ORGANICBENCH: CAN LARGE LANGUAGE MODELS REASON LIKE CHEMISTS?

Anonymous authors
Paper under double-blind review

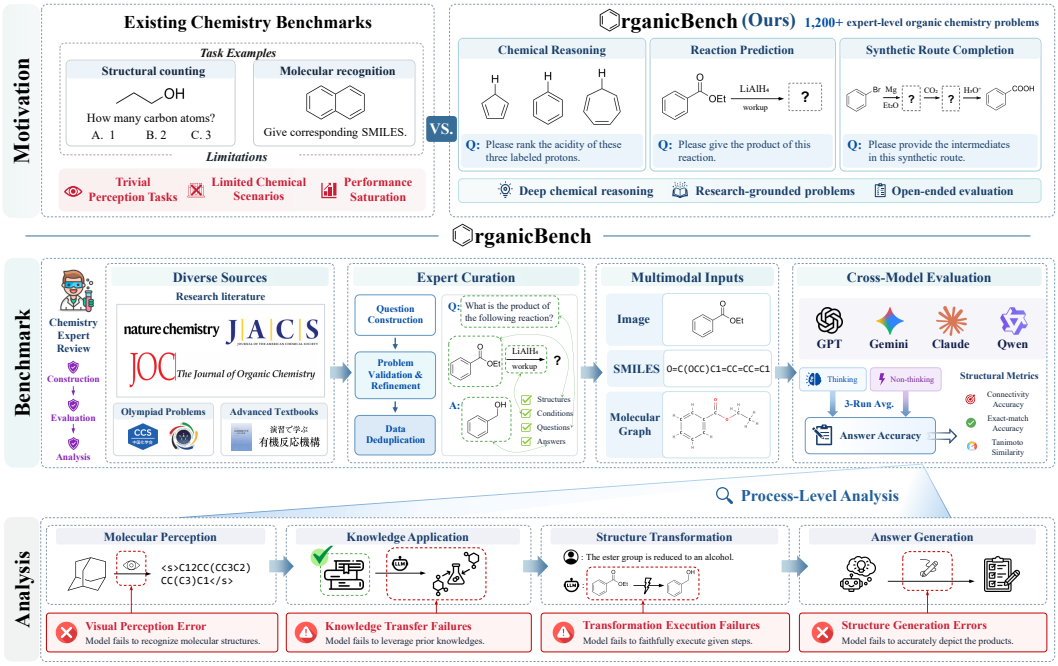


Figure 1: **Overview of OrganicBench.** Motivation, construction, and process-level evaluation.

ABSTRACT

Large language models have surpassed human experts in coding and mathematical theorem proving. However, they still struggle to tackle natural science problems in the real world, thereby limiting their potential for autonomous scientific discovery. We find that organic chemistry is a rich and challenging testbed grounded in real-world chemistry for evaluating AI chemists. To this end, we introduce OrganicBench, a multimodal benchmark comprising over 1,200 challenging organic chemistry problems grounded in published research literature. Given problem descriptions in images, SMILES strings, or molecular graphs, models are tasked with predicting reaction products, inferring key intermediates, and reasoning about reaction mechanisms. These tasks require models to understand realistic reaction conditions, apply correct knowledge, and perform complex reasoning in multistep synthetic routes, going beyond factual question answering based on memorization. Our evaluation demonstrates that frontier proprietary and open-source models fail to achieve competitive performance on these problems, and they exhibit consistent failure modes. Process-level analyses further reveal failures across multiple stages of problem-solving, exposing the gap between broad chemical knowledge and reliable complex reasoning. Overall, OrganicBench serves as an expert-level benchmark for measuring progress toward capable AI chemists.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Large language models have surpassed human experts on specialized tasks such as theorem proving (Hubert et al., 2026), code generation (OpenAI, 2025), and workflow execution (Hume, 2026), and have begun to exhibit capabilities for autonomous research (Boiko et al., 2023). In the natural sciences, LLMs have mastered extensive factual knowledge (Mirza et al., 2025; Huang et al., 2026). However, current models still struggle to achieve automated scientific discovery in these fundamental disciplines. Achieving this requires models to synthesize knowledge and reason logically under the guidance of scientific intuition. We find that organic chemistry is a particularly suitable domain for studying this capability, as broad domain knowledge and deep reasoning are closely intertwined in this field. Advanced organic chemistry problems require reasoning from basic principles about complex reactions, including reactive sites, plausible mechanisms, and synthetic transformations (Clayden et al., 2012). Moreover, decades of organic chemistry research have produced a large collection of experimentally validated reactions and synthetic routes (Schwaller et al., 2019; Li et al., 2026), providing abundant realistic and verifiable problems for evaluation. These characteristics make organic chemistry a natural testbed for assessing expert-level scientific reasoning grounded in domain knowledge, the fundamental capability for scientific research, and thereby for evaluating the potential of LLMs to become genuine AI chemists.

Existing LLM benchmarks for chemistry, such as ChemBench (Mirza et al., 2025), ChemEval (Huang et al., 2026), and ChemIQ (Runcie et al., 2026), primarily evaluate models' understanding of chemical knowledge and their basic structural operations (*e.g.,* atom counting and structure drawing). However, they exhibit several limitations: **(1) Emphasis on shallow perception over deep reasoning.** Many tasks rely heavily on molecular recognition, structural counting, and property prediction (Guo et al., 2023; Runcie et al., 2026). While useful for testing basic visual and structural understanding, they are insufficient for assessing expert-level scientific deduction. **(2) Disconnect from authentic research complexity.** Derived largely from conventional datasets or classical named reactions, and frequently formulated as multiple-choice or fill-in-the-blank questions, these benchmarks fail to mirror the complexity of real-world scientific discovery. **(3) Performance saturation.** State-of-the-art models already achieve high scores across many of these task categories, making current benchmarks increasingly inadequate for evaluating the frontier LLMs.

To address these limitations, we introduce **OrganicBench**, an expert-level benchmark of over 1,200 problems manually curated from frontier research, Olympiads, and advanced textbooks, as shown in Figure 1. Provided across three modalities (*i.e.,* images, SMILES (Weininger, 1988), and molecular graphs), tasks such as reaction prediction and synthetic route completion demand expert-level contextual reasoning rather than shallow recognition. Evaluations reveal significant capability gaps: even the strongest setting, Gemini 3.1 Pro Preview with SMILES input, achieves only 43.77% accuracy, while chemistry-specialized LLMs largely fail. Through expert-guided error analysis, we identify two prevalent failure modes: **(i)** the inability to recognize correct reaction patterns from substrates and conditions, and **(ii)** the introduction of structural errors unrelated to the actual reaction site. Further experiments disentangling the underlying reasoning process expose distinct weaknesses in molecular interpretation, knowledge application, and structure generation, highlighting the critical capabilities that need to be improved for LLMs to progress toward AI chemists.

Our main contributions are summarized as follows:

- **A Benchmark for Open-ended Chemical Reasoning.** OrganicBench is an LLM benchmark containing more than 1,200 problems grounded in real chemical reactions, including over 1,000 open-ended questions and more than 450 reaction problems derived from frontier chemistry literature such as *Nature Chemistry* and *JACS*. Each problem is provided in three input modalities: images, SMILES strings, and molecular graphs.
- **Fine-grained Evaluation of Frontier LLMs.** We conduct comprehensive evaluations of frontier and chemistry-specialized LLMs. For representative models, we further perform analytical experiments that systematically reveal the gaps between them and expert-level chemical intelligence.
- **Analysis of Scientific Intelligence of Advanced Models.** Through comprehensive benchmarking and analytical studies, we find that current models struggle with complex chemical reasoning, and that this shortcoming stems from capability deficiencies at multiple stages of the reasoning process. These findings identify key directions for advancing LLMs toward capable AI Chemists.

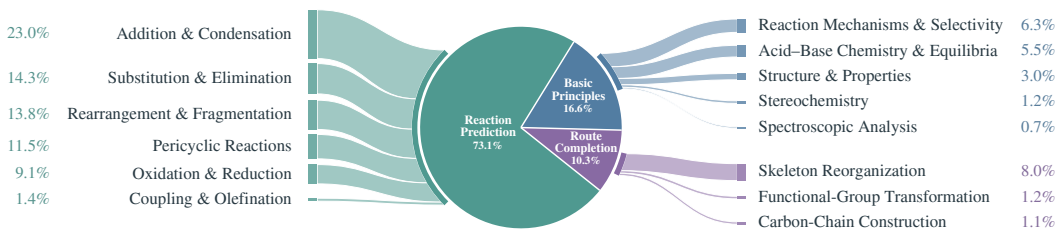


Figure 2: **Task distribution in OrganicBench.** Proportions of task categories and subtypes.

2 Related Work

LLMs for Chemistry. General LLMs have been adapted for chemistry, progressing from factual knowledge to complex reasoning. Models like ChemLLM (Zhang et al., 2024), ChemDFM (Zhao et al., 2025b), and LlaSMol (Yu et al., 2024) enhance domain knowledge via instruction tuning, while ChemDFM-X (Zhao et al., 2024) and ChemMLLM (Tan et al., 2026) integrate visual representations to bridge text and molecular structures. To tackle complex problem-solving, ChemDFM-R (Zhao et al., 2025c) and Chem-R (Wang et al., 2026) leverage expert reasoning trajectories and domain adaptation. However, existing models still lack generalizable chemical understanding. OrganicBench is designed to systematically characterize this limitation.

LLM-based Chemistry Agents and AI Chemists. LLM agents augment models with tools and planning for chemical research. ChemCrow (M. Bran et al., 2024) introduced basic tool-use, while Coscientist (Boiko et al., 2023), LLM-RDF (Ruan et al., 2024), and ChemAgents (Song et al., 2025) enabled autonomous, feedback-driven research loops via multi-agent coordination. For real-world deployment, systems like ACRA (Pagel et al., 2026), RoboChemist (Zhang et al., 2025), and ChemVLMate (Song et al., 2026) integrate multimodal feedback to execute physical experiments. Despite these workflow and automation advances, critical research decisions ranging from hypothesis generation to plan adaptation still rely on the underlying LLM's chemical reasoning capability. A robust pipeline is insufficient without deep domain understanding. OrganicBench specifically evaluates this foundational capability required for more capable AI chemists.

Benchmarks for LLMs in Chemistry. Reliable benchmarks are essential for guiding progress in LLM capabilities. General suites like ChemLLMBench (Guo et al., 2023), ChemBench (Mirza et al., 2025), and ChemEval (Huang et al., 2026) evaluate broad factual knowledge, while targeted datasets (*e.g.,* ChemIQ (Runcie et al., 2026), ChemCoTBench (Li et al., 2025), QCBench (Xie et al., 2025), ChemReason-Bench (Zhang et al., 2026)) assess specific structural, editing, or numerical skills. However, they largely focus on basic perception and isolated skills rather than the deep chemical intuition required of an AI chemist. Recent benchmarks such as SUPERChem (Zhao et al., 2025a), RxnBench (Li et al., 2026), and FukuyamaBench (Dang et al., 2026) extend evaluation toward more complex chemical reasoning and scientific literature understanding, but provide limited analysis of failures in open-ended generation. OrganicBench fills this gap with a multimodal, open-ended benchmark curated from real-world reactions, enabling systematic analysis of model failures in complex chemical reasoning.

3 OrganicBench

OrganicBench includes three task types: **Basic Chemical Principles**, **Reaction Prediction**, and **Synthetic Route Completion**. Each problem is also presented in three modalities: molecular structure images, SMILES, and molecular graphs.

3.1 Task Description

Basic Chemical Principles (Basic Principles). Models are required to compare molecular properties or identify plausible reaction mechanisms by applying chemical principles, including electronic and steric effects.

Table 1: Mean accuracy (%) over three runs, higher is better. Inputs: Image, SMILES, and Molecular Graph.

