriptors.

E.2 Overall Mechanistic Validity Score

The overall mechanistic validity score aggregates the three mechanistic submetrics: rule violation, mechanism consistency, and causal-chain completeness. Since rule violation is a negative metric, we convert it into a rule-compliance term before aggregation. Given a generated proposal D for task X, the overall mechanistic validity score is computed as:

$$\begin{aligned} S_{\text{mech}}(D \mid X) = {} & w_{\text{rule}} \cdot (1 - v_{\text{rule}}(D \mid X)) \\ & + w_{\text{cons}} \cdot v_{\text{cons}}(D \mid X) \\ & + w_{\text{chain}} \cdot v_{\text{chain}}(D \mid X) \end{aligned} \tag{20}$$

where v_{rule} is the rule violation score, v_{cons} is the mechanism consistency score, and v_{chain} is the causal-chain completeness score. In our main experiments, we set the weight as $w_{\text{rule}} = 0.40$, $w_{\text{cons}} = 0.35$, and $w_{\text{chain}} = 0.25$. We assign the largest weight to rule compliance because explicit contradictions with known CO_2RR mechanisms are considered the most severe failure mode. Mechanism consistency receives the second largest weight because a proposal must provide an internally coherent rationale across material structure, modulation strategy, target product, and claimed catalytic effect. Causal-chain completeness is also important, but it is weighted slightly lower because some valid proposals may present a concise mechanism without explicitly enumerating every intermediate link.

The weights are determined based on expert calibration over a small development subset. Domain experts first rank candidate proposals for the same task according to overall scientific reasonableness. We then select the weight combination that best preserves expert preferences while maintaining interpretability of the three submetrics. We also verify that the main conclusions are stable under nearby weight settings in Table 8.

E.3 Synthesis Feasibility

Synthesis feasibility evaluates whether the proposed catalyst design is supported by an experimentally executable preparation route. This metric is used as a final assessment dimension, which checks whether the generated synthesis description

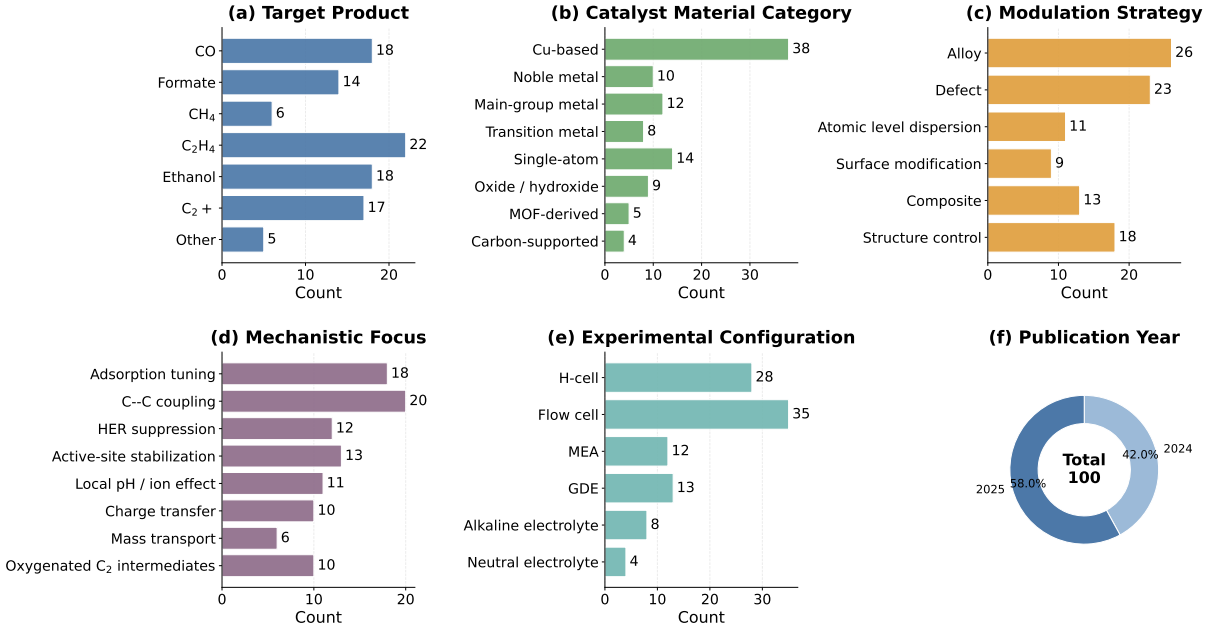


Figure 8: Distribution of the Real-World Set from multiple perspectives.

w_{rule}	w_{cons}	w_{chain}	RAG	Ours	Δ
0.35	0.40	0.25	0.774	**0.871**	+0.097
0.40	0.35	0.25	0.781	**0.878**	+0.097
0.45	0.30	0.25	0.789	**0.884**	+0.095
0.40	0.30	0.30	0.776	**0.881**	+0.105
0.35	0.35	0.30	0.778	**0.875**	+0.097

Table 8: Sensitivity analysis of the mechanism-validity score under nearby weight settings. The performance gap between Ours and RAG-Only remains stable.

is sufficiently plausible for evaluating a catalyst proposal.

Fig. 9 describes the whole evaluation procedure. First, the synthesis description in the generated proposal is parsed into a structured protocol (Zhong et al., 2024):

$$P_{\text{syn}} = \{(o_i, m_i, c_i, \theta_i)\}_{i=1}^{L}, \tag{21}$$

where o_i denotes the operation, m_i denotes the material or reagent, c_i denotes the condition, and θ_i denotes process parameters at step i. The parser identifies common preparation operations such as precursor dissolution, deposition, hydrothermal treatment, and so on.

Second, the system retrieves relevant synthesis rules or templates from the synthesis rule database. These rules describe required operations, valid operation sequences, and reasonable parameter ranges for common CO_2RR catalyst preparation procedures. The retrieved rule set is denoted as $\mathscr{R}_{\text{syn}}(D)$.

Third, the structured protocol is compared with the retrieved rules. We check five common error types: missing key operations, incorrect operation order, missing process parameters, invalid parameter values, and fictitious synthesis steps. The original synthesis description, the parsed protocol, and the matched rules are then provided to GPT-5.2 to assign a feasibility score $q_{\text{syn}}(D)$. The score is normalized to $[0, 1]$, where a higher value indicates a more executable synthesis route. The evaluation prompt is provided in Appendix D.4.

We further evaluate different LLM-based synthesis feasibility assessors on the synthesis-feasibility test set. The evaluation covers five common error types in catalyst preparation protocols: missing operation, incorrect operation, missing parameter, invalid parameter, and hallucinated step. For each error type, we report three metrics: Judgement accuracy (J), Correction effectiveness (C), and Explanation quality (E). J measures whether the assessor correctly identifies the feasibility issue, C measures whether the suggested correction can effectively address the detected error, and E measures the clarity and scientific grounding of the explanation based on human expert scoring.

Table 9 shows that rule-grounded assessment improves synthesis feasibility evaluation across nearly all error types and metrics. Compared with direct GPT-5.2 assessment, Rule + GPT-5.2 achieves higher judgement accuracy, correction effectiveness, and explanation quality in all five categories.

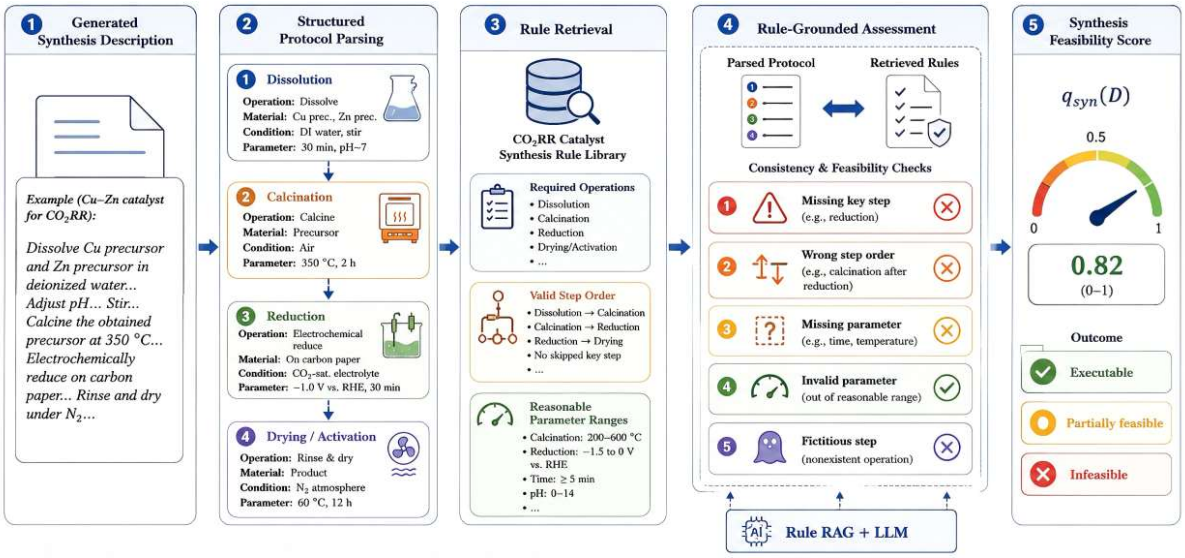


Figure 9: Workflow of the synthesis feasibility assessment.

Evaluator	Missing Operation			Incorrect Operation			Missing Parameter			Invalid Parameter			Hallucinated Step		
	J	*C*	*E*	*J*	*C*	*E*	*J*	*C*	*E*	*J*	*C*	*E*	*J*	*C*	*E*
GPT-5.2	0.89	0.79	0.92	0.92	0.87	0.90	0.82	0.51	0.65	0.65	0.59	0.68	0.54	0.35	0.49
Gemini-2.5-Flash	0.86	0.71	0.84	0.89	0.81	0.83	0.80	0.35	0.57	0.52	0.46	0.55	0.35	0.28	0.38
Rule + GPT-5.2	**0.92**	**0.88**	**0.95**	**0.94**	**0.90**	**0.94**	**0.89**	**0.63**	**0.82**	**0.82**	**0.66**	**0.84**	**0.62**	**0.48**	**0.61**
Rule + Gemini-2.5-Flash	0.90	0.86	0.92	0.92	**0.90**	0.91	0.85	0.59	0.77	0.81	0.64	0.80	0.42	0.34	0.47

Table 9: Evaluation of synthesis feasibility assessors across five synthesis error types. *J* denotes judgement accuracy, *C* denotes correction effectiveness, and *E* denotes explanation quality scored by human experts.

The improvement is particularly clear for missing parameter and invalid parameter errors, where retrieved rules provide explicit constraints on required parameters and reasonable value ranges. Rule grounding also improves hallucinated-step detection, although this remains the most challenging category because fictitious procedures can be phrased fluently and may require broader synthesis knowledge to identify.

The comparison between Gemini-2.5-Flash and Rule + Gemini-2.5-Flash shows a similar trend, indicating that the benefit comes from rule grounding rather than only from using a stronger backbone model. Overall, these results support our use of Rule RAG + LLM as the synthesis feasibility assessment module, since it provides more reliable error judgement, more actionable correction suggestions, and better grounded explanations than direct LLM assessment.

E.4 Soft Match Evaluation

The Real-World Set is not evaluated as a single-reference generation benchmark. Since multiple catalyst designs may be scientifically plausible for the same research goal, we use soft match evaluation to assess whether a generated proposal aligns with the key design logic of the source paper at a conceptual level. This method of calculating correlation scores through the matching of logical terms has been validated in the work of (Wang et al., 2024) and (Yang et al., 2025b).

For each Real-World Set instance, domain experts annotate a set of key logic constraints:

$$\mathscr{L}_X = \{\ell_1, \ell_2, \ldots, \ell_M\}, \tag{22}$$

where each ℓ_j describes an essential design principle, such as the intended modulation strategy, target intermediate regulation, active-site construction, local environment effect, or product-oriented mechanism. Given a generated proposal D, we extract its design logic statements:

$$\hat{\mathscr{L}}(D) = \{\hat{\ell}_1, \hat{\ell}_2, \ldots, \hat{\ell}_N\}. \tag{23}$$

We compute the soft match score as the fraction of expert-annotated logic constraints that are matched by the generated proposal:

$$S_{\text{soft}}(D \mid X) = \frac{1}{|\mathscr{L}_X|} \sum_{\ell_j \in \mathscr{L}_X} \mathbb{I}_{\text{match}}(\ell_j, \hat{\mathscr{L}}(D)). \quad (24)$$

The matching indicator $\mathbb{I}_{\text{match}}$ is set to 1 when the generated proposal expresses a scientifically equivalent or compatible design logic.

E.5 Coverage-Isolated Evaluation Protocol

A potential concern of guideline-grounded refinement is evaluator–refiner circularity. If the same guideline entries are used both to generate refinement feedback and to compute final scores, the improvement may partly reflect optimization against the evaluator rather than genuine improvement in scientific validity. To reduce this risk, all main automatic evaluations in Table 1 and Table 2 are conducted under a coverage-isolated protocol. The key idea is to use high-level mechanism guidelines for refinement-time verification, while reserving fine-grained constraints for final evaluation including material, strategy, product and condition.

Guideline split. We split the guideline library according to task purpose. The refinement-time verifier only accesses high-level mechanism guidelines, which describe general CO_2RR principles such as the role of local *CO coverage in C–C coupling, competition between CO_2RR and HER, intermediate adsorption strength, and adjacent-site requirements for multi-carbon products. The final evaluator uses held-out fine-grained constraints that are more specific to material category, modulation strategy, target product, and reaction condition. For example, a high-level guideline may state that sufficient local *CO coverage is important for C–C coupling, while a held-out fine-grained constraint may evaluate whether a defect-engineered oxide-derived Cu proposal correctly links defect-rich Cu sites, *CO retention, adjacent-site coupling, and C_2H_4 selectivity.

The initial guideline pool contains 1,454 high-level mechanism guidelines and 3,046 candidate fine-grained constraints. After overlap filtering and manual review, 2,873 fine-grained constraints are retained for final evaluation as shown in Fig. 6. The removed constraints are not used in either refinement or evaluation.

Overlap filtering. We apply both lexical and semantic filtering to remove direct duplicates or near-paraphrases between the refinement and evaluation splits. For each high-level guideline and candidate fine-grained constraint pair, we compute unigram Jaccard similarity after lowercasing, lemmatization, and normalization of chemical terms. Pairs are flagged for manual inspection if either condition holds:

$$\text{Jaccard}(g_h, g_f) \geq 0.35 \quad (25)$$

We also encode each entry using the same domain encoder used in retrieval and compute nearest-neighbor cosine similarity. Pairs are flagged for semantic inspection if:

$$\cos(\mathbf{g}_h, \mathbf{g}_f) \geq 0.82. \quad (26)$$

A total of 620 candidate pairs are flagged by lexical or semantic filters and manually inspected by two annotators with CO_2RR background. A fine-grained constraint is removed if it is an exact duplicate, a near-exact paraphrase, or a direct restatement of a refinement-time guideline without additional reasoning. Constraints are retained if they share general scientific background but require additional task-specific reasoning. This is expected in scientific domains because fine-grained constraints necessarily build on common mechanistic principles.

Manual audit and deletion criteria. Among the 620 flagged pairs, 173 fine-grained constraints are removed or merged, corresponding to 5.7% of the 3,046 candidate fine-grained constraints. The main reasons for removal are near-identical mechanism statements, duplicated target-product rules, and fine-grained entries that did not add task-specific conditions beyond a high-level guideline. To estimate whether the filter misses substantial overlap, we additionally sample 200 low-overlap pairs below all thresholds for manual audit. Only 5 pairs are judged to contain weak conceptual overlap, and none is considered a direct paraphrase of a refinement-time guideline.

Residual overlap after filtering. After filtering, the retained fine-grained constraints have low lexical overlap with refinement-time guidelines. The average unigram Jaccard similarity is 0.071, with a median of 0.054 and a 95th percentile of 0.184. The average ROUGE-L overlap is 0.118, with a median of 0.103 and a 95th percentile of 0.284. For semantic similarity, most retained constraints remain below the inspection threshold. Approximately 81.6% of retained constraints have nearest-neighbor cosine similarity below 0.70, 15.7% fall

Guideline Partition	Initial Count	Removed	Final Usage
High-level mechanism guidelines	1,454	0	Refinement-time verification only
Fine-grained constraints	3,046	173	2,873 for final evaluation only

Table 10: Coverage-isolated guideline split. High-level mechanism guidelines are used for feedback generation, while retained fine-grained constraints are used only for final scoring.