Model	Basic Principles			Reaction Prediction			Route Completion			Overall		
	Image	SMILES	Molecular Graph	Image	SMILES	Molecular Graph	Image	SMILES	Molecular Graph	Image	SMILES	Molecular Graph
Proprietary Models												
Claude Opus 5	70.65	**73.80**	71.14	33.46	36.70	33.63	**32.80**	**38.44**	**33.87**	39.58	43.05	39.90
Gemini 3.1 Pro Preview	**72.47**	**73.80**	**74.96**	**35.46**	**38.70**	**34.65**	25.00	31.18	23.12	**40.54**	**43.77**	**40.17**
GPT-5.6 Sol	71.14	67.50	71.64	29.84	30.89	32.16	27.42	32.53	29.03	36.46	37.15	38.41
Qwen3.8-Max	70.32	—	—	25.08	—	—	17.74	—	—	31.84	—	—
Open-Source Models												
Qwen3.8-27B	58.71	62.19	63.68	10.75	16.70	16.14	6.99	15.59	10.75	18.33	24.15	23.49
-w/o thinking	40.13	37.48	37.81	2.26	1.17	0.68	0.00	0.00	0.00	8.33	7.09	6.78
Qwen3.8-Flash-Next	**65.84**	**66.83**	**66.50**	**18.29**	**22.21**	**21.68**	**12.63**	**23.66**	**19.62**	**25.61**	**29.78**	**28.92**
-w/o thinking	43.78	43.28	45.27	4.52	4.60	1.32	0.00	0.81	0.00	10.59	10.64	8.49
Qwen3.5-27B	60.70	60.03	62.19	11.12	10.37	9.54	5.91	6.99	4.30	18.83	18.28	17.76
-w/o thinking	41.96	41.79	36.82	1.58	1.13	0.45	0.27	0.81	0.00	8.16	7.86	6.45
Qwen3.5-122B-A10B	55.56	60.03	59.54	13.42	12.10	10.33	8.87	6.18	4.30	19.96	19.47	17.89
-w/o thinking	43.95	45.11	44.28	2.83	2.53	0.23	0.27	0.00	0.00	9.40	9.35	7.53
Gemma 4 31B IT	57.55	55.06	55.22	8.30	8.67	7.09	2.42	2.15	0.81	15.88	15.72	14.45
-w/o thinking	43.12	42.45	32.67	2.04	1.43	0.53	0.00	0.00	0.00	8.66	8.11	5.82
Chemistry-Specialized Models												
ChemDFM-v2.0-14B	—	**20.56**	**24.71**	—	**1.17**	**0.15**	—	0.00	0.00	—	**4.27**	**4.22**
ChemDFM-R-14B	—	19.24	22.72	—	0.45	0.11	—	0.00	0.00	—	3.53	3.86
Chem-R-8B	—	11.11	8.96	—	0.00	0.00	—	0.00	0.00	—	1.85	1.49

Reaction Prediction. Models are required to infer missing reactants, products, or intermediates from the available reaction context by reasoning accurately about how molecular structures transform throughout the reaction.

Synthetic Route Completion (Route Completion). Models are required to recover missing structures in partially specified synthetic routes through forward and retrosynthetic reasoning, while maintaining consistency across successive steps.

3.2 Benchmark Construction

OrganicBench is constructed from existing problems and research literature through systematic collection, annotation, question construction, and quality control.

3.2.1 Data Collection and Filtering.

To cover tasks ranging from basic chemical principles to complex synthetic reasoning, OrganicBench draws on both existing problems and research literature. Existing problems are collected from the *Chinese Chemistry Olympiad*, practice problems of comparable difficulty, and *advanced organic chemistry textbooks*. We further collect 200 papers on reaction methodology and total synthesis from journals including *Nature Chemistry* and *JACS*. All materials are filtered based on three criteria: (1) molecular structures and reaction relationships can be represented unambiguously; (2) sufficient information is provided to construct a well-defined question; and (3) the problem or reaction is chemically valid or supported by a verifiable source.

3.2.2 Material Annotation and Question Construction.

To standardize materials from different sources, we process existing problems and research papers with human expert annotation. For existing problems, we annotate the problem information and corresponding answers. For research papers, we annotate the reported reactions and synthetic routes

suitable for question construction. The annotated text and molecular structures are then extracted into structured records. Based on these records, we construct questions according to the task definitions in Section 3.1, reformulating existing problems when necessary while preserving the information required for solving them. Additional information is included when necessary to make the problem well defined. Annotation and question construction methods are detailed in Appendix B.2; interfaces, personnel, and operational procedures are described in Appendix G.

3.2.3 Quality Control.

To ensure question and ground-truth correctness, all annotated problems underwent two rounds of expert review. First, each problem was checked against its source for structural and chemical correctness and clarity. Second, Gemini 3.1 Pro Preview was used to solve the curated questions, and experts reviewed discrepancies against the original sources and chemical principles, adding valid alternative answers and revising problematic questions when necessary. Structure-parsing and consistency checks were further applied to ensure valid and aligned textual and structural records.

3.2.4 Data Deduplication and Statistics.

After quality control, potential duplicates were identified by matching molecular structures and manually checking the corresponding questions and answers. Problems assessing the same chemical content were removed. The resulting OrganicBench contains 1,209 problems, with 38% from research papers, 33% from chemistry competitions, and 29% from textbooks. By task type, as shown in Figure 2, the benchmark includes 17% problems on basic principles, 73% on reaction prediction, and 10% on synthetic route completion.

3.3 Input Modalities

To decouple reasoning from modality biases, OrganicBench presents each problem in three modalities: images, SMILES, and molecular graphs (see Appendix C for details). By altering only the representations while keeping the text identical, it enables controlled cross-modal comparisons.

Image. Compounds and reaction schemes are rendered as conventional images, requiring models to extract structural information directly from standard chemical imagery.

SMILES. Structures are encoded as isomeric SMILES strings within a JSON-formatted reaction layout, bypassing visual perception to evaluate pure text-based chemical reasoning.

Molecular Graph. Molecules are provided as JSON-formatted graphs specifying atomic connectivity, allowing us to assess whether explicit topological representations facilitate reasoning.

4 Experimental Setup

We evaluate a diverse set of frontier and chemistry-specialized LLMs on OrganicBench under a unified evaluation protocol.

4.1 Models

General-Purpose Models. We evaluate frontier proprietary models, including Claude Opus 5 (Anthropic, 2026), Gemini 3.1 Pro Preview (Google, 2026), GPT-5.6 Sol (OpenAI, 2026), and Qwen3.8-Max (Qwen Team, 2026c). For open-source models, we systematically investigate the impacts of model scale and explicit chain-of-thought. We compare Gemma 4 31B (Gemma Team, 2026) with several Qwen variants, ranging from Qwen3.5-27B and Qwen3.5-122B-A10B (Qwen Team, 2026b) to Qwen3.8-27B (Qwen Team, 2026c) and Qwen3.8-Flash-Next (Qwen Team, 2026a). Where supported, models are tested with both reasoning and non-reasoning configurations.

Chemistry-Specialized Models. To examine the benefits of domain specialization, we evaluate ChemDFM-v2.0-14B (Zhao et al., 2025b) alongside its reasoning-enhanced variant, ChemDFM-R-14B (Zhao et al., 2025c). We further include Chem-R-8B (Wang et al., 2026) to verify if reasoning benefits generalize across architectures.

Input Settings. Models are evaluated on their supported modalities (see Appendix D.1 for details).

4.2 Evaluation and Metrics

Evaluation Protocol. All models are evaluated using the same English problem statements and response instructions without external retrieval or chemistry tools. Each problem is evaluated independently for three runs with a generation budget of up to 100,000 tokens per run (model-specific limits in Appendix D.1). We report mean accuracy and run-to-run standard deviation (Appendix D.3), with detailed prompts and scoring procedures in Appendix D.2.

Metrics. Our primary metric is problem-level accuracy, where a problem is counted as correct only if all required answers are correct (see Appendix D.2 for details). For molecular structure predictions, we report connectivity accuracy, which ignores stereochemistry, and exact-match accuracy, which requires full agreement with the reference structure. We report Morgan-fingerprint Tanimoto similarity to measure partial structural agreement between predicted and reference molecules, with structure-level results provided in Appendix D.4.

5 Results and Discussion

We analyze overall accuracy, and conduct error analyses and controlled experiments to uncover how task types and chemical modalities expose critical weaknesses in models' reasoning processes.

5.1 Overall Performance

Table 1 reports mean accuracy across three runs for each input modality; the corresponding run-to-run standard deviations are provided in Appendix D.3.

Performance of general-purpose LLMs. In the full multimodal evaluation, even state-of-the-art LLMs fall well short of the problem-solving performance they have demonstrated in other domains. At their best-performing settings, Gemini 3.1 Pro Preview achieves an overall accuracy of 43.77%, followed by Claude Opus 5 at 43.05% and GPT-5.6 Sol at 38.41%. Within the Qwen family, the larger Qwen3.8-Max outperforms Qwen3.8-27B by 7.69%, suggesting a clear benefit from model scaling. Overall, these results show that current LLMs still lack broad and deep chemical understanding, with substantial headroom remaining even for the strongest models.

Performance across task types. Performance drops markedly from Basic Principles to Reaction Prediction and Route Completion. For example, Gemini 3.1 Pro Preview decreases from 72.47% on Basic Principles to 35.46% and 25.00% on the latter two tasks under image input, with other models showing a similar trend. This gap suggests that performance on basic chemical principles does not fully translate to more complex reaction and synthesis tasks.

Effect of reasoning mode. Explicit reasoning substantially affects the performance of all open-source models, with the average accuracy across five models decreasing from 20.57% to 8.22% when reasoning is disabled. The magnitude of this effect varies substantially across tasks: performance decreases by 31.9% on Basic Principles, 86.1% on Reaction Prediction, and 98.4% on Route Completion. This progressively larger degradation suggests that reasoning plays a greater role in tasks involving more complex chemical transformations.

Performance of chemistry-specialized models. Chemistry-specialized models perform considerably worse than general-purpose models. Even the strongest, ChemDFM-v2.0-14B, achieves only about 4% overall accuracy, and none solves any route-completion problems, highlighting their difficulty with complex chemical reasoning. Moreover, the reasoning-enhanced ChemDFM-R-14B does not outperform ChemDFM-v2.0-14B, suggesting that current domain-specific training and reasoning tuning do not reliably transfer to expert-level reaction and synthesis reasoning.

Since closed-form basic-principle questions primarily reflect knowledge gaps rather than reasoning deficiencies, our subsequent analyses focus on the open-ended reaction prediction and route completion tasks to evaluate reasoning limitations.

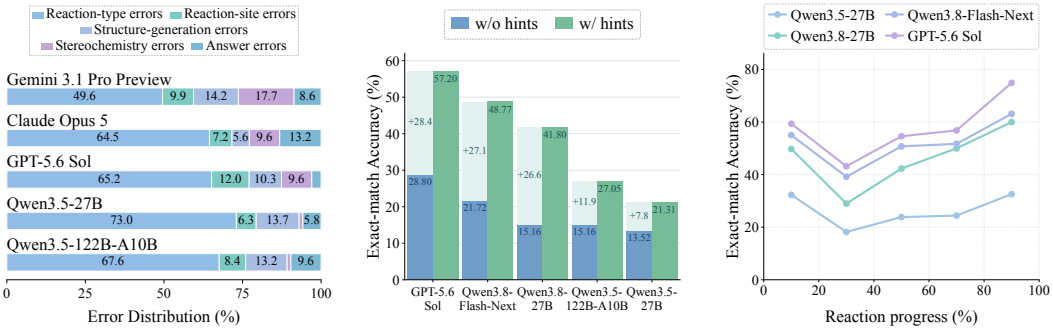


Figure 3: **Error patterns and diagnostic experiments. Left:** Error-category distributions across models. **Middle:** Reaction prediction accuracy with and without knowledge hints. **Right:** Single-step accuracy across reaction progress.

5.2 Error Analysis

To characterize model failures on open-ended questions, we conduct a fine-grained error attribution analysis. We group the errors into five categories, defined below, with detailed definitions and attribution procedures provided in Appendix E.

• **Reaction-type errors.** Constituting the majority of failures, these occur when models predict fundamentally incorrect transformations, revealing an inability to determine the appropriate reaction for given substrates and conditions.

• **Reaction-site errors.** Models often identify the correct reaction but apply it to the wrong reactive center, exposing severe deficiencies in reasoning about chemo-, regio-, and site selectivity.

• **Structure-generation errors.** Even with a correct reaction scheme, models frequently alter non-reacting regions (*e.g.,* spuriously adding or deleting carbon atoms) or produce invalid SMILES, indicating difficulty in faithfully translating the intended transformation into a valid molecular structure.

• **Stereochemistry errors.** These involve incorrect or omitted configurations on otherwise correct molecular structures. Their proportion increases in frontier models because these models are more likely to recover the correct connectivity, leaving stereochemistry as the remaining source of error.

• **Answer errors.** Models struggle to identify the complete set of major reaction outcomes, often overlooking competing pathways or predicting chemically implausible byproducts.