Audit Category	Count	Outcome
Flagged by lexical overlap	312	Manual inspection
Flagged by semantic overlap	428	Manual inspection
Union of flagged pairs	620	Manual inspection
Removed or merged constraints	173	Excluded from final evaluation
Random low-overlap audit pairs	200	No direct paraphrase found

Table 11: Manual audit statistics for coverage isolation. Counts refer to candidate guideline pairs, except removed constraints, which refer to fine-grained entries excluded from the held-out evaluation set.

Overlap Measure	Mean	Median	95th Percentile
Unigram Jaccard	0.071	0.054	0.184
ROUGE-L	0.118	0.103	0.284
Nearest-neighbor cosine	0.612	0.604	0.781

Table 12: Residual overlap between refinement-time guidelines and held-out fine-grained evaluation constraints after filtering.

between 0.70 and 0.80, and 2.7% fall between 0.80 and 0.82. The remaining higher-similarity cases are manually checked and retained only when they contain additional fine-grained material, strategy, product, or condition information not available to the refinement-time verifier.

Real-world soft-match isolation. For the Real-World Set, soft-match score evaluation uses expert-annotated key logic constraints derived from the source paper. These constraints describe the reported design logic, including the modulation strategy, active-site effect, intermediate regulation, and product-selectivity mechanism. They are used only after generation is complete. During generation and refinement, the model receives only the design target derived from the paper introduction. It does not receive the source paper's actual catalyst design, final soft-match labels, or evaluation evidence.

Feedback masking. The feedback module is also constrained to prevent answer leakage. Feedback messages contain only the failure type and a diagnostic clue, such as [CRITICAL ERROR], [LOGIC FLAW], or [MISSING MECHANISM]. They do not include held-out fine-grained rule IDs, full evaluation constraints, or reference designs.

Potential Leakage Source	Control Mechanism
Verifier retrieves final evaluation rules	Held-out fine-grained constraints are removed from the refinement-time retrieval index.
Duplicate or paraphrased constraints	Entry-level split, lexical filtering, semantic filtering, and manual audit are applied before evaluation.
Generator sees final rule text or IDs	Final rule text, rule IDs, matched evidence, and scoring decisions are never included in generation or feedback prompts.
Feedback templates reveal target answers	Templates provide diagnostic clues only and do not expose complete constraints or reference designs.
Real-world source paper leaks into generation	Only the design target is given to the generator; the paper's actual catalyst design and expert logic constraints are hidden.
Best-of-*N* optimizes the final evaluator	Best-of-*N* selection uses refinement-time verifier scores, while final evaluation uses held-out fine-grained constraints.

Table 13: Leakage-control mechanisms in the coverage-isolated evaluation protocol.

For example, a feedback message may indicate that a claim conflicts with the need for sufficient *CO availability in C–C coupling, but it does not reveal the held-out constraint that specifies a complete defect-engineered Cu mechanism under a particular condition. Table 13 summarizes the main leakage risks and the corresponding controls.

Interpretation. The coverage-isolated protocol does not claim that refinement-time guidelines and held-out constraints are scientifically unrelated. Such independence would be unrealistic because valid CO_2RR constraints share common mechanistic background. Instead, the protocol prevents direct access to the same, near-identical, or answer-level constraints used for final scoring. Therefore, improvements under this setting indicate that high-level mechanistic feedback helps the model better satisfy more specific held-out scientific constraints,

rather than simply optimizing against the final evaluator. This provides stronger evidence that the method improves independent scientific validity in a guideline-driven evaluation setting.

F Feedback Generation Details

This section provides additional details on the failure-mode-aware feedback module. Unlike LLM-as-verifier method, it is implemented as a rule-driven feedback mapper that converts verifier outputs into template-based natural-language feedback. The goal is to provide targeted revision guidance while avoiding direct exposure of complete answer-level guidelines.

Given a generated proposal D, the verifier outputs mechanistic submetric scores and structured diagnostic evidence:

$$\mathscr{O}_v = \{S_{\text{mech}}, e, g_f, \mathscr{E}_v\}, \tag{27}$$

where S_{mech} is the overall mechanistic validity score, e is the detected failure mode, g_f is the feedback label, and $\mathscr{E}_v$ denotes structured evidence such as violated guidelines, inconsistent mechanism links, or missing causal transitions. The feedback mapper first buckets the relevant submetric scores into severity levels, then selects a failure-mode-specific template, and finally fills the template with concise evidence.

F.1 Severity Bucketing

For each feedback type, the mapper uses the corresponding mechanistic submetric as the primary signal. We use three severity levels: mild, moderate, and severe. For rule violation, higher values indicate more severe errors; for mechanism consistency and causal-chain completeness, lower values indicate more severe errors.

$$\text{sev}_{\text{rule}} = \begin{cases} \text{severe}, & v_{\text{rule}} \geq \tau^h_{\text{rule}}, \\ \text{moderate}, & \tau^l_{\text{rule}} \leq v_{\text{rule}} < \tau^h_{\text{rule}}, \\ \text{mild}, & v_{\text{rule}} < \tau^l_{\text{rule}}. \end{cases} \tag{28}$$

$$\text{sev}_{\text{cons}} = \begin{cases} \text{severe}, & v_{\text{cons}} \leq \tau^l_{\text{cons}}, \\ \text{moderate}, & \tau^l_{\text{cons}} < v_{\text{cons}} \leq \tau^h_{\text{cons}}, \\ \text{mild}, & v_{\text{cons}} > \tau^h_{\text{cons}}. \end{cases} \tag{29}$$

$$\text{sev}_{\text{chain}} = \begin{cases} \text{severe}, & v_{\text{chain}} \leq \tau^l_{\text{chain}}, \\ \text{moderate}, & \tau^l_{\text{chain}} < v_{\text{chain}} \leq \tau^h_{\text{chain}}, \\ \text{mild}, & v_{\text{chain}} > \tau^h_{\text{chain}}. \end{cases} \tag{30}$$

In our implementation, the thresholds are calibrated on a small development subset. The severity level controls the strength of the feedback message. Severe errors produce direct warnings about invalid scientific claims, while mild errors produce local revision suggestions.

F.2 Failure-Mode Mapping

The feedback mapper assigns one of three feedback labels according to the dominant diagnostic signal:

$$g_f = \begin{cases} [\text{CRITICAL ERROR}], & v_{\text{rule}} > \tau_{\text{rule}}, \\ [\text{LOGIC FLAW}], & v_{\text{cons}} < \tau_{\text{cons}}, \\ [\text{MISSING MECHANISM}], & v_{\text{chain}} < \tau_{\text{chain}}, \\ [\text{NONE}], & \text{otherwise}. \end{cases} \tag{31}$$

When multiple signals are triggered, the mapper generates a composite feedback signal while preserving the priority of the individual signals: [CRITICAL ERROR], [LOGIC FLAW], and [MISSING MECHANISM]. This priority reflects practical catalyst design requirements. The selected feedback label is then combined with structured evidence to fill a template.

F.3 Feedback Template Library

The feedback templates are designed to be clue-guided rather than answer-providing. They identify the type and approximate location of the scientific weakness, which encourages the generator to revise the proposal while preserving the open-ended nature of catalyst design.

Templates for [CRITICAL ERROR]**.** This feedback type is used when the proposal explicitly violates known mechanistic rules or domain constraints.

Critical Error Template C_1

The proposal contains a critical mechanistic conflict:
`<violated_claim>` contradicts the known constraint that `<matched_guideline>`.
Revise this claim before refining the rest of the mechanism.

Critical Error Template C_2

The proposed relation between <design_strategy> and <target_product> is inconsistent with established CO_2RR mechanism knowledge.
Reconsider the claimed intermediate adsorption or pathway effect.

Critical Error Template C_3

The explanation violates a mechanism guideline associated with <material_category> and <catalytic_property>.
Remove or correct the conflicting mechanistic claim.

Critical Error Template C_4

The proposal claims that <mechanism_effect> improves <target_product>, but the matched mechanism indicates the opposite trend under <condition>.
Revise the mechanism to satisfy this domain constraint.

Critical Error Template C_5

A critical rule violation is detected in the link from <material_structure> to <reaction_pathway>.
The revised proposal should avoid this contradiction and rebuild the mechanism around compatible intermediate behavior.

Templates for [LOGIC FLAW]**.** This feedback type is used when the proposal is internally inconsistent or contains an unsupported causal relation.

Logic Flaw Template L_1

The mechanism is internally inconsistent: <claim_A> does not support <claim_B>.
Clarify how the proposed structure or modulation strategy leads to the claimed catalytic effect.

Logic Flaw Template L_2

The proposed <design_strategy> is not well aligned with the stated <target_pathway> / <target_product>.
Revise the rationale so that the material design and product selectivity follow a coherent mechanism.

Logic Flaw Template L_3

The proposal contains an unsupported causal relation between <material_property> and <mechanistic_effect>.
Add a scientifically grounded explanation or revise the claim.

Logic Flaw Template L_4

The explanation mixes incompatible mechanism directions.
Reconcile the role of <active_site> with the claimed effect on <intermediate> and <target_product>.

Logic Flaw Template L_5

The current rationale does not consistently connect <material_structure>, <modulation_strategy>, and <catalytic_outcome>.
Revise the mechanism to make these components mutually supportive.

Templates for [MISSING MECHANISM]**.** This feedback type is used when the proposal provides an incomplete causal chain from material modification to CO_2RR performance. Unlike the other two feedback types, this feedback may include key missing intermediate mechanisms, but it still avoids prescribing a full target design.

Missing Mechanism Template M_1

The proposal is plausible but the causal chain is incomplete.
Add the missing link from <design_strategy> to <intermediate_effect> and then to <target_product>.

Missing Mechanism Template M_2

The explanation should clarify how <material_modification> changes <electronic_or_geometric_effect>, how this affects <intermediate>, and why this supports <product_selectivity>.

Missing Mechanism Template M_3

The mechanism skips an important intermediate step.
Include how the proposed catalyst regulates <adsorption_behavior> or <reaction_pathway> before claiming improved <catalytic_outcome>.

Missing Mechanism Template M_4

The rationale should include the missing transition between <active_site_property> and <key_intermediate>.
Without this link, the claimed product selectivity is not sufficiently justified.

Missing Mechanism Template M_5

The proposal needs a more complete mechanism chain:
<structural_modulation> →
<surface/electronic_effect> →
<intermediate_adsorption> →
<pathway_or_selectivity_change>.
Add the missing steps while keeping the original design goal.

G Additional Experimental Results

This section provides additional experimental results that complement the main paper. We include extended ablation studies, detailed inference analysis across refinement iterations, and evaluation on an additional benchmark. These results further examine the contribution of individual components, the behavior of iterative refinement, and the generalization of the proposed method beyond the two main test sets.

G.1 Additional Main Results

We provide the run-to-run standard deviations for both test sets and full baseline comparison results in Table 14, 15, 16 and 17.

G.2 Detailed Ablation Results

In addition to the ablation results reported in Sec 5.2, we provide a more detailed comparison over the three mechanistic submetrics. We compare the full method with variants that remove factual retrieval, remove fine-grained feedback, remove multi-dimensional scoring, and replace the guideline-grounded verifier with an LLM-as-a-judge baseline. All variants use the same backbone generator and evaluation protocol.

G.3 Detailed Inference Analysis

We further analyze how each evaluation dimension changes with the number of refinement iterations. The initial factual-only output is treated as iteration 0. Each subsequent iteration applies guideline-grounded verification and failure-mode-aware feedback before generating the next proposal. In Sec. 5.3, we report the overall trend of mechanistic validity. Here, we provide the full metric trajectory, including rule violation, mechanism consistency, causal-chain completeness, and synthesis feasibility.

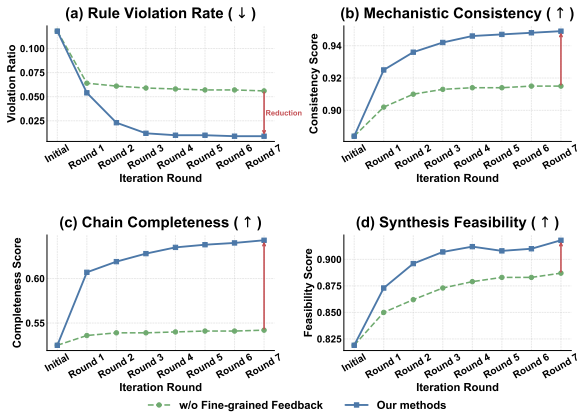


Figure 10: The full metric trajectory across refinement iterations.

Fig. 10 shows that the largest improvement occurs in the early refinement rounds. Rule violation decreases sharply after the first iteration, indicating that explicit mechanistic conflicts can often be corrected once they are identified by the verifier. Mechanism consistency also improves quickly, while causal-chain completeness increases more gradually because it requires the model to add missing intermediate reasoning links. Synthesis feasibility improves moderately during refinement, although it is not the primary feedback target. After three iterations, all metrics become relatively stable, suggesting that a small iteration budget is sufficient in most cases.

G.4 Evaluation across modulation strategies

Fig. 11 analyzes mechanistic validity across material modulation strategies on both test sets. Compared with RAG-only generation, our method consistently reduces rule violation and improves mechanism consistency and causal-chain completeness across most subgroups. This trend holds for dif-

Model	Chemical-structure Validity Std.				Rule Violation Std.				Mechanism Consistency Std.				Causal-chain Completeness Std.				Synthesis Feasibility Std.			
	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours
Qwen3-8B	0.021	0.015	0.010	0.009	0.028	0.022	0.015	0.012	0.024	0.017	0.013	0.011	0.019	0.016	0.014	0.012	0.027	0.020	0.016	0.013
DeepSeek-v3.1	0.018	0.012	0.009	0.008	0.024	0.018	0.013	0.010	0.020	0.015	0.011	0.009	0.017	0.014	0.012	0.010	0.023	0.017	0.014	0.011
GPT-4o	0.019	0.013	0.010	0.009	0.026	0.020	0.014	0.011	0.021	0.016	0.012	0.010	0.018	0.015	0.013	0.011	0.024	0.018	0.015	0.012
Gemini-2.5-pro	0.018	0.012	0.009	0.008	0.025	0.018	0.013	0.010	0.019	0.014	0.011	0.009	0.017	0.014	0.012	0.010	0.022	0.016	0.013	0.011
ChemLLM-7B	0.017	0.012	0.009	0.008	0.022	0.016	0.012	0.009	0.021	0.016	0.012	0.010	0.018	0.015	0.013	0.011	0.026	0.019	0.015	0.012
CatalystLLM-7B	0.016	0.011	0.008	0.007	0.020	0.015	0.010	0.007	0.022	0.015	0.011	0.009	0.019	0.014	0.012	0.010	0.027	0.019	0.015	0.011

Table 14: Standard deviations of the main results on the Synthetic Set. Each metric is reported under four generation settings: generate-only, RAG-only, Best-of-N, and our guideline-grounded refinement method. The standard deviations are computed over repeated runs using the same evaluation protocol as Table 1. Lower values indicate more stable performance across runs.

Model	Soft Match Std.				Rule Violation Std.				Mechanism Consistency Std.				Causal-chain Completeness Std.				Synthesis Feasibility Std.			
	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours
Qwen3-8B	0.031	0.025	0.020	0.017	0.030	0.024	0.017	0.014	0.026	0.020	0.016	0.013	0.023	0.019	0.016	0.014	0.029	0.022	0.018	0.015
DeepSeek-v3.1	0.027	0.021	0.017	0.014	0.026	0.020	0.015	0.012	0.023	0.018	0.014	0.011	0.021	0.017	0.015	0.013	0.026	0.020	0.016	0.013
GPT-4o	0.028	0.022	0.018	0.015	0.028	0.022	0.016	0.013	0.024	0.019	0.015	0.012	0.022	0.018	0.015	0.013	0.027	0.021	0.017	0.014
Gemini-2.5-pro	0.027	0.021	0.017	0.014	0.027	0.021	0.015	0.012	0.023	0.018	0.014	0.011	0.022	0.018	0.015	0.013	0.026	0.020	0.016	0.013
ChemLLM-7B	0.029	0.023	0.019	0.016	0.024	0.018	0.014	0.011	0.024	0.019	0.015	0.012	0.022	0.018	0.015	0.013	0.028	0.021	0.017	0.014
CatalystLLM-7B	0.028	0.022	0.018	0.015	0.022	0.017	0.012	0.009	0.025	0.019	0.015	0.012	0.023	0.018	0.015	0.013	0.029	0.021	0.017	0.014

Table 15: Standard deviations of the main results on the Real-World Set. Each metric is reported under four generation settings: generate-only, RAG-only, Best-of-N, and our guideline-grounded refinement method. The standard deviations are computed over repeated runs using the same evaluation protocol as Table 2. The relatively small standard deviations indicate that the observed improvements are stable across repeated evaluations.