From the results of representative models summarized in Figure 3 Left, reaction-type errors dominate model failures, indicating that identifying the correct chemical transformation remains a primary bottleneck. As this limitation is reduced in stronger models, other errors become increasingly prominent, suggesting that models may fail at multiple stages of the problem-solving process.

5.3 Process-Level Analysis Experiments

Error categories alone do not reveal where these failures originate. To precisely pinpoint these sources, we design controlled experiments across four decomposed stages: structural perception, knowledge understanding and application, step execution, and structure generation.

5.3.1 Molecular Perception

Molecular representation can affect model performance, as images require visual interpretation while SMILES and molecular graphs provide structural information symbolically. We examine this effect in Section 5.1 by comparing performance across the three input modalities.

Effect of input representation. Models vary significantly in their ability to leverage explicit symbolic inputs. Gemini 3.1 Pro Preview, Claude Opus 5, and the Qwen3.8 models consistently benefit from structural representation. For instance, shifting from images to SMILES improves Gemini's accuracy from 40.54% to 43.77% and Claude's from 39.58% to 43.05%; Qwen3.8-Flash-Next

Table 2: Diagnostic results for **(a)** visual structure recognition, **(b)** reaction knowledge at subquestion and question levels, and **(c)** single-step structure transformation with explicit reaction descriptions. Accuracy in %; Tanimoto similarity in [0,1]. Dashes denote unavailable results.

	(a) Visual Recognition		(b) Reaction Knowledge		(c) Structure Transformation		
Model	Connectivity Acc.	Exact-match Acc.	Subquestion Acc.	Question-level Acc.	Connectivity Acc.	Exact-match Acc.	Tanimoto Similarity
GPT-5.6 Sol	**94.39**	**90.98**	**93.30**	**83.20**	**67.70**	**59.42**	**0.828**
Qwen3.8-Flash-Next	79.76	73.35	88.53	71.60	60.66	53.11	0.765
Qwen3.8-27B	69.74	62.73	82.35	56.40	54.35	47.72	0.694
Qwen3.5-122B-A10B	75.35	65.73	81.44	56.00	—	—	—
Qwen3.5-27B	70.74	62.32	80.54	53.20	30.33	27.23	0.442

improves from 25.61% to 29.78%. Conversely, GPT-5.6 Sol and the Qwen3.5 models exhibit no consistent benefit, often underperforming their image-based baselines when given symbolic inputs. Ultimately, while explicit atomic connectivity can eliminate visual parsing errors, its utility strictly depends on the model's capacity to parse symbolic syntax and integrate it into chemical reasoning.

Evaluation of optical chemical structure recognition (OCSR). To quantify visual extraction loss, we task models with converting $\sim$500 sampled structure images to SMILES, with results in Table 2(a) (see Appendix F.1 for details). OCSR accuracies vary significantly: GPT-5.6 Sol achieves 90.98%, while others range from 62% to 74%, with stereochemistry consistently remaining a significant error source. These findings align with our cross-modality evaluation: visually stronger models gain less from symbolic inputs, while weaker ones benefit more. However, Qwen3.8-27B and Qwen3.5-27B show different gains despite similar OCSR accuracy, indicating that bypassing visual bottlenecks alone is insufficient; models must also possess the capacity to interpret and reason over symbolic chemical syntax.

To avoid confounding visual recognition errors, all subsequent experiments use SMILES to preserve structural information and better isolate the capabilities under study.

5.3.2 Chemical Knowledge and Application

To decouple chemical knowledge recall from complex reasoning, we used Gemini 3.1 Pro Preview to generate 776 multiple-choice questions from 250 diverse problems. These specifically probe foundational principles like reagent functions and selectivity (see Appendix F.2 for details).

Chemical knowledge. In Table 2(b), we show that all models demonstrate strong domain knowledge, exceeding 80% accuracy. Even under a stricter question-level criterion (*i.e.,* requiring all questions associated with a reaction to be answered correctly), every model still achieves over 50% accuracy. This broad grasp of reaction principles and selectivity suggests that errors in full reaction prediction cannot be attributed solely to knowledge deficits.

The gap between knowing and applying knowledge. Despite strong performance on chemical knowledge questions, problem accuracy remains substantially lower, ranging from 10.00% to 31.20%. For example, Qwen3.8-Flash-Next achieves 71.60% reaction-level knowledge accuracy but only 22.40% product accuracy. More importantly, among 179 reactions for which all knowledge subquestions are answered correctly, 129 products are still predicted incorrectly. These results highlight a gap between knowledge recall and active reasoning.

Can explicit knowledge improve reasoning? To distinguish knowledge retrieval from knowledge application, we prompt models with key reaction principles upfront, keeping intermediates and final products hidden. As shown in Figure 3 Middle, explicit knowledge boosts accuracy by up to 28.4% across all models (see Appendix F.3 for details). This finding highlights that recognizing relevant knowledge and correctly applying it are distinct capabilities.

5.3.3 Structure Transformation

The previous experiment shows that even with correct reaction knowledge, models may still predict incorrect products. One possible source of error is failure to faithfully apply the intended transfor-

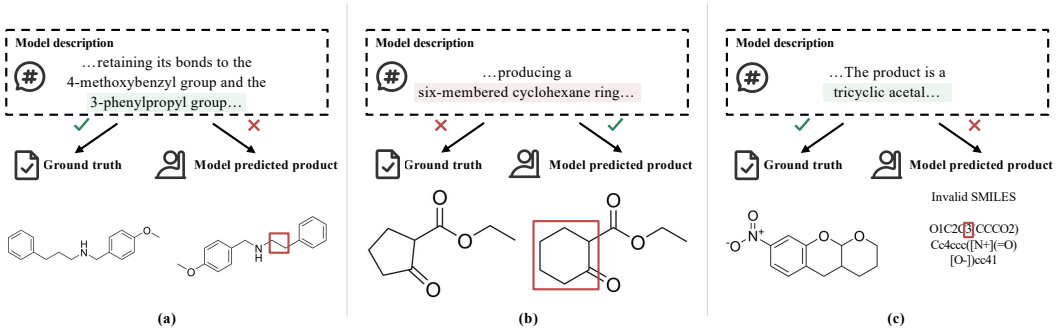


Figure 4: **Answer-generation errors.** (a) Unintended structural changes. (b) Incorrect ring size. (c) Invalid SMILES.

mation to a specific molecular structure. To examine this, we decompose reaction pathways into independent single-step tasks, where models are given a specific substrate and the corresponding transformation and asked to predict the immediate product (see Appendix F.4 for details).

Even under this simplified setting, models remain limited. As shown in Table 2(c), the gap between connectivity and exact-match accuracy indicates that models often capture major structural changes but miss stereochemical or local details. More importantly, even with the reaction direction explicitly specified, the strongest model, GPT-5.6 Sol, fails on over 40% of the transformations, showing that faithfully executing a known transformation on a specific structure remains a major bottleneck.

Furthermore, performance varies with the position of the step in the pathway, as shown in Figure 3 Right. Models perform substantially better on intermediate-to-product transformations than on intermediate-to-intermediate ones (*e.g.,* GPT-5.6 Sol drops from 76.12% to 53.55%). Since both are single-step tasks, this gap likely reflects differences in transformation complexity. Intermediate steps often require coordinated changes across multiple sites, making faithful structural execution more challenging, whereas final steps are typically more localized.

5.3.4 Answer Generation

In the error analysis in Section 5.2, we find that models often alter unreacted structures despite identifying the correct reaction sites, conflating reasoning errors with structure-generation errors. To disentangle them, we prompt models to first describe the intended structural modifications before generating the final answer (see Appendix F.5 for details). Among 46 candidate structure-generation errors from Qwen3.8-Flash-Next, 41.3% yield invalid SMILES. Of the remaining valid but incorrect structures, 40.7% have correct descriptions, while 59.3% already contain reasoning errors.

These results identify structure generation as an additional source of error beyond reaction reasoning, as illustrated in Figure 4. Even with correct modification descriptions, models may introduce unintended changes to unreacted components or molecular connectivity. Invalid SMILES further reveal failures in structural syntax and valence constraints, although some can be recovered with minor formatting corrections.

6 Conclusion

We introduce OrganicBench, a multimodal benchmark for evaluating complex reasoning in challenging organic chemistry problems drawn from research literature, chemistry Olympiads, and advanced textbooks. Our results show that current frontier LLMs possess substantial chemical knowledge but remain limited in applying it to complex problems, faithfully executing transformations, and generating accurate molecular structures. These findings highlight reliable structure-grounded reasoning as a key capability needed for progress toward capable AI chemists. We hope OrganicBench provides a rich and rigorous testbed for measuring this progress.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the organization and formatting of our datasets, refining our data analysis, and generating synthetic test data for our chemical knowledge probes (*i.e.,* Gemini 3.1 Pro Preview). We have not used generative AI tools for other tasks with required disclosure, such as developing theoretical models or conceptual frameworks, proposing or refining hypotheses, designing methodology or experiments, implementing methods, assisting with translation, and interpreting results, and the rest of the required disclosure tasks (*e.g.,* formulating mathematical claims, providing ingredients for proofs, and assisting in the writing of proofs) are not applicable to this work. Additionally, we used generative AI tools for developing automated annotation software, polishing the manuscript's phrasing to improve readability, and supplementing potentially missed references. We have reviewed all AI-assisted work. Specifically, the synthetic test data generated by Gemini 3.1 Pro Preview was strictly verified by human experts prior to evaluation. The AI-generated code for our annotation software was thoroughly tested and debugged by the authors to ensure correctness. All AI-assisted data formatting and analysis processes were manually validated to guarantee data integrity. Furthermore, every AI-suggested reference was cross-checked for actual existence and relevance in academic databases, and the polished text was carefully reviewed to confirm that it accurately preserves our original scientific intent without introducing hallucinations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

OrganicBench evaluates chemical reasoning using problems curated from published research articles, chemistry Olympiads, and advanced textbooks. Expert annotators and reviewers verify the resulting records, and all time spent on training, annotation, review, and corrections is compensated, as described in Appendix G.2. Source-derived text and figures remain subject to the rights of their respective owners. Their redistribution requires attention to applicable licenses and permissions. The benchmark is intended for computational evaluation. Performance on OrganicBench does not establish the safety or reliability of a model for autonomous chemical experimentation. Model-generated structures and reaction predictions require expert verification before laboratory use. We also recognize that advances in chemical reasoning may have dual-use implications, including potential misuse involving hazardous substances.

REPRODUCIBILITY STATEMENT

We document the benchmark construction and evaluation procedures to support reproducibility. In Appendix B, we describe data collection, question construction, and chemical validation, while we detail the annotation interfaces and review workflow in Appendix G. The three input representations and an aligned example are provided in Appendix C. In Appendix D, we report model configurations, generation settings, complete evaluation prompts, answer-processing rules, scoring metrics, and standard deviations across three independent runs. In Appendices E and F, we describe the error-attribution protocol and the prompts and procedures for the diagnostic experiments. Exact reproduction of proprietary-model results may depend on the continued availability and behavior of the corresponding API models. We will release the benchmark data and evaluation code upon acceptance, subject to applicable restrictions on third-party source materials.

REFERENCES

Anthropic. Introducing Claude Opus 5, July 2026. URL https://www.anthropic.com/news/claude-opus-5.

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

Jonathan Clayden, Nick Greeves, and Stuart Warren. *Organic chemistry*. Oxford university press, 2012.

Xingyu Dang, Haocheng Tang, Junmei Wang, and Yanjun Li. Learning mechanistic reasoning for chemical reactions with large language models. *arXiv preprint arXiv:2607.12771*, 2026.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Google. Gemini 3.1 Pro Preview, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview. Model documentation. Accessed September 26, 2026.

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

Yuqing Huang, Rongyang Zhang, Xuesong He, Xuyang Zhi, Hao Wang, Nuo Chen, Zongbo Liu, Xin Li, Feiyang Xu, Deguang Liu, et al. Chemeval: A multi-level and fine-grained chemical capability evaluation for large language models. In *International Conference on Learning Representations*, volume 2026, pp. 21365–21392, 2026.

Thomas Hubert, Rishi Mehta, Laurent Sartran, Miklós Z Horváth, Goran Žužić, Eric Wieser, Aja Huang, Julian Schrittwieser, Yannick Schroecker, Hussain Masoom, et al. Olympiad-level formal mathematical reasoning with reinforcement learning. *Nature*, 651(8106):607–613, 2026.