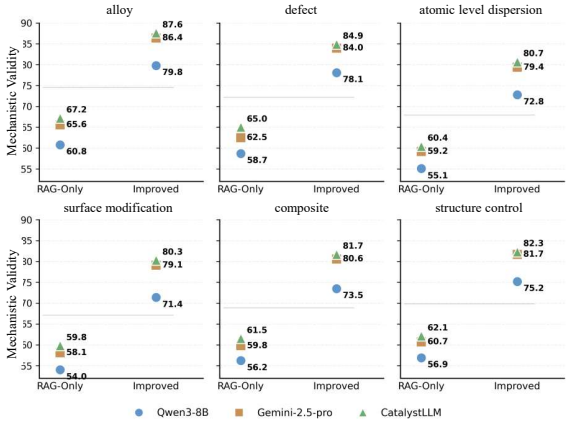


Figure 11: Mechanistic validity across material modulation strategies. The plot compares the overall mechanistic scores of three backbone models in their "RAG-Only" state with the "Improved" state (using our method).

ferent catalyst families and different modulation strategies, suggesting that the refinement module does not only benefit a narrow subset of CO_2RR tasks.

G.5 Evaluation on Additional Benchmark

To demonstrate that the capabilities of our method extend beyond a single evaluation framework, we conduct additional experiments on CatalystBench (Chen et al., 2026), a multi-task benchmark tailored for catalysis science. We focus specifically on the catalyst design question & answer task to demonstrate that our method truly enables LLMs to understand the relevant knowledge and mechanisms of catalyst design.

Table 19 reports the extended evaluation results on CatalystBench. Compared with general-purpose models, CatalystLLM already shows strong performance on the catalyst design question-answering task. Our method further improves CatalystLLM across all evaluation metrics. The consistent improvement under both semantic similarity and LLM-based scoring indicates that the proposed method does not merely optimize for a specific evaluator. Instead, guideline-grounded refinement helps the model produce answers that are more semantically aligned with expected catalyst design knowledge and more strongly preferred by independent LLM judges.

H Expert Agreement Study

To validate whether the automatic verifier reflects expert scientific judgment, we conduct an expert agreement study. The study evaluates both score-level agreement and ranking-level agreement between the verifier and human experts.

H.1 Expert Background and Annotation Protocol

We invite five domain experts with research experience in CO_2RR electrocatalysis to annotate a subset of catalyst design proposals. The annotation

Model	Chemical-structure Validity ↑				Rule Violation ↓				Mechanism Consistency ↑				Causal-chain Completeness ↑				Synthesis Feasibility ↑			
	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours
Open-Source Models																				
Qwen3-8B	0.649	0.785	0.906	0.930	0.464	0.412	0.214	0.163	0.593	0.857	0.871	0.902	0.101	0.401	0.462	0.542	0.652	0.756	0.804	0.853
Llama2-7B	0.621	0.762	0.891	0.921	0.512	0.444	0.245	0.176	0.548	0.822	0.832	0.886	0.083	0.379	0.431	0.519	0.431	0.759	0.762	0.822
Mistral-7B	0.637	0.773	0.892	0.924	0.485	0.429	0.228	0.158	0.568	0.839	0.845	0.895	0.093	0.386	0.446	0.526	0.508	0.815	0.832	0.849
DeepSeek-v3.1	0.703	0.899	0.929	0.948	0.431	0.279	0.181	0.098	0.636	0.866	0.912	0.938	0.110	0.451	0.517	0.579	0.722	0.792	0.834	0.869
Closed-Source Models																				
GPT-4o	0.708	0.902	0.934	0.940	0.466	0.401	0.202	0.152	0.602	0.864	0.881	0.922	0.108	0.468	0.528	0.582	0.746	0.827	0.865	0.898
Gemini-2.5-pro	0.712	0.936	0.940	0.942	0.422	0.258	0.167	0.116	**0.658**	0.890	0.926	**0.954**	0.117	0.479	0.539	0.591	**0.764**	0.839	0.882	**0.912**
Claude-3	0.709	0.910	0.937	0.943	0.410	0.248	0.160	0.108	0.648	0.887	0.924	0.946	0.114	0.472	0.534	0.587	**0.758**	0.836	0.878	0.909
Domain-Specialized Models																				
ChemLLM-7B	0.724	0.923	0.949	0.960	0.398	0.142	0.102	0.054	0.542	0.848	0.865	0.904	0.100	0.412	0.473	0.551	0.492	0.706	0.758	0.795
CatalystLLM-7B	**0.727**	0.921	0.962	**0.969**	**0.371**	0.129	0.064	**0.014**	0.562	0.885	0.892	0.951	**0.129**	0.531	0.547	**0.639**	0.498	0.819	0.861	0.907

Table 16: Full results on the Synthetic Set. Each metric is reported under four generation settings: generate-only, RAG-only, Best-of-*N*, and our guideline-grounded refinement method. Rule Violation is reported as a violation rate, where lower is better. All other metrics are normalized scores, where higher is better.

Model	Soft Match ↑				Rule Violation ↓				Mechanism Consistency ↑				Causal-chain Completeness ↑				Synthesis Feasibility ↑			
	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours	Gen.	RAG	B-o-N	Ours
Open-Source Models																				
Qwen3-8B	0.06	0.40	0.52	0.66	0.499	0.436	0.241	0.175	0.514	0.813	0.846	0.887	0.088	0.372	0.436	0.515	0.618	0.712	0.775	0.828
Llama2-7B	0.04	0.33	0.45	0.59	0.531	0.468	0.276	0.198	0.469	0.775	0.808	0.852	0.070	0.338	0.396	0.472	0.402	0.683	0.731	0.787
Mistral-7B	0.05	0.36	0.48	0.61	0.507	0.452	0.257	0.181	0.486	0.789	0.821	0.866	0.078	0.351	0.409	0.487	0.452	0.731	0.778	0.823
DeepSeek-v3.1	0.11	0.57	0.69	0.82	0.451	0.298	0.196	0.112	0.553	0.836	0.881	0.914	0.097	0.426	0.490	0.566	0.728	0.804	0.852	0.894
Closed-Source Models																				
GPT-4o	0.13	0.64	0.75	0.85	0.484	0.319	0.216	0.165	0.543	0.822	0.862	0.892	0.092	0.327	0.453	0.525	0.712	0.801	0.851	0.892
Gemini-2.5-pro	**0.15**	0.68	0.78	**0.88**	0.441	0.279	0.181	0.132	**0.582**	0.861	0.901	**0.928**	0.106	0.452	0.513	0.586	**0.735**	0.813	0.861	**0.901**
Claude-3	0.14	0.66	0.77	0.86	0.436	0.274	0.174	0.124	0.575	0.857	0.896	0.924	0.104	0.448	0.507	0.579	0.732	0.809	0.857	0.898
Domain-Specialized Models																				
ChemLLM-7B	0.10	0.55	0.66	0.81	0.417	0.163	0.121	0.076	0.462	0.802	0.838	0.889	0.101	0.413	0.474	0.561	0.445	0.758	0.812	0.861
CatalystLLM-7B	0.14	0.62	0.73	0.87	**0.388**	0.126	0.090	**0.042**	0.490	0.845	0.885	0.927	0.114	0.502	0.539	**0.646**	0.483	0.812	0.849	0.889

Table 17: Full results on the Real-World Set. Each metric is reported under four generation settings: generate-only, RAG-only, Best-of-*N*, and our guideline-grounded refinement method.

set contains 180 samples in total. We first sample 150 generated proposals from outputs produced by different methods, including generate-only, factual-only, Best-of-*N*, and our guideline-grounded refinement method. To make the evaluation cover more boundary cases, experts additionally construct 15 innovative but scientifically plausible proposals and 15 clearly invalid proposals with obvious mechanistic errors. This sampling strategy ensures that the evaluated proposals span a wide quality range.

Each annotation instance contains the input task and the generated catalyst design proposal. Experts are asked to evaluate the proposal independently, without access to the automatic verifier score. For each proposal, experts assign scores along four dimensions: rule violation, mechanism consistency, causal-chain completeness, and overall scientific reasonableness. In addition, for a subset of 50 tasks, experts are asked to rank multiple candidate proposals generated for the same input task.