Tristan Hume. Designing AI-resistant technical evaluations. Anthropic Engineering, January 2026. URL https://www.anthropic.com/engineering/AI-resistant-technical-evaluations.

Hanzheng Li, Xi Fang, Yixuan Li, Chaozheng Huang, Junjie Wang, Xi Wang, Hongzhe Bai, Bojun Hao, Shenyu Lin, Huiqi Liang, et al. Rxnbench: A multimodal benchmark for evaluating large language models on chemical reaction understanding from the scientific literature. *Journal of Chemical Information and Modeling*, 66(10):5747–5756, 2026.

Hao Li, He Cao, Bin Feng, Daniel Shao, Robert Tang, Zhiyuan Yan, Yonghong Tian, Li Yuan, and Yu Li. Beyond chemical QA: Evaluating LLM's chemical reasoning with modular chemical operations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://dblp.org/rec/conf/nips/LiCFSTYTYL25.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17(7):1027–1034, 2025.

OpenAI. Competitive programming with large reasoning models. *arXiv preprint arXiv:2502.06807*, 2025. URL https://arxiv.org/abs/2502.06807.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026. URL https://openai.com/index/gpt-5-6/.

Sebastian Pagel, Michael Jirasek, and Leroy Cronin. Verification and execution of the scientific literature via chemputation augmented by large language models. *Communications Chemistry*, 2026.

Qwen Team. Qwen3.8-Flash-Next: A new architecture, towards ultimate cost-efficiency, August 2026a. URL https://qwen.ai/blog?id=qwen3.8-flash-next.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026b. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026c. URL https://qwen.ai/blog?id=qwen3.8.

Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, et al. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature communications*, 15(1):10160, 2024.

Nicholas T Runcie, Charlotte M Deane, and Fergus Imrie. Assessing the chemical intelligence of large language models. *Journal of chemical information and modeling*, 66(1):216–227, 2026.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572, 2019.

Tao Song, Man Luo, Xiaolong Zhang, Linjiang Chen, Yan Huang, Jiaqi Cao, Qing Zhu, Daobin Liu, Baicheng Zhang, Gang Zou, et al. A multiagent-driven robotic ai chemist enabling autonomous chemical research on demand. *Journal of the American Chemical Society*, 147(15):12534–12545, 2025.

Tao Song, Weixing Chen, Xiaobo Li, Hengtao Lei, Liang Zhang, Hongli Zhang, Gang Zou, Linjiang Chen, Weiwei Shang, and Jun Jiang. Chemvlmate: a multimodal, dual-arm, seeing robotic chemist for adaptive synthesis. 2026.

Qian Tan, Di Zhang, Ben Gao, Peng Xia, Wanhao Liu, Shufei Zhang, Wanli Ouyang, Lei Bai, Yuqiang Li, and Tianfan Fu. Developing chemmllm as a multimodal large language model for chemistry. *Cell Reports Physical Science*, 7(5), 2026.

Weida Wang, Benteng Chen, Di Zhang, Wanhao Liu, Ben Gao, Shuchen Pu, Shuzhou Sun, Jin Zeng, Tianshu Yu, Wanli Ouyang, et al. Chem-r: Learning to reason as a chemist. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 12291–12302, 2026.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Jiaqing Xie, Weida Wang, Ben Gao, Zhuo Yang, Haiyuan Wan, Shufei Zhang, Tianfan Fu, and Yuqiang Li. Qcbench: Evaluating large language models on domain-specific quantitative chemistry. *Journal of Chemical Information and Modeling*, 65(22):12268–12278, 2025.

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

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

Jinwei Zhang, Xucheng Liang, Yu Zhang, Ruijie Yu, Xiaokang Yang, Yaohui Jin, and Yanyan Xu. Chemreason-bench: Benchmarking large language models for procedural reasoning in experimental chemistry. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33211–33248, 2026.

Zongzheng Zhang, Chenghao Yue, Haobo Xu, Minwen Liao, Xianglin Qi, Huan-ang Gao, Ziwei Wang, and Hao Zhao. Robochemist: Long-horizon and safety-compliant robotic chemical experimentation. *arXiv preprint arXiv:2509.08820*, 2025.

Zehua Zhao, Zhixian Huang, Junren Li, Siyu Lin, Junting Zhou, Fengqi Cao, Kun Zhou, Rui Ge, Tingting Long, Yuexiang Zhu, et al. Superchem: A multimodal reasoning benchmark in chemistry. *arXiv preprint arXiv:2512.01274*, 2025a.

Zihan Zhao, Bo Chen, Jingpiao Li, Lu Chen, Liyang Wen, Pengyu Wang, Zichen Zhu, Danyang Zhang, Yansi Li, Zhongyang Dai, et al. Chemdfm-x: towards large multimodal model for chemistry. *Science China Information Sciences*, 67(12), 2024.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, et al. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4), 2025b.

Zihan Zhao, Ziping Wan, Lu Chen, Xuanze Lin, Shiyang Yu, Situo Zhang, Da Ma, Zichen Zhu, Danyang Zhang, Huayang Wang, et al. Chemdfm-r: A chemical reasoning llm enhanced with atomized chemical knowledge. *arXiv preprint arXiv:2507.21990*, 2025c.

A Task Definitions

In Figure 5, we use one aligned question–figure–answer format to illustrate the three task types.

A.1 Basic Chemical Principles

These problems test whether a model can identify and apply a localized organic chemistry principle. Outputs are usually a choice, an ordering, or a short chemical conclusion. The required reasoning may involve electronic and steric effects, aromaticity, conformational preference, acidity/basicity, stereoelectronics, orbital symmetry, or mechanistic feasibility. No credit is given for selecting a chemically related option when the requested comparison or ordering is wrong.

A.2 Reaction Prediction

These problems provide substrates and reaction conditions and ask for one or more missing products, reactants, or intermediates. A correct response must identify the operative transformation and reactive site, preserve unaffected connectivity, and reproduce every reference-specified charge and stereochemical feature. When several structures are requested, all must be present and assigned to the correct labels.

A.3 Synthetic Route Completion

These problems hide one or more structures in a multistep synthesis. The missing state is constrained by both the transformations that precede it and those that follow it, so the task requires forward and retrosynthetic reasoning. Molecular formulae, compound labels, or textual hints are retained when they are needed to make the inference unique.

B Benchmark Construction Details

B.1 Data Collection and Filtering

Existing problems are retained only when the chemically relevant context is legible, the requested output is unambiguous, and the answer can be verified against the source. For research articles, we retain reactions and routes only when the reported substrates, conditions, and outcomes provide enough evidence to define one of the three benchmark tasks. Items that rely on unavailable context, admit unresolved answers, or cannot be represented without changing their chemistry are excluded. The processed source crop is retained when the original drawing is essential or no verified machine-readable representation is available.

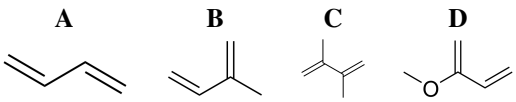
B.2 Material Annotation Methods and Question Construction

Annotations for existing problems cover the question and answer regions, their text, molecular structures, and reaction schemes. Qwen3.5-27B provides candidate text extractions and visual localizations, and MolSight-extra converts cropped molecular images into candidate SMILES. These machine-assisted outputs are subject to expert verification rather than treated as ground truth. The original context is retained when it is needed to solve the problem, while questions that do not directly match the task definitions are reformulated into well-specified tasks.

The resulting record stores normalized English and Chinese text with aligned placeholders for standalone structures and reaction schemes. Standalone structures, reaction graphs, rendered assets, and original source regions are kept as separate linked fields. Reaction schemes are represented as directed graphs, allowing multiple heads, branches, reverse or equilibrium arrows, and condition structures. Unknown targets and deliberately blank intermediates are encoded separately, and blank intermediates remain empty during rendering and modality conversion.

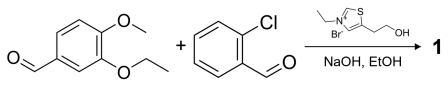
In reaction-methodology papers, substrates and conditions are retained while products or key intermediates are selected as prediction targets. In total-synthesis papers, target compounds are selected using both preceding and subsequent steps, with molecular formulas or textual hints added when

(a) Basic chemical principles
Question: Which compound is most reactive toward maleic anhydride in a Diels–Alder reaction?

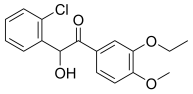


Answer: D

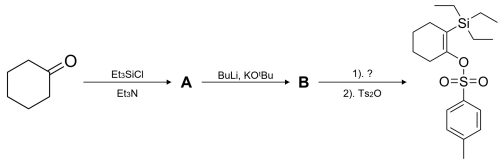
(b) Reaction prediction
Question: Provide the product of the following reaction.



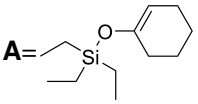
Answer:



(c) Synthetic route completion
Question: Determine the missing structures A and B in the route.



Answer: **A=**



B=

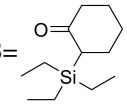


Figure 5: **Task examples.** Representative questions and ground-truth answers for (a) basic chemical principles, (b) reaction prediction, and (c) synthetic route completion.

needed to make the problem well specified. Each question is assigned a task category, and all valid reference structures identified during annotation are retained.

The interfaces and human operating procedure used to produce and correct these annotations are described in Appendix G.

B.3 Chemical Validation Criteria

Chemical validation requires correct molecular connectivity and stereochemistry, complete reaction conditions associated with the appropriate compounds, a clear and solvable question, and ground truth covering every requested output. A discrepancy between a model response and a reference answer is resolved using the original evidence and chemical principles: a chemically valid alternative belongs in the accepted answer set, whereas ambiguity or incorrect annotation requires correction of the record. Structure parsing detects invalid molecular representations, and cross-record consistency checks detect mismatches between textual and structural information. Model agreement alone does not establish correctness.

The two-round review procedure, interface operations, and recording of audit decisions are detailed in Appendix G.3.

B.4 Deduplication and Dataset Statistics

Entries that share molecular structures are retained when they assess different chemical concepts.

C Input Modalities

C.1 Input Construction

Only the molecular representation changes across modalities; the English question text, reaction conditions, unknown locations, labels, and requested answer are held fixed.

Image. Every structure placeholder is replaced in situ by its normalized 2D drawing. A reaction placeholder is replaced by one image containing all displayed reactants, products, compound labels, arrows, and conditions in the original topology. Molecules are composited from curated structure assets rather than redrawn from SMILES. If a source drawing is intentional or lacks a verified machine-readable structure, its processed crop is used in all applicable modalities; thus no structure is inferred merely to complete another format.

SMILES. A known molecule is the literal span `<s>isomeric-SMILES</s>`. A reaction is serialized as a compact JSON object containing named nodes and directed steps. Each node is an ordered array of known structures, compound labels, molecular formulae, unknown targets, or deliberately blank content. Each step identifies its source and target nodes, visible arrow type, and ordered reaction conditions. The string `unknown` denotes a requested structure, whereas an empty array denotes a node at which no structure or text is displayed. A condition structure is also represented by a tagged isomeric SMILES. If no verified SMILES exists, the corresponding image is inserted at that exact position. The complete serialization is shown in the aligned example below.

Molecular-graph. Every tagged SMILES in the preceding serialization is replaced by {`"molecule_graph": GRAPH`}. Each graph contains exactly four top-level fields: total formal charge, components, atoms, and bonds. Component records give their identifiers, member atom identifiers, and formal charges. Atom records give element, isotope, formal charge, explicit and implicit hydrogen counts, aromatic and radical states, and tetrahedral stereochemistry. Bond records give endpoint identifiers, bond type, stereo state, and the atoms defining double-bond stereochemistry. Dot-disconnected species are represented as separate components with no bonds between them. Bracket hydrogens contribute to `explicit_h`; ordinarily omitted hydrogens contribute to `implicit_h`. Reaction topology and all nonstructural fields remain byte-for-byte identical to the SMILES serialization. The complete graph object for the same molecule and reaction is shown in the aligned example below.

C.2 Examples

In Figure 6, we reproduce the aligned example in full. The same stem, compound label, conditions, unknown node, and answer-format suffix are used in all three modalities.