To reduce annotation bias, the proposals are anonymized and randomly shuffled before being assigned to experts. When multiple experts annotate the same instance, we use the average expert score as the reference label. For ranking evaluation, majority preference is used when experts disagree. Cases with substantial disagreement are further inspected qualitatively to identify ambiguous mechanisms. Each proposal is annotated by at least two experts, and a randomly selected subset of 90 proposals is annotated by all five experts to estimate inter-annotator agreement. Inter-annotator agreement is measured using **ICC score**, yielding 0.82 for overall mechanistic validity and 0.87, 0.84, 0.73 for rule violation, mechanism consistency, and causal-chain completeness.

H.2 Scoring Rubric

Experts evaluate each proposal using a normalized scoring rubric. The scores are assigned on a 5-point scale and then normalized to the range [0, 1] when calculating correlation scores.

Rule violation. Experts assess whether the proposal explicitly contradicts known CO_2RR mechanisms. A high violation score indicates that the proposal contains severe mechanistic contradictions.

Mechanism consistency. Experts evaluate whether the proposal is internally coherent. A

Variant	Rule Violation ↓	Mechanism Consistency ↑	Causal-chain Completeness ↑	Overall Mech. Score ↑
Full Method	**0.073**	**0.921**	**0.612**	**0.882**
w/o Factual Retrieval	0.108	0.897	0.581	0.842
w/o Fine-grained Feedback	0.124	0.884	0.553	0.831
w/o Multi-dimensional Scoring	0.166	0.842	0.497	0.782
LLM-as-a-judge	0.151	0.856	0.512	0.794
Factual-only	0.305	0.643	0.451	0.541
Generate-only	0.371	0.562	0.129	0.464

Table 18: Detailed ablation results on mechanistic validity.

Model	STS	LLM Scores	
	Scores	gpt-4o	deepseek-r1
Closed-source Models			
claude-3	0.73	0.74	0.73
Gemini-2.5-pro	0.74	0.75	0.76
gpt-4o	0.73	0.71	0.72
Open-source Models			
deepseek-v3.1	0.73	0.84	0.86
LLaMA2-7B	0.63	0.57	0.52
Qwen3-8B	0.62	0.70	0.72
Mistral-7B	0.66	0.68	0.42
ChemLLM	0.68	0.64	0.68
CatalystLLM	0.79	0.82	0.86
CatalystLLM-our method	**0.82**	**0.85**	**0.88**

Table 19: Extended evaluation on the CatalystBench in semantic understanding tasks. The best model is in bold font, and the second-best is underlined.

consistent proposal should align the material structure, modulation strategy, target product, and claimed catalytic effect.

Causal-chain completeness. Experts assess whether the proposal provides sufficient intermediate reasoning from material modification to catalytic outcome. Proposals that skip key intermediate mechanisms receive lower scores.

Overall mechanistic validity. Experts provide an overall judgment of whether the proposal is scientifically plausible and useful as a catalyst design candidate. This score considers chemical plausibility, mechanistic validity, and the clarity of the design rationale.

H.3 Ranking Evaluation Setup

In addition to score-level annotation, we conduct a ranking-based evaluation to test whether the verifier can identify the best proposal among multiple candidates for the same task. For each selected input task, we collect candidate proposals generated by different methods. Experts are asked to rank the candidates according to overall scientific reasonableness, with particular attention to mechanistic validity.

We compare the verifier ranking with the expert ranking using two metrics. The first is **Top-1 accuracy**, which measures whether the proposal ranked highest by the verifier is also selected as the best proposal by experts. The second is **pairwise accuracy**, which measures whether the verifier agrees with expert preferences for pairs of candidate proposals. Formally, for a pair of proposals (D_i, D_j), the verifier is considered correct if its relative ranking is consistent with the expert preference.

The verifier achieves a Top-1 accuracy of 0.86 and a pairwise accuracy of 0.82. These results indicate that the verifier can reliably identify higher-quality catalyst design proposals among multiple candidates. The ranking agreement is especially important for our method because the verifier is also used for selecting and guiding refinement directions.

H.4 Correlation Analysis

We further analyze score-level agreement between the automatic verifier and expert ratings. For each annotated proposal, we compute verifier scores for rule violation, mechanism consistency, causal-chain completeness, and overall mechanistic validity. These scores are then compared with the corresponding expert annotations using Pearson correlation.

As shown in Fig. 12, the verifier shows strong alignment with expert judgment across all evaluated dimensions. The **Pearson correlation** is $r = 0.84$ for rule violation, $r = 0.89$ for mechanism consistency, $r = 0.80$ for causal-chain completeness, and $r = 0.86$ for overall mechanistic validity. The strongest agreement appears in mechanism consistency, indicating that the verifier can reliably judge whether the material structure, modulation strategy, and catalytic rationale form a coherent explanation.

Causal-chain completeness shows a slightly lower correlation, which is expected because open-ended catalyst design may admit multiple valid

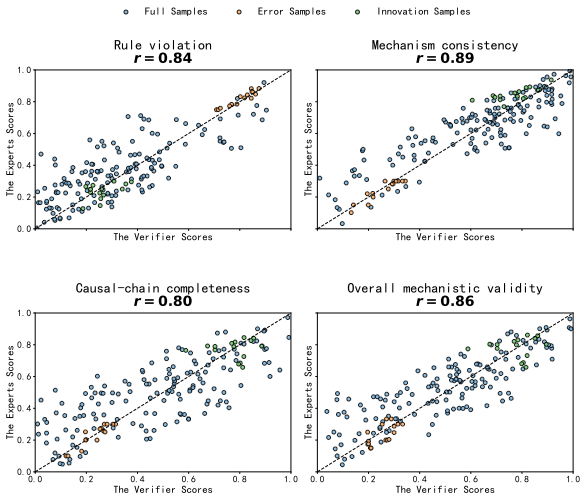


Figure 12: Correlation between verifier scores and expert scores across four mechanistic evaluation dimensions. Blue points denote the full annotated sample set, orange points denote expert-constructed error samples, and green points denote expert-constructed innovation samples. The dashed diagonal line indicates perfect agreement.

mechanistic routes and experts may accept different levels of mechanistic detail. The expert-constructed error samples concentrate in the low-quality regions for consistency, completeness, and overall validity, while the innovation samples are generally assigned higher scores when they remain scientifically plausible. This suggests that the verifier is not only sensitive to obvious scientific errors, but can also recognize creative proposals when they satisfy mechanistic constraints. Overall, the correlation results support the use of the verifier as an automatic approximation of expert mechanistic judgment for large-scale evaluation and refinement.

I Case Studies

I.1 Successful Studies

This section presents representative case studies to illustrate how the proposed method identifies and corrects different types of mechanistic failures in CO_2RR catalyst design. We show examples for explicit rule violation, mechanism inconsistency, missing causal chains, and a successful catalyst design. For each case, we report the input task, the initial proposal issue, the feedback signal, and the refined proposal.

As shown in Fig. 16 and Fig. 17, we further examine whether the proposed method can support the discovery of experimentally actionable catalyst designs. We select two refined catalyst proposals that receive high mechanistic validity and synthesis feasibility scores, and synthesize the corresponding materials for experimental validation. For each case, we report the design task, the final proposal, the predicted mechanism chain, and the corresponding experimental evidence. To the best of our knowledge, these two catalyst formulations have not been previously reported for CO_2RR under comparable design settings. It demonstrates the practical usefulness of our method in producing experimentally testable catalyst hypotheses.

I.2 Failure Case Analysis

In addition to successful refinement examples, we analyze two representative failure cases to clarify the current limitations of the framework. These cases show that guideline-grounded refinement improves scientific validity but does not fully solve all aspects of open-ended catalyst design. The first case illustrates conservative refinement, where the final proposal is mechanistically valid but lacks novelty. The second case illustrates limited feedback specificity for emerging mechanisms that are not sufficiently covered by the current guideline library.

Failure Case 1: Conservative refinement. As shown in Fig. 18, The guideline-grounded refinement can improve scientific validity, but may also encourage conservative designs that stay close to well-established mechanisms. The current verifier focuses on validity rather than novelty. Future work can incorporate novelty-aware scoring, literature-distance estimation, or diversity-promoting proposal selection to better balance scientific correctness and exploratory design.