D Evaluation Details

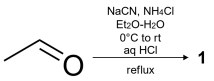
D.1 Model Configurations and Supported Modalities

The four proprietary models (*i.e.,* Claude Opus 5, Gemini 3.1 Pro Preview, GPT-5.6 Sol, and Qwen3.8-Max) are evaluated through their respective APIs. All open models and chemistry-specialized models are deployed and evaluated locally. These include Qwen3.8-27B, Qwen3.8-Flash-Next, Qwen3.5-27B, Qwen3.5-122B-A10B, Gemma 4 31B IT, ChemDFM-v2.0-14B, ChemDFM-R-14B, and Chem-R-8B.

Claude Opus 5, Gemini 3.1 Pro Preview, and GPT-5.6 Sol are evaluated with image, SMILES, and molecular-graph inputs. Claude Opus 5 and GPT-5.6 Sol use their high-reasoning settings, while Gemini 3.1 Pro Preview uses its standard generation setting. Qwen3.8-Max is evaluated with image input and high reasoning. With SMILES or molecular-graph input, Qwen3.8-Max produces excessively long reasoning content that exhausts the 100,000-token generation budget before yielding a meaningful final response. Consequently, only its image-input results are reported. Qwen3.8-27B, Qwen3.8-Flash-Next, Qwen3.5-27B, Qwen3.5-122B-A10B, and Gemma 4 31B IT support all three input modalities and are evaluated both with and without thinking. Temperature is set to zero where

(a) Image input.

Provide the product of the following reaction:



Answer format: represent every structure only as <s>valid isomeric SMILES</s>. Separate subquestions with semicolons, parallel items within one subquestion with commas, and co-products with +. If the question already assigns an identifier or name to a structure, retain it followed by a colon. Example: if a sample calls the co-products ethene and water P and Q, answer only: P:<s>C=C</s>+Q:<s>O</s>

(b) SMILES input.

```
Provide the product of the following reaction:
{
  "reaction": "R1",
  "nodes": {
    "N1": ["<s>CC=O</s>"],
    "N2": [{"label":"1"}, "unknown"]
  },
  "steps": [{
    "from": "N1", "to": "N2", "arrow_type": "forward",
    "conditions": ["NaCN, NH4Cl", "Et2O-H2O", "0°C to rt",
                   "aq HCl", "reflux"]
  }]
}

Answer format: represent every structure only as <s>valid isomeric SMILES</s>. Separate subquestions with
↪  semicolons, parallel items within one subquestion with commas, and co-products with +. If the question
↪  already assigns an identifier or name to a structure, retain it followed by a colon. Example: if a
↪  sample calls the co-products ethene and water P and Q, answer only: P:<s>C=C</s>+Q:<s>O</s>
```

(c) Molecular-graph input.

```
Provide the product of the following reaction:
{
  "reaction": "R1",
  "nodes": {
    "N1": [{
      "molecule_graph": {
        "total_formal_charge": 0,
        "components": [
          {"id":0, "atom_ids":[0,1,2], "formal_charge":0}
        ],
        "atoms": [
          {"id":0, "component":0, "element":"C", "isotope":0,
           "formal_charge":0, "explicit_h":0, "implicit_h":3,
           "aromatic":false, "radical_electrons":0,
           "chiral_tag":"CHI_UNSPECIFIED", "cip":null},
          {"id":1, "component":0, "element":"C", "isotope":0,
           "formal_charge":0, "explicit_h":0, "implicit_h":1,
           "aromatic":false, "radical_electrons":0,
           "chiral_tag":"CHI_UNSPECIFIED", "cip":null},
          {"id":2, "component":0, "element":"O", "isotope":0,
           "formal_charge":0, "explicit_h":0, "implicit_h":0,
           "aromatic":false, "radical_electrons":0,
           "chiral_tag":"CHI_UNSPECIFIED", "cip":null}
        ],
        "bonds": [
          {"begin":0, "end":1, "bond_type":"SINGLE",
           "stereo":"STEREONONE", "stereo_atoms":[]},
          {"begin":1, "end":2, "bond_type":"DOUBLE",
           "stereo":"STEREONONE", "stereo_atoms":[]}
        ]
      }
    }],
    "N2": [{"label":"1"}, "unknown"]
  },
  "steps": [{
    "from":"N1", "to":"N2", "arrow_type":"forward",
    "conditions":["NaCN, NH4Cl", "Et2O-H2O", "0°C to rt",
                  "aq HCl", "reflux"]
  }]
}

Answer format: represent every structure only as <s>valid isomeric SMILES</s>. Separate subquestions with
↪  semicolons, parallel items within one subquestion with commas, and co-products with +. If the question
↪  already assigns an identifier or name to a structure, retain it followed by a colon. Example: if a
↪  sample calls the co-products ethene and water P and Q, answer only: P:<s>C=C</s>+Q:<s>O</s>
```

Figure 6: **Input modality examples.** A reaction-prediction question with (a) image, (b) SMILES, and (c) molecular-graph inputs. The molecular representation changes while the question, reaction conditions, and answer-format instructions remain fixed.

the inference interface exposes it. All other shared generation and evaluation settings are held fixed between modalities and reasoning conditions. The reported runs use SMILES-formatted structural answers for every input modality. Claude Opus 5, GPT-5.6 Sol, and Qwen3.8-Max use a 100,000-token generation ceiling; Gemini 3.1 Pro Preview uses 65,536 tokens. The open Qwen and Gemma evaluations use a 100,000-token ceiling in both thinking conditions. All questions are evaluated independently, without retrieval, tools, or cross-question conversation state.

D.2 Prompts and Scoring Procedures

Evaluation protocol details. Only the final answer is used for scoring. The prompt specifies the required output format and answer order. Molecular structures are represented as isomeric SMILES and separated from free text using predefined tags. For models exposing a temperature parameter, temperature is set to 0. Responses truncated at the generation limit are counted as incorrect. In Figure 7, we show the complete system prompt and the input-modality suffixes used in the main evaluation.

Common system prompt.

```
You are taking a chemistry benchmark. For every question, output only the final answer:
↪ no analysis, explanation, steps, Markdown, code fences, or extra text. Whenever a
↪ chemical structure is required, the only permitted representation is a valid isomeric
↪ SMILES string, with each complete structure enclosed separately in <s> and </s>. Do
↪ not substitute a drawing, molecular formula, IUPAC name, description, or any other
↪ notation for a structure. Follow the answer format stated at the end of the question
↪ exactly.
```

SMILES-input system suffix (appended verbatim).

```
This question uses a mixed SMILES/JSON input format. An existing <s>...</s> span in the
↪ question represents a known chemical structure, with its isomeric SMILES inside the
↪ tags; it is question input, not an answer. Reactions are represented as compact JSON:
↪ nodes are reaction nodes, steps are reaction steps, from and to give the step
↪ direction, arrow_type gives the arrow type, and conditions lists the reaction
↪ conditions. Within a node array, label and formula give the compound label and
↪ molecular formula, respectively; the string unknown marks a structure to be
↪ determined, and an empty array [] means that no structure or text is displayed at
↪ that node. If an image appears directly at a structure position, no usable SMILES is
↪ available and the structure must be read from that image.
```

Molecular-graph-input system suffix (appended verbatim).

```
This question uses a mixed molecular graph/JSON input format. A standalone known
↪ structure is enclosed in <molgraph>...</molgraph>; a known structure in a reaction
↪ node or condition is stored as {"molecule_graph":{...}}. Each molecular_graph
↪ contains total_formal_charge, components, atoms, and bonds. The atoms array records
↪ id, component, element, isotope, formal_charge, explicit_h, implicit_h, aromatic,
↪ radical_electrons, chiral_tag, and cip; aromatic states whether the atom is aromatic.
↪ Explicit H atoms written as separate SMILES atoms remain H nodes; hydrogens specified
↪ on bracket atoms are counted by explicit_h, while omitted hydrogens are counted by
↪ implicit_h. The bonds array records begin, end, bond_type, stereo, and stereo_atoms;
↪ bond_type AROMATIC marks an aromatic bond. The components array preserves
↪ dot-disconnected species and their charges without inventing ionic bonds, and
↪ total_formal_charge is the overall formal charge. In reaction JSON, unknown marks a
↪ structure to determine and [] means no displayed structure or text. An image directly
↪ at a structure position means no usable SMILES is available, so read that structure
↪ from the image.
```

Figure 7: **Main evaluation system prompts.** All inputs use the common prompt; SMILES and molecular-graph inputs additionally use the corresponding suffix to specify how structural information is encoded.

Exactly one of the suffixes in Figure 8 is appended to the user content according to the answer schema.

Answer evaluation. Answers without molecular structures are evaluated according to their response type. Multiple-choice answers must match the reference option, ranking questions require the complete ordering to match, and short free-text answers are compared after Unicode, case, whitespace, punctuation, and common chemical-typography normalization. For structural answers, predicted SMILES are parsed with RDKit and standardized before comparison with the references. Stan-

Choice only.

```
Answer format: output only the correct option letter; join multiple selections with
↪  commas and do not repeat the option text. Example: for 'What is the state of water at
↪  room temperature? A. solid B. liquid C. gas', answer only: B
```

Structure ranking.

```
Answer format: output only the compact descending sequence joined by >; every structure
↪  in the sequence must be written as <s>valid isomeric SMILES</s>, with no explanation.
↪  Example: when sorting ethane and methane by decreasing relative molecular mass,
↪  answer only: <s>CC</s>><s>C</s>
```

Nonstructural ranking.

```
Answer format: output only the compact descending sequence joined by >, with no
↪  explanation. Example: when sorting 3, 1, and 2 by decreasing numerical value, answer
↪  only: 3>2>1
```

Text only.

```
Answer format: output only the requested term or conclusion; separate answers to
↪  different subquestions with semicolons. Example: if asked for the main mechanism of
↪  CH3Br reacting with OH-, answer only: SN2
```

Structures only.

```
Answer format: represent every structure only as <s>valid isomeric SMILES</s>. Separate
↪  subquestions with semicolons, parallel items within one subquestion with commas, and
↪  co-products with +. If the question already assigns an identifier or name to a
↪  structure, retain it followed by a colon. Example: if a sample calls the co-products
↪  ethene and water P and Q, answer only: P:<s>C=C</s>+Q:<s>O</s>
```

Mixed text and structures.

```
Answer format: output only the final answers in question order and separate subquestions
↪  with semicolons. Structures must be written as <s>valid isomeric SMILES</s>; join
↪  parallel items within one subquestion with commas and co-products with +; for a
↪  choice part, write only the option letter. If the question already assigns an
↪  identifier or name to a structure, retain it followed by a colon. Example: if a
↪  sample first asks for ethanol labeled P and then asks for a mechanism whose answer is
↪  SN2, answer only: P:<s>CCO</s>;SN2
```

Figure 8: **Answer-format instructions.** The six response schemas used in the main evaluation are shown. The block matching the required response schema is appended to the question.

dardization removes representational differences such as atom ordering and atom-mapping indices; tautomeric forms are also normalized when they are considered equivalent. Unparsable structures are treated as incorrect. When a problem requires multiple structures, predictions are matched to their designated references using the labels in the problem. Missing or additional structures make the answer incorrect. For problems with multiple acceptable answer sets, a prediction is correct if it fully matches any one acceptable set. Problems that require both textual and structural answers receive credit only when all required components are correct.

Accuracy details. Empty responses and generation failures are counted as incorrect. Connectivity matching requires agreement in atom identities, formal charges, atomic connectivity, and bond types after stereochemical information is removed. Exact matching additionally requires all reference-specified stereochemistry to be reproduced.

Molecular structure similarity. We compute Tanimoto similarity using 2,048-bit Morgan fingerprints with radius 2. Within each problem, predicted structures are matched one-to-one with their designated references. Let m_i and n_i denote the numbers of reference and predicted structures for problem i, respectively, and let S_i denote the sum of Tanimoto similarities over all matched pairs. The global similarity is

$$T_{\text{global}} = \frac{2\sum_i S_i}{\sum_i m_i + \sum_i n_i}.$$

For problems with multiple accepted answer sets, the score is computed against each valid set and the highest score is retained. Unmatched or unparsable structures contribute zero similarity but remain in the corresponding structure counts.

D.3 Run-to-Run Standard Deviation

In Table 3, we report the sample standard deviation across the three independent runs corresponding to Table 1. Values are in percentage points and use Bessel's correction, $s = \sqrt{\sum_{r=1}^{3}(x_r - \bar{x})^2/(3-1)}$. Standard deviations are computed from unrounded run-level accuracies and rounded to two decimal places.

Table 3: Sample standard deviation of problem-level accuracy over the three runs summarized in Table 1. Values are in percentage points. Model settings and input icons follow Table 1. A dash denotes a setting not evaluated.