Failure Case 2: Emerging mechanism beyond guideline coverage. Fig. 19 illustrates a limitation of guideline-grounded verification: the quality of feedback depends on the coverage of the guideline library. For emerging catalyst strategies, especially molecular interphase engineering, ionomer effects, and dynamic surface reconstruction, the verifier may detect incompleteness but cannot always provide fine-grained mechanistic guidance. Future work can expand the guideline library with newly reported mechanisms and integrate external tools such as spectroscopy-aware evidence retrieval or molecular dynamics simulations.

Failure mode frequency. We further estimate the frequency of these failure patterns through man-

Design Task (Synthetic Set): Design a Cu-based CO_2RR electrocatalyst for selective C_2H_4 production under alkaline electrolyte conditions. The catalyst should involve surface defect engineering and maintain high *CO coverage for C--C coupling.

Initial Proposal:
A defect-rich Cu surface with abundant oxygen vacancies is introduced by partial electrochemical reduction of CuO. The oxygen vacancies are used as the main active sites to strongly stabilize *COOH and promote direct CO formation. The enhanced CO formation is expected to suppress *CO surface coverage and reduce C--C coupling barriers, thereby improving C_2H_4 selectivity. The catalyst can be prepared by oxidizing Cu foam in air, followed by partial electrochemical reduction in CO_2-saturated $KHCO_3$ electrolyte.

Retrieved guidelines from Verifier:
- For Cu-based CO_2RR toward C_2 products, sufficient local *CO coverage on adjacent Cu sites is a prerequisite for C--C coupling.
- Defect-rich or oxide-derived Cu surfaces may promote C_2 formation by increasing *CO retention and enabling neighboring carbon-containing intermediates to couple.
- A design rationale that claims suppressed *CO coverage improves C_2H_4 selectivity conflicts with the known requirement of *CO-mediated C--C coupling on Cu.

$v_{rule} = 0.42$ $v_{cons} = 0.61$ $v_{chain} = 0.48$ $v_{mech} = 0.57$

Error Type: *[CRITICAL ERROR]*
Feedback: The proposal contains a mechanistic contradiction: C_2H_4 formation on Cu generally requires sufficient *CO coverage and favorable C--C coupling. Revising the mechanism should avoid claiming that suppressed *CO coverage promotes C_2H_4 selectivity.

Revised Proposal:
A defect-rich oxide-derived Cu catalyst is constructed by controlled oxidation and electrochemical reduction of Cu foam. The reduction process generates grain boundaries, undercoordinated Cu sites, and residual subsurface oxygen species. These surface defects tune the adsorption strength of *CO and help maintain a high local *CO coverage on adjacent Cu sites. The increased *CO residence and spatial proximity of carbon-containing intermediates promote C--C coupling toward *OCCO formation, thereby improving C_2H_4 selectivity under alkaline CO_2RR conditions. The catalyst can be prepared by thermal oxidation of Cu foam followed by electrochemical reduction in CO_2-saturated alkaline electrolyte.

Figure 13: Case study of correcting an explicit rule violation under CatalystLLM-7B as the backbone generator. The initial proposal designs a defect-rich oxide-derived Cu catalyst for C_2H_4 production, but incorrectly claims that suppressing local *CO coverage can promote C–C coupling. The verifier matches guidelines indicating that sufficient *CO coverage on adjacent Cu sites is required for C_2 formation and assigns a [CRITICAL ERROR] feedback signal. After refinement, the proposal revises the mechanism by linking defect-rich Cu sites to enhanced *CO retention, *CO dimerization, and improved C_2H_4 selectivity.

ual inspection of refined outputs. Specifically, We inspect all 404 refined outputs from the 304 Synthetic Set tasks and 100 Real-World Set tasks after guideline-grounded refinement and categorize the remaining weaknesses according to the dominant failure type. As shown in Table 20, most refined proposals do not contain major scientific errors, while the remaining failures are mainly associated with conservative refinement and insufficiently specific feedback for emerging mechanisms. This suggests that the framework is effective at reducing explicit mechanistic errors, but its remaining limitations are more related to novelty control and guideline coverage.

Failure Type	Count	Percentage
No major failure	307	76.0%
Conservative refinement / low novelty	47	11.6%
Emerging mechanism beyond guideline coverage	30	7.4%
Residual mechanism inconsistency	13	3.2%
Residual synthesis ambiguity	7	1.7%
Total	404	100.0%

Table 20: Manual analysis of remaining failure modes after guideline-grounded refinement.

Design Task (Synthetic Set): Design a Cu--Ag bimetallic CO_2RR catalyst for enhanced C_2+ product formation. The design should use tandem catalysis between Ag and Cu sites.

Initial Proposal:
A Cu--Ag bimetallic catalyst is prepared by depositing Ag nanoparticles on a rough Cu surface. Ag sites accelerate CO_2 conversion to *CO, while Cu sites promote C--C coupling. The Ag nanoparticles are designed to bind *CO very strongly and keep CO intermediates on Ag, which prevents CO spillover to Cu and increases ethylene production on Cu. The rough Cu surface further improves local pH and suppresses hydrogen evolution.

Retrieved guidelines from Verifier:
- In Cu--Ag tandem CO_2RR catalysts, Ag mainly promotes CO generation, while Cu provides sites for *CO adsorption and C--C coupling.
- Effective tandem catalysis requires spatial or transport coupling between CO-generating sites and C--C coupling sites, so that CO produced on Ag can become available to neighboring Cu sites.
- Excessively strong retention of *CO on non-Cu sites can weaken the tandem effect if it prevents CO transfer or spillover to Cu active sites.

$v_{rule} = 0.18$ $v_{cons} = 0.46$ $v_{chain} = 0.52$ $v_{mech} = 0.62$

Error Type: *[LOGIC FLAW]*
Feedback: The tandem mechanism is internally inconsistent. If *CO is retained too strongly on Ag and does not transfer to Cu, the proposed Ag-to-Cu tandem pathway cannot support C--C coupling on Cu. Clarify how CO generated on Ag becomes available to adjacent Cu sites.

Revised Proposal:
A Cu--Ag tandem catalyst is constructed by decorating a rough Cu surface with dispersed Ag nanoparticles. Ag sites preferentially convert CO_2 to CO with high efficiency, while adjacent Cu sites provide active ensembles for *CO adsorption and C--C coupling. The Ag loading is controlled to generate CO near Cu without completely covering the Cu surface. The local CO produced on Ag can diffuse or spill over to neighboring Cu sites, increasing *CO availability on Cu and promoting *CO dimerization toward C_2+ products. The rough Cu surface further increases the density of low-coordination sites and strengthens local CO retention, while the Ag component helps suppress competing hydrogen evolution.

Figure 14: Case study of correcting mechanism inconsistency under CatalystLLM-7B as the backbone generator. The initial Cu–Ag tandem catalyst proposal states that Ag sites retain *CO strongly and prevent CO spillover to Cu, while also claiming that Cu sites promote C–C coupling. The verifier identifies this as a logical inconsistency because an effective tandem mechanism requires CO generated on Ag to become available to neighboring Cu sites. The [LOGIC FLAW] feedback guides the revised proposal to clarify Ag-to-Cu CO transfer and the cooperative role of Ag and Cu sites in C_{2+} formation.

Design Task (Real-World Set): Design a nitrogen-doped carbon-supported Cu single-atom or subnanocluster catalyst for selective CO production from CO_2RR under neutral electrolyte conditions.

Initial Proposal:
Cu species are anchored on nitrogen-doped carbon to form isolated Cu active sites. Nitrogen doping improves the electronic structure of the support and enhances CO_2RR activity. The catalyst is expected to show high CO selectivity because Cu sites can activate CO_2 efficiently. The material can be synthesized by mixing a Cu precursor with a nitrogen-rich carbon support, followed by pyrolysis under inert atmosphere.

Retrieved guidelines from Verifier:
- For atomically dispersed Cu sites on N-doped carbon, Cu-N_x coordination can stabilize isolated Cu centers and suppress aggregation during synthesis or operation.
- CO production requires sufficient stabilization of the *COOH intermediate, followed by moderate *CO binding that allows CO desorption instead of further reduction.
- A complete mechanism for single-atom or subnanocluster CO_2RR catalysts should connect coordination environment, electronic structure modulation, intermediate binding, and product selectivity.