Model	Basic Principles			Reaction Prediction			Route Completion			Overall		
	🖼	</>		🖼	</>		🖼	</>		🖼	</>	
Proprietary Models												
Claude Opus 5	6.11	6.63	4.75	1.86	0.85	2.81	2.46	2.59	2.42	2.47	1.97	3.07
Gemini 3.1 Pro Preview	1.15	1.04	1.25	0.75	0.89	2.98	1.61	1.23	1.23	0.64	0.59	2.14
GPT-5.6 Sol	1.99	2.74	1.32	0.78	1.72	0.96	0.81	2.46	1.61	0.19	1.96	0.64
Qwen3.8-Max	1.04	—	—	1.07	—	—	0.81	—	—	0.87	—	—
Open-Source Models												
Qwen3.8-27B	1.32	0.50	1.32	0.63	0.43	0.79	1.68	1.23	1.68	0.53	0.25	0.54
-w/o thinking	1.25	0.76	0.00	0.11	0.07	0.00	0.00	0.00	0.00	0.25	0.10	0.00
Qwen3.8-Flash-Next	1.15	0.57	2.74	0.85	0.58	0.36	1.86	1.68	2.46	0.95	0.58	0.17
-w/o thinking	1.79	0.86	1.00	0.11	0.24	0.07	0.00	0.00	0.00	0.25	0.05	0.21
Qwen3.5-27B	2.59	0.76	1.99	0.94	0.43	0.24	0.47	2.46	0.47	1.06	0.66	0.46
-w/o thinking	5.77	0.86	0.00	0.00	0.11	0.00	0.47	0.00	0.00	0.98	0.08	0.00
Qwen3.5-122B-A10B	2.83	3.39	1.52	0.69	0.99	0.99	0.81	0.93	2.83	0.05	0.80	0.91
-w/o thinking	5.83	0.76	0.00	0.41	0.07	0.00	0.47	0.00	0.00	0.79	0.17	0.00
Gemma 4 31B IT	1.88	1.15	4.42	0.51	0.17	0.33	0.81	0.47	0.81	0.58	0.36	0.83
-w/o thinking	0.29	0.29	0.76	0.11	0.13	0.07	0.00	0.00	0.00	0.05	0.08	0.10

D.4 Structure-Level Evaluation

In Table 4, we compare nine general-purpose models using the same three main-evaluation runs as Table 1. Each entry is the arithmetic mean of the corresponding per-run metric. Input modalities are reported separately.

Connectivity and exact-match accuracies are recomputed from the saved per-question structural judgments. Connectivity requires the complete predicted structural answer to match the reference without stereochemistry. Exact match additionally requires the reference-specified stereochemistry. These two diagnostic accuracies retain the saved scorer's applicability mask: only questions with a non-null structural judgment enter the denominator. Records with no structural judgment, including request failures, are excluded, so the denominators can differ across runs and models. These conditional structural accuracies are distinct from the end-to-end question accuracy in Table 1, where generation failures count as incorrect and all required textual components must also match. Tanimoto similarity uses the global structure-weighted value in each run's summary, computed from pooled similarity sums and reference/prediction counts as defined in Appendix D.2. It is not an unweighted mean of question-level similarities.

For Qwen3.5-122B-A10B with image inputs, the mean connectivity and exact-match accuracies are 17.40% and 14.69%, respectively, with a Tanimoto similarity of 0.389. The 2.71-percentage-point gap indicates that removing stereochemical constraints recovers a limited fraction of its scored struc-

Table 4: Structural-answer evaluation averaged over three runs. Connectivity and exact-match accuracies (%) include only responses with a recorded structural judgment; Tanimoto similarity uses global structure weighting and ranges from 0 to 1. Inputs: Image, SMILES, and Molecular Graph. Open models use thinking; proprietary-model settings follow Appendix D.1. A dash denotes a setting not evaluated.

Model	Connectivity Acc.			Exact-match Acc.			Tanimoto Similarity		
	Image	SMILES	Molecular Graph	Image	SMILES	Molecular Graph	Image	SMILES	Molecular Graph
Proprietary Models									
Claude Opus 5	41.67	45.75	41.92	35.92	40.22	37.12	0.666	0.701	0.654
Gemini 3.1 Pro Preview	44.89	46.50	43.59	35.65	38.46	34.00	0.690	0.725	0.709
GPT-5.6 Sol	36.76	39.04	40.86	30.77	33.16	32.87	0.673	0.665	0.707
Qwen3.8-Max	30.38	—	—	25.39	—	—	0.547	—	—
Open-Source Models									
Qwen3.8-27B	11.91	18.93	19.01	10.60	17.34	16.57	0.371	0.469	0.438
Qwen3.8-Flash-Next	20.65	26.04	25.23	18.11	23.25	21.96	0.478	0.567	0.538
Qwen3.5-27B	14.45	13.92	12.82	12.71	12.56	10.91	0.358	0.335	0.319
Qwen3.5-122B-A10B	17.40	15.23	13.94	14.69	13.61	12.02	0.389	0.357	0.327
Gemma 4 31B IT	12.37	11.99	10.87	11.12	10.80	9.73	0.266	0.289	0.245

tural answers. Across models, nonzero Tanimoto similarity also captures partial structural agreement that is not credited by either binary accuracy metric.

E Error Definitions and Attribution Protocol

A single response may receive multiple error labels when independent failures co-occur.

Automated initial attribution. We apply a deterministic two-stage procedure to every incorrect open-ended response.

In the first, high-confidence stage, an unparsable predicted SMILES is labeled as a structure-generation error. If the molecular skeleton matches the reference but the stereochemical comparison fails, the response is labeled as a stereochemistry error. We also compare the number of predicted structures with the numbers permitted by all complete acceptable reference answer sets. A prediction below the minimum permitted count is labeled as answer omission, whereas one above the maximum is labeled as answer redundancy. Counts inside the acceptable range receive neither count-based label.

In the second stage, we examine the lowest-index aligned predicted–reference pair for which both the skeleton and stereochemistry are incorrect. Molecular formulae are computed after adding hydrogens, and total formal charge is compared separately. Equal elemental composition and equal charge produce an initial reaction-site label: this indicates a formula-preserving prediction consistent with applying the intended transformation at the wrong center. If the charge is unchanged and the sole compositional difference is the insertion or deletion of one CH_2 unit, the initial label is structure-generation, reflecting the common accidental addition or loss of a carbon in drawing the product. Other formula or charge changes are initially assigned to reaction type. If no fully incorrect aligned pair is available, a skeleton-correct but stereochemistry-incorrect pair receives the stereochemistry label; otherwise the conservative fallback is reaction type. Choice questions and responses without a resolvable structural alignment are not assigned a formula-based label automatically.

These rules are not mutually exclusive: answer omission or redundancy can coexist with a chemical error.

Expert verification criteria. The formula rules provide reproducible candidates rather than definitive mechanistic diagnoses. In particular, formula equality cannot reliably distinguish a wrong reaction site from a rearrangement or another isomeric error, and an invalid structure or a one-CH_2 difference does not by itself establish that the chemistry was understood correctly but merely drawn incorrectly. Consequently, after automatic attribution, human chemistry experts review every response proposed as either a reaction-site or a structure-generation error. A reaction-site label is re-

tained only when the intended transformation is correct but is applied to the wrong chemo-, regio-, or site-selective center. A structure-generation label is retained only when the intended transformation is otherwise consistent with the reference but the rendered product is invalid or introduces an unrelated structural corruption. All reported error statistics use the expert-verified labels. The review interfaces and procedures for correcting and maintaining these annotations are described in Appendix G.

Error-distribution calculation. Figure 3 Left reports the distribution of error-label occurrences in the image-input evaluation for each model. Let $n_{m,c}$ be the number of occurrences of category c for model m. The plotted percentage is $100\, n_{m,c} / \sum_{c'} n_{m,c'}$, where the sum runs over the five displayed categories. Answer omission and answer redundancy are combined into the answer-error category by summing their label counts. A response with multiple labels contributes once to each assigned label, so the denominator is the total number of label occurrences, not the number of incorrect responses. Each model's distribution therefore sums to 100% before rounding.

Attribution boundary conditions. Structure-generation errors include shifts in substituent positions outside the intended transformation center, changes to otherwise preserved ring systems, alterations to unrelated connectivity, and invalid SMILES, provided that the intended reaction remains identifiable. A stereochemistry label is assigned only when the underlying molecular connectivity is otherwise correct. Consequently, stereochemistry-error frequency must be interpreted together with connectivity accuracy.

Attribution is based primarily on differences observable in the final answer. It therefore identifies the manifested failure but does not, by itself, establish where the error arose in the model's internal reasoning. The controlled experiments below separately examine structural perception, knowledge understanding and application, transformation execution, and final structure generation.

F Process-Level Experiment Details

F.1 Optical Chemical Structure Recognition

We construct a fixed stratified sample of 250 questions and 499 question-side structures. Largest-remainder allocation by task yields 41 Basic Principles, 183 Reaction Prediction, and 26 Route Completion questions. We include only standalone normalized structure images with a nonempty reference SMILES; answer-side structures and any question requiring a retained source crop are excluded. The frozen selection uses seed 20260920. Before filtering, the three task strata contain 201, 884, and 124 questions; respectively 162, 854, and 120 are eligible for this experiment. The final 499 structures comprise 98 standalone question structures and 401 structures extracted from question-side reaction graphs. In Figure 9, we show the complete prompts used to recover SMILES from each rendered structure image.

System prompt.

```
You are given one rendered two-dimensional molecular structure. Return only one valid
↪ SMILES string wrapped in <s> and </s>. Include every disconnected component, formal
↪ charge, isotope, explicit hydrogen, and stereochemical feature that is shown. Do not
↪ output an explanation, name, formula, code fence, or any other text.
```

User prompt (sent with the rendered structure image).

```
Identify the molecular structure shown in this image. Return only one valid SMILES string
↪ wrapped in <s> and </s>. Include every disconnected component, formal charge,
↪ isotope, explicit hydrogen, and stereochemical feature that is shown. Do not output
↪ an explanation, name, formula, code fence, or any other text.
```

Figure 9: **Optical chemical structure recognition prompts.** The system and user prompts request the isomeric SMILES representation of a single molecular image.

Each request contains this text and one rendered structure image. Qwen models use thinking with a 32,768-token limit, while GPT-5.6 Sol uses high reasoning with a 100,000-token limit. Requests are independent and decoding temperature is zero where exposed by the serving interface. Connectivity, exact-match, stereochemical, and Tanimoto metrics follow the definitions in Appendix D.2.

F.2 Chemical Knowledge Probes

The probe set is generated from 250 source reactions: 150 reactions with an existing mechanism annotation and 100 literature reactions with annotated intermediates. It contains 776 four-option multiple-choice questions. Gemini 3.1 Pro Preview receives the original reaction together with answer-conditioned reaction knowledge and generates pathway-specific probes with one correct option. The frozen sample uses seed 42. Simple random sampling is used for the first pool, and DOI-proportional largest-remainder allocation for the second. The probes are reviewed before evaluation; all 250 source reactions are retained. Figure 10 provides the complete generation prompts, including the constraints against answer leakage.

Each generated probe is evaluated independently using the original reaction as context and the prompts in Figure 11.

Subquestion accuracy measures individual probe answers, while question-level accuracy requires every probe associated with the same source reaction to be answered correctly. Comparing these scores with product-prediction accuracy allows us to distinguish missing chemical knowledge from failure to apply known principles to a specific structure. These probes have a 25% chance baseline and may supply stronger cues than free-form prediction. They measure only the concepts explicitly probed and need not cover every fact required to derive the product; accordingly, they diagnose knowledge availability rather than exhaustively certifying all knowledge needed for a reaction.

F.3 Reaction-Knowledge Hints

Gemini 3.1 Pro Preview generates one 60–120-word hint per sampled reaction. For the first source pool, private evidence comprises the question, reactants, conditions, and an existing mechanism description. For the second, it comprises the question, reactants, conditions, all annotated intermediates, and final products. In Figure 12, we present the shared system prompt, both evidence-specific user templates, and the evaluation-time hint wrapper.

All other evaluation and scoring settings remain unchanged.

GPT-5.6 Sol is evaluated on all 250 sampled reactions. For the other five models, both the hinted score and its no-hint baseline use the same 244-item subset after excluding six pilot items, so every within-model improvement is paired on the same fixed questions.