$v_{rule} = 0.09$ $v_{cons} = 0.74$ $v_{chain} = 0.31$ $v_{mech} = 0.70$

Error Type: ***[MISSING MECHANISM]***
Feedback: The proposal is generally plausible but the causal chain is incomplete. It should explain how N coordination stabilizes Cu sites, how this affects CO_2 activation and *COOH formation, and why *CO desorption is favored for CO selectivity.

Revised Proposal:
Cu atoms are anchored on nitrogen-doped carbon through Cu--Nx coordination to form atomically dispersed Cu sites. The N-rich coordination environment stabilizes isolated Cu centers and prevents aggregation during pyrolysis. Electron donation from the N-doped carbon support tunes the Cu electronic structure, facilitating CO_2 adsorption and activation. The Cu--Nx sites stabilize the *COOH intermediate sufficiently to promote CO formation, while maintaining moderate *CO binding so that CO can desorb instead of undergoing further reduction. This balance between *COOH stabilization and *CO release improves CO selectivity under neutral CO_2RR conditions. The catalyst can be prepared by impregnating a nitrogen-rich carbon precursor with a Cu salt, followed by controlled pyrolysis and acid washing to remove unstable Cu particles.

Figure 15: Case study of completing a missing causal chain under CatalystLLM-7B as the backbone generator. The initial proposal for Cu sites on nitrogen-doped carbon is broadly plausible but only states that N doping improves the electronic structure, without explaining the intermediate mechanism. The verifier detects incomplete reasoning and assigns a [MISSING MECHANISM] feedback signal. The revised proposal completes the causal chain by connecting Cu–N_x coordination, isolated Cu stabilization, CO_2 activation, *COOH formation, moderate *CO binding, and selective CO desorption.

latex/Figures/appendix/successful case_1.pdf

Figure 16: Successful catalyst design case study on Synthetic Set. The design task asks for a Cu-based CO_2RR catalyst for selective ethanol production under alkaline flow-cell conditions. The final proposal constructs a sulfur-terminated molecularly modified CuO_x catalyst, where the organic modifier forms a Cu–S-related metal-organic interphase on the CuO_x surface. The mechanism chain links molecular modification, interfacial Cu coordination, local reaction-environment regulation, oxygenated C_2 intermediate stabilization, and enhanced ethanol selectivity. The experimental characterization panel summarizes corresponding evidence, including metal-organic interphase formation, Cu coordination modulation, and improved ethanol or C_{2+} production on modified CuO_x catalysts.

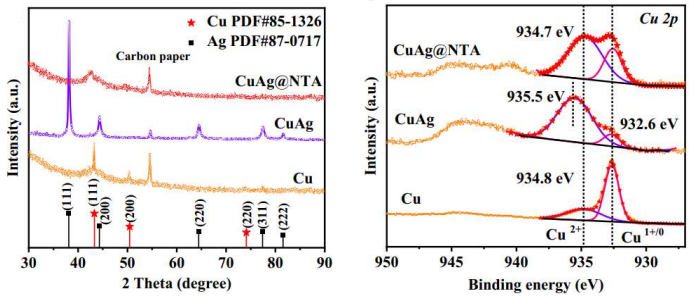
Design Task (Real-World Set): Design a Cu-based CO_2RR catalyst for highly selective ethanol production at industrial current density. The design should stabilize C_2 oxygenated intermediates, promote the hydrogenation pathway toward ethanol, and maintain catalyst stability during long-term electrolysis.

Final Proposal:
A CuAg bimetallic catalyst is coordinated and stabilized by nitrilotriacetic acid (NTA) to form a CuAg@NTA interfacial structure. The CuAg catalytic centers provide adjacent metallic sites for *CO adsorption and C--C coupling, while the NTA-derived coordination environment anchors the surface Cu and Ag species through Cu--N, Cu--O, or related metal--ligand interactions. This coordination layer suppresses uncontrolled reconstruction or dissolution-redeposition of the CuAg surface during high-current CO$_2$RR. Mechanistically, the CuAg sites promote the formation of C_2 intermediates, and the NTA-stabilized interface favors the hydrogenation of $*CH_2CHO$ toward $*CH_3CH_2O$ rather than C--O bond cleavage toward ethylene. The resulting catalyst is expected to improve ethanol selectivity while maintaining stability under industrial current density. A feasible preparation route is to electrodeposit Cu and Ag precursors on carbon paper in the presence of NTA, followed by electrochemical activation under CO_2RR conditions to form the stabilized CuAg@NTA catalyst.

$v_{rule} = 0.05$ $v_{cons} = 0.94$ $v_{chain} = 0.79$ $v_{mech} = 0.91$

Mechanism Chain:

CuAg + NTA
↓
CuAg@NTA coordination interface
↓
stabilized CuAg catalytic centers
↓
Lower C-C coupling and $*CH_2CHO$ hydrogenation barriers
↓
enhanced ethanol selectivity



The catalyst achieves ~87% faradaic efficiency for C_2H_5OH products.

Figure 17: Successful catalyst design case study on Real-World Set. The design task asks for a Cu-based CO_2RR catalyst for selective ethanol production at industrial current density. The final proposal constructs an NTA-stabilized CuAg bimetallic interface, where CuAg sites promote C–C coupling and the NTA coordination environment stabilizes catalytic centers while favoring hydrogenation of oxygenated C_2 intermediates toward ethanol. The mechanism chain connects molecular coordination, active-site stabilization, C_2 intermediate regulation, ethanol selectivity, and long-term operational stability. The experimental panel summarizes corresponding evidence from CuAg@NTA, including high ethanol Faradaic efficiency, high ethanol partial current density, and stable operation over 300 h.

Design Task (Real-World Set): Design a Cu-based CO_2RR catalyst for selective C_2H_4 production under alkaline conditions. The design should use interface engineering to improve C--C coupling and suppress hydrogen evolution.

Initial Proposal:
A Cu/Cu_2O interface is constructed by partially reducing Cu_2O nanocubes. The Cu^0 domains provide active sites for *CO adsorption and C--C coupling, while residual Cu^+ sites stabilize oxygenated intermediates. The interface improves C_2H_4 selectivity by increasing *CO coverage and reducing competing HER. The catalyst can be prepared by growing Cu_2O nanocubes and applying mild electrochemical reduction under CO_2RR conditions.

$v_{rule} = 0.04$ $v_{cons} = 0.88$ $v_{chain} = 0.77$ $v_{mech} = 0.86$

Observed Limitation:
The proposal is **scientifically valid** but **overly conventional**. It relies on a common oxide-derived Cu/Cu_2O design pattern and does not introduce a clearly differentiated structural feature or modulation strategy. The verifier assigns a high score because the proposal satisfies known mechanistic constraints, but the evaluation framework does not explicitly penalize low novelty.

Figure 18: A representative failure case of our method. It shows conservative refinement: the final proposal satisfies known CO_2RR mechanistic constraints but remains close to conventional oxide-derived Cu/Cu_2O designs.

Design Task (Synthetic Set): Design a molecularly modified Cu catalyst for selective ethanol production from CO_2RR under high-current-density flow-cell conditions. The design should regulate the local interfacial microenvironment and stabilize oxygenated C_2 intermediates.

Initial Proposal:
A CuO_x-derived catalyst is modified with a zwitterionic organic molecule containing both sulfonate and amine groups. The sulfonate group anchors near the Cu surface through electrostatic interaction, while the amine group enriches local CO_2 and modulates interfacial water. The organic layer stabilizes oxygenated C_2 intermediates through hydrogen bonding and promotes ethanol formation. The catalyst can be prepared by coating the organic modifier on oxide-derived Cu followed by electrochemical activation in alkaline electrolyte.

$v_{rule} = 0.07$ $v_{cons} = 0.74$ $v_{chain} = 0.49$ $v_{mech} = 0.72$

Error Type: *[MISSING MECHANISM]*
Feedback: The proposal should clarify how the molecular modifier remains stable under alkaline CO_2RR conditions and how the local microenvironment specifically changes adsorption or protonation of oxygenated C_2 intermediates.

Observed Limitation:
The feedback correctly identifies that the mechanism chain is incomplete, but it remains **generic**. Because molecular interphase effects and interfacial water regulation are only partially covered in the current guideline library, the verifier cannot provide highly specific diagnostic evidence for this emerging design direction.

Figure 19: A representative failure case of our method. It shows limited feedback specificity for an emerging molecular-interphase strategy, where the verifier detects an incomplete causal chain but cannot provide highly specific guidance due to limited guideline coverage.