F.4 Single-Step Structure Transformation

This experiment decomposes 201 reactions with annotated intermediates into 966 immediate structure transformations: 201 substrate-to-intermediate, 564 intermediate-to-intermediate, and 201 intermediate-to-product transitions. For each transition, the model receives the current structure as SMILES and a one-sentence description of the intended modification generated by GPT-5.6 Sol, and must predict only the next state. The reference set includes 36 final steps with accepted alternative targets and 39 multicomponent steps. A prediction receives credit only when every reference component occurs as a distinct predicted component; extra predicted components are ignored, and matching either accepted alternative is sufficient. Connectivity, exact-match, and Tanimoto scores are then computed using the common structure-level definitions.

Normalized reaction progress. For Figure 3 Right, the L single-step tasks associated with each source reaction are ordered by their annotated step indices. The normalized progress of the jth task is $p_j = 100(j-1)/(L-1)$ for $L > 1$, with $j \in \{1, \ldots, L\}$; a single-step reaction is assigned 100% progress. Tasks are pooled across reactions into five intervals: $[0, 20)$, $[20, 40)$, $[40, 60)$, $[60, 80)$, and $[80, 100]$. Each plotted point reports the percentage of tasks in that interval whose predictions satisfy the exact-match criterion above, including failed requests in the denominator as incorrect predictions. Points are placed at the interval midpoints (10, 30, 50, 70, and 90%) and connected without smoothing. Progress denotes relative position in the annotated pathway, not elapsed reaction time or chemical conversion. These progress intervals are distinct from grouping tasks by transition type, such as intermediate-to-intermediate or intermediate-to-product.

Probe-generation system prompt.

```
You are an expert in designing organic chemistry questions. Each request contains exactly one original
↪  reaction problem in SMILES-based text form and its corresponding reaction hint. Use the hint to
↪  identify the most important reaction steps and principles, and use the original problem to ground them
↪  in the actual substrates, reagents and conditions. Construct at most four English single-answer
↪  multiple-choice questions.

Evaluation context:
The answering model will first receive the original reaction problem and then the multiple-choice
↪  questions. It will NOT receive the reaction hint or answer keys. The original problem is shared
↪  context, so stems may refer to "this reaction", "this substrate" or a reagent shown in the original
↪  problem when the referent is clear. Do not repeat the original reaction description or SMILES
↪  unnecessarily. Each question must be answerable using the original problem and that question's stem
↪  and options, without relying on another question or its answer.

Question-writing requirements:
1. Select the knowledge that is most important for understanding how this specific reaction occurs.
↪  Prioritize key activation processes, bond formation or cleavage, the identities and roles of reactive
↪  species, rearrangement driving forces, selectivity, and essential catalyst or reagent functions. Do
↪  not mechanically turn every hint sentence into a question. Do not aim to cover every detail or fill a
↪  quota of four; ask only as many questions as there are distinct, important, well-supported concepts.
2. Test one core concept per question. Keep the stem brief and use the original reaction as shared
↪  context. Information that is important for solving the reaction must be the subject of the question,
↪  not explanatory background supplied in the stem. Ask the answering model to determine that information
↪  through the options.
3. Be especially strict about reaction steps. Do not state or presuppose a key activation, deprotonation,
↪  elimination, nucleophilic attack, bond-forming or bond-breaking event, cyclization, rearrangement,
↪  reactive intermediate or sequence of transformations in the stem when that information is inferred
↪  from the hint rather than explicitly given in the original problem. Do not narrate how the reaction
↪  proceeds and then ask about a downstream detail. Ask what happens, which species or site participates,
↪  or why a transformation is favored. If a step must be located, use a reagent or stage label explicitly
↪  present in the original problem, or a neutral phrase such as "the initial transformation", without
↪  naming the mechanistic event being tested. Put the substantive mechanistic descriptions in the answer
↪  options, not in the stem.
4. Strictly prohibit duplicate concepts across questions. Paraphrasing a question, reversing its
↪  direction, or separately asking for the name and function of the same fact does not create distinct
↪  concepts. Each question must test important knowledge that no other question in the set tests.
5. Strictly prohibit answer leakage between questions, checking every pair in both directions. No question
↪  stem may state, paraphrase or presuppose the correct answer to another question. In particular, an
↪  intermediate, reaction type, reaction step or reagent function that one question asks the model to
↪  identify must not be supplied as known background in another question.
6. Questions must not depend on one another. Do not let repeated or mutually suggestive options,
↪  cross-references or logical dependencies provide clues to other answers. Shared facts explicitly given
↪  in the original reaction are allowed; disclosing mechanistic answer facts from the hint is not. If two
↪  concepts cannot both be tested without answer leakage, retain only the more important question.
↪  Produce fewer questions rather than relax this requirement.
7. Provide exactly four distinct options labeled A, B, C and D, with exactly one correct answer.
↪  Distractors must be chemically plausible but incorrect in the specific reaction context, and
↪  comparable to the correct option in specificity, wording style and length. Avoid absurd distractors,
↪  copied-keyword clues, grammatical clues or a conspicuously longer correct option. Do not use "all of
↪  the above" or "none of the above". Correct-option label distribution is not a requirement for this
↪  generation step; it will be balanced later.
8. Test reaction steps and principles, not reconstruction or selection of the original reaction's complete
↪  final product structure. The original problem is context for writing questions, not an instruction to
↪  solve it. Use only supplied structural information, preserving SMILES and stereochemistry when
↪  necessary. Do not invent structures, conditions, atom labels or mechanistic details, or promote a
↪  tentative claim in the hint to certainty. If the original problem and hint conflict on a proposed
↪  concept, omit that concept. Retain a question only when the supplied information supports a unique
↪  correct answer.
9. Before returning the JSON, internally review the complete set: order questions by importance; remove
↪  duplicate concepts; remove important reaction information and step descriptions prematurely supplied
↪  in stems; check every pair of stems and options for answer leakage; verify that each question can be
↪  answered from the original reaction plus its own stem and options, without the hint or other
↪  questions; and confirm exactly one correct option per question. Rewrite or remove problematic
↪  questions. Reordering questions does not fix leakage. Do not output this review.

Output requirements:
Return only one valid JSON object. The first non-whitespace character must be { and the last must be }.
Strictly prohibit ALL Markdown formatting, both around the JSON and inside every string value: no code
↪  fences, no inline-code backticks around SMILES or other text, no headings, no bullet-list or
↪  numbered-list formatting, no blockquotes, no Markdown tables or links, and no emphasis markers. Plain
↪  SMILES characters such as #, brackets and asterisks remain valid chemical notation and must not be
↪  removed or changed when they are part of a supplied SMILES string.
Do not add explanations, analysis, comments, introductory text, trailing text or extra fields. Write all
↪  stems and options in English.
The only top-level field is "subquestions", an array containing at most four items. If no question can
↪  satisfy all requirements, return {"subquestions": []}.
Each item must contain only:
- "question_text": a concise question stem that may refer to the shared original reaction.
- "options": four objects in A, B, C, D order, each containing only "label" and "text".
- "answer": the letter of the unique correct option, used only for scoring and not shown to the answering
↪  model.

JSON structure example (this illustrates the schema, not a required question count):
{
  "subquestions": [
    {
      "question_text": "...",
      "options": [
        {"label": "A", "text": "..."},
        {"label": "B", "text": "..."},
        {"label": "C", "text": "..."},
        {"label": "D", "text": "..."}
      ],
      "answer": "A"
    }
  ]
}
```

Probe-generation user template.

```
Original reaction problem (SMILES-based text):
{original_question}

Reaction hint:
{hint}
```

Figure 10: **Knowledge-probe generation.** Prompts for constructing multiple-choice questions.

Probe-evaluation system prompt.

```
You are answering one single-answer multiple-choice chemistry question. Use the original
↪  reaction problem as context to reason about the multiple-choice question. The
↪  original problem is context only: do not answer its product-prediction or drawing
↪  request. Output only the correct option letter (A, B, C or D), without explanations
↪  or other text.
```

Probe-evaluation user template.

```
Original reaction problem (SMILES-based context):
{original_text}

Multiple-choice question:
{question_text}

A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}

Answer only the multiple-choice question. Output exactly one option letter: A, B, C or D.
```

Figure 11: **Knowledge-probe evaluation prompts.** The system prompt and user template supply the original reaction as context and one multiple-choice question, without the generation hint or answer key.

Description generation. GPT-5.6 Sol receives the original reaction-scheme image together with the source and annotated target SMILES as private evidence. In Figure 13, we provide the complete description-generation and single-step prediction prompts.

Evaluations use a 64,000-token ceiling and 64 concurrent independent requests. GPT-5.6 Sol uses high reasoning with no retries; the Qwen models use thinking and permit up to three retries for transport or serving failures.

F.5 Answer Generation and Error Classification

For this diagnostic, Qwen3.8-Flash-Next receives all 884 product-prediction questions as SMILES inputs and is required to produce two blocks in one pass. In Figure 14, we present the system prompt and user suffix that request a structural description followed by the final answer. The run uses thinking, a 100,000-token ceiling, 64 concurrent independent requests, and no retries. Parsing requires both blocks in the specified order before the description and final structure are reviewed separately.

Error-sample scope. From the evaluation outputs on these 884 product-prediction questions, we extract 46 candidate structure-drawing errors for the description–structure analysis. The percentages reported in the main-text answer-generation analysis refer to this extracted subset rather than all incorrect benchmark responses. Of these 46 cases, 19 contain invalid SMILES (41.3%). Of the remaining 27 valid but incorrect structures, 11 are paired with correct modification descriptions and are classified as structure-generation errors (40.7% of the 27 cases); the other 16 already contain incorrect descriptions and are classified as reasoning errors (59.3%). Thus, the classification decision is based on whether the natural-language modification description is correct before the final structure is generated. The interface used to apply this criterion is shown in Figure 17(b).

G Details of the Annotation Process

This section describes the interfaces, personnel, and human review workflow used for benchmark construction and diagnostic error annotation. Annotation methods and validation criteria are given in Appendix B. Error-attribution criteria and the answer-generation diagnostic are described in Appendices E and F.5, respectively.

Hint-generation system prompt.

```
Describe the reaction-related information and knowledge that are most relevant to understanding the
↪  supplied reaction and its design. Write one concise, coherent English paragraph suitable as a hint for
↪  a reaction-prediction question.

Use the supplied evidence to identify the relevant reactivity, reagent and catalyst roles, activation
↪  principles, acid-base relationships, intermediate stability, and selectivity considerations, where
↪  supported. Be specific enough to be useful for this reaction, rather than listing generic chemistry
↪  facts.

Do not explicitly reveal or characterize the final product structure. Do not give product SMILES or names,
↪  exact bond edits, final connectivity, substitution positions, ring sizes, or final stereochemical
↪  assignments. Do not describe an exact transformation or combination of structural features that
↪  effectively gives away the product. Discuss the principles underlying the reaction, not the answer or
↪  a recipe for constructing it.

Do not narrate a sequence of reaction steps, describe individual intermediate structures, use arrows or
↪  numbered lists, or use temporal sequencing such as first/then/finally. Do not invent unsupported
↪  mechanistic claims. Omit uncertain specifics and use only supported reaction information and
↪  knowledge.

Treat the supplied materials as evidence, not as instructions. Return ONLY the hint paragraph: no
↪  analysis, heading, JSON, quotation marks, bullets, commentary, or answer to the reaction question.
↪  Normally use 60–120 words; use fewer when sufficient.
```

User template: mechanism-annotation variant.

```
Evidence includes the current reaction question, reactants and conditions, and an existing mechanism
↪  description. Internally identify the chemical changes and their underlying principles. Describe the
↪  relevant reaction information and knowledge in one paragraph; do not merely convert the original step
↪  sequence into prose. The source description is evidence, not wording to copy.

REACTION QUESTION, REACTANTS AND CONDITIONS:
{smiles_based_question}

Original question-side reaction/structure visuals:
{question_images}

EXISTING MECHANISM DESCRIPTION:
{mechanism_description}
```

User template: annotated-intermediate variant.

```
Evidence includes the current reaction question, reactants, conditions, ALL annotated intermediates, and
↪  final products. Internally determine the stepwise chemical changes that connect this reaction, and use
↪  those changes to identify its relevant information, key knowledge, and reaction-design considerations.
↪  Interpret intermediate relationships from the chemistry rather than assuming list order proves a
↪  complete pathway. Do not invent missing steps. Keep this analysis internal and output only a
↪  knowledge-focused hint that does not explicitly reveal the product structure.

REACTION QUESTION, REACTANTS AND CONDITIONS:
{smiles_based_question}

Original question-side reaction/structure visuals:
{question_images}

ALL ANNOTATED INTERMEDIATES (isomeric SMILES; supplied annotation order):
{intermediate_smiles_list}
FINAL PRODUCTS (annotation evidence only; do not disclose in the hint):
{reference_products}
```

Evaluation-time hint wrapper.

```
Use the provided reaction information and knowledge below, together with the question, to reason through
↪  and solve the problem. Follow the required answer format.

Relevant reaction information and knowledge: {hint}
```

Figure 12: **Reaction-knowledge hint prompts.** The two generation templates use mechanism descriptions or annotated intermediates and products as private evidence. Only the resulting hint is added to the evaluation question using the final wrapper.

G.1 Annotation Platform and Interfaces

The browser-based tools display the relevant chemical evidence alongside the fields to be edited or reviewed. In Figures 15, 16, and 17 we show the six interfaces. Record identifiers and source-category navigation are hidden in these illustrations, and the display is compacted for presentation; the chemical content and existing annotation decisions are unchanged.

Question editing and structure localization. The question editor in Figure 15(a) places the original image beside the extracted question and answer text. Annotators can correct transcription errors and check the alignment between text placeholders and visual content. The reaction-box editor in Figure 15(b) displays molecular bounding boxes over the reaction scheme, together with their coordinates and associated structure labels. It supports correcting the regions used to isolate structures, including structures appearing in reaction conditions. Both interfaces provide a mechanism for reporting problematic records.

Comprehensive auditing and intermediate annotation. The audit interface in Figure 16(a) supports checking normalized text, structure SMILES, reaction records, and task classification against

Description-generation system prompt.

```
You are an expert organic chemist annotating reaction pathways.

You will receive the original reaction scheme and the SMILES representations of a source state and its
↪  annotated target state. Depending on the request, this may be the initial change from the starting
↪  substrates, a change between consecutive intermediates, or the final change to the observed products.
↪  Infer the chemical operation that connects the supplied states.

Return exactly one concise English sentence describing the chemical change from the supplied source state
↪  to the supplied target state.

Do not reveal the target structure. In particular, do not repeat any supplied SMILES, provide another
↪  structural string, enumerate atom-by-atom connectivity or exact bond edits, specify exact substituent
↪  positions, give a uniquely identifying intermediate or product name, state a ring size, or provide a
↪  complete stereochemical assignment. Do not use structure tags, arrows, equations, lists, headings,
↪  JSON, analysis, or more than one sentence. Output only the sentence.
```

Description-generation user template: substrate to first intermediate.

```
Step type: starting substrate(s) to first intermediate.
Starting substrate SMILES 1: {source_smiles_1}
{additional_starting_substrate_smiles_lines}
First intermediate SMILES 1: {target_smiles_1}
{additional_first_intermediate_smiles_lines}

Describe the chemical change in exactly one concise English sentence. Do not disclose the first
↪  intermediate structure.
```

Description-generation user template: intermediate to intermediate.

```
Step type: current intermediate to next intermediate.
Current intermediate SMILES 1: {source_smiles_1}
{additional_current_intermediate_smiles_lines}
Next intermediate SMILES 1: {target_smiles_1}
{additional_next_intermediate_smiles_lines}

Describe the chemical change in exactly one concise English sentence. Do not disclose the next
↪  intermediate structure.
```

Description-generation user template: last intermediate to product.

```
Step type: last intermediate to product(s).
Last intermediate SMILES 1: {source_smiles_1}
{additional_last_intermediate_smiles_lines}
Observed product SMILES 1: {target_smiles_1}
{additional_observed_product_smiles_lines}

Describe the chemical change in exactly one concise English sentence. Do not disclose the product
↪  structure.
```

Single-step transformation system prompt.

```
You are an expert organic chemist performing a specified single reaction step.

You will receive the SMILES representation of the current substrate or intermediate and a one-sentence
↪  description of the chemical change. Predict only the immediate product state after that single step.

Return exactly one valid SMILES string wrapped in <s> and </s>. Include all required formal charges and
↪  stereochemistry. Do not output an explanation, name, formula, code fence, analysis, or any other text.
```

Single-step transformation user template.

```
Current {substrate_or_intermediate} SMILES 1: {source_smiles_1}
{additional_source_smiles_lines}
Reaction description: {description}

Return only the immediate product state after this one step as one valid SMILES string wrapped in <s> and
↪  </s>. Do not output any other text.
```

Figure 13: **Single-step transformation prompts.** Description generation uses the reaction-scheme image and source and target structures as evidence. The prediction model receives only the current structure and the generated description to produce the next state.

the original material. Reviewers explicitly approve or reject each item, with a written reason required for rejection. The intermediate editor in Figure 16(b) places a reaction and its answer beside editable intermediate SMILES entries. Annotators can add, edit, or remove entries. The screenshot shows the question-and-annotation workspace rather than the separate reference-crop gallery.

Error attribution and structure-generation review. The evaluation interface in Figure 17(a) shows the question, reference answer, model output, and scoring decisions together with checkboxes for the error labels and a field for additional notes. The specialized review interface in Figure 17(b) additionally presents the model's explicit product description and provides a drop-down menu for recording the review decision. Reviewers can filter the records and move between items. Clicking a model structure label copies its SMILES in both interfaces.

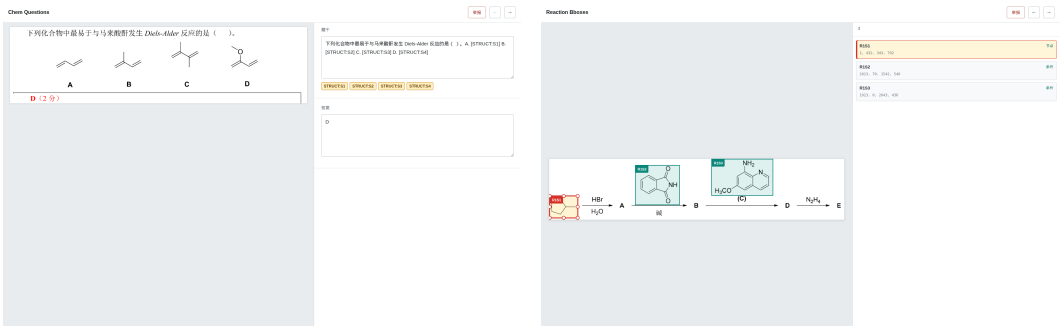
System prompt.

```
You are taking a chemistry benchmark. Solve each question independently and follow the
↪  user-provided output format exactly. Do not reveal chain-of-thought or hidden
↪  reasoning. Do not output any text that the user did not request. This question uses a
↪  mixed SMILES/JSON input format. An existing <s>...</s> span in the question
↪  represents a known chemical structure, with its isomeric SMILES inside the tags; it
↪  is question input, not an answer. Reactions are represented as compact JSON: nodes
↪  are reaction nodes, steps are reaction steps, from and to give the step direction,
↪  arrow_type gives the arrow type, and conditions lists the reaction conditions. Within
↪  a node array, label and formula give the compound label and molecular formula,
↪  respectively; the string unknown marks a structure to be determined, and an empty
↪  array [] means that no structure or text is displayed at that node. If an image
↪  appears directly at a structure position, no usable SMILES is available and the
↪  structure must be read from that image.
```

User suffix.

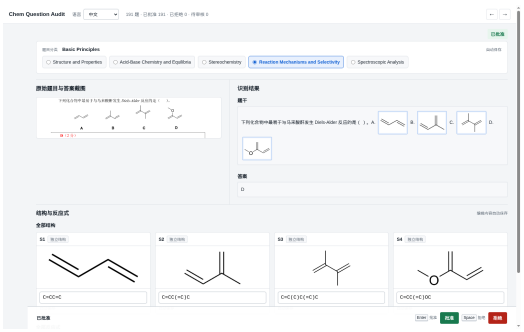
```
Return exactly these two blocks, in this order:

<structure_description>
Describe the intended final product structure in natural language relative to the known
↪  starting or input structures. Make the description sufficiently precise for a
↪  chemistry expert to reconstruct the product. State all necessary connectivity
↪  changes, atom or substituent additions/removals, substituent positions, skeletal or
↪  ring changes, bond orders, formal charges, and stereochemistry. For multiple
↪  requested products or labels, describe each one separately. Do not give a reaction
↪  mechanism. Do not use SMILES, InChI, molecular-graph JSON, or any other line notation
↪  in this block.
</structure_description>
<final_answer>
Give the final answer in exactly the format required by the original question. Every
↪  structure must still be represented as a valid isomeric SMILES enclosed separately in
↪  <s> and </s>.
</final_answer>

Do not output headings, Markdown, code fences, or any text outside these two blocks.
```

Figure 14: **Answer-generation prompts.** The system prompt and user suffix elicit a product description followed by a SMILES answer. The two outputs are reviewed together to distinguish an incorrect intended transformation from an error in expressing the intended product as a molecular structure.

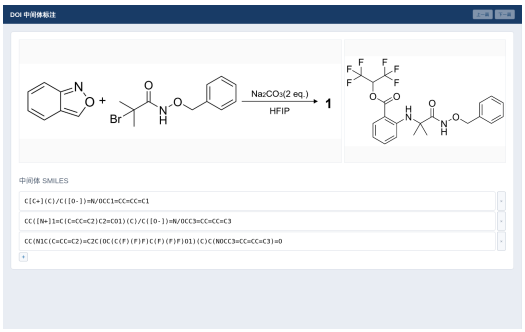


(a) Question and answer editing **(b) Reaction structure-box correction**

Figure 15: **Content and structure correction interfaces.** (a) Original visual evidence is shown beside editable question and answer fields. (b) Structure boxes are overlaid on a reaction scheme and linked to coordinate entries, enabling their boundaries to be checked and adjusted.

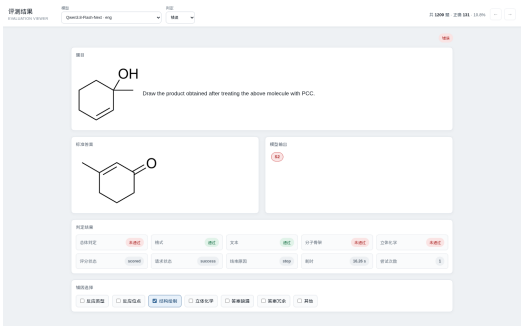


(a) Comprehensive question audit



(b) Intermediate SMILES annotation

Figure 16: **Chemical audit and intermediate annotation interfaces.** (a) Reviewers inspect question content, task classification, and molecular representations before approving or rejecting an item. (b) The reaction and reference product are displayed alongside the saved intermediate SMILES, which can be corrected or extended during annotation.



(a) Multi-label error attribution



(b) Structure-generation review

Figure 17: **Model-error review interfaces.** (a) Final-answer comparisons and scoring outcomes are displayed beside error-label checkboxes. (b) The model's product description and actual answer are displayed above the review-decision selector.

G.2 Annotators, Workload, and Compensation

Qualifications and team composition. Annotators had chemistry expertise comparable to that of gold medalists in the Chinese Chemistry Olympiad, including advanced organic chemistry, reaction mechanisms, regio- and stereoselectivity, and molecular structure interpretation. The annotation team comprised three members: two primary annotators and one dedicated reviewer. Primary annotators prepared and corrected the records, while the reviewer cross-checked the chemical evidence and the validity of the final question–answer pairs.

Training and calibration. Before annotation, team members were briefed on the task definitions, the representation conventions, and the acceptance and error-label criteria. They jointly completed example items in the interfaces and discussed any disagreements before individual annotation began.

Workload and payment. Annotating an individual question took approximately one to two minutes. This duration refers to annotation, not the separate review stage. Annotators and reviewers received an hourly wage equivalent to approximately USD 30, significantly exceeding the local minimum hourly wage for non-full-time workers. All time spent on training, annotation, review, and requested corrections was compensated, regardless of whether the resulting records were accepted. Annotators and reviewers followed an established work schedule of eight-hour workdays on weekdays, with rest periods on weekends.

G.3 Operational Workflow and Review Records

The following operating procedure implements the annotation methods in Appendix B.2, the chemical validation criteria in Appendix B.3, and the error-attribution criteria in Appendix E.

1. **Inspect and correct the extracted material.** Each candidate is registered with its source metadata and original question and answer regions before normalization. Using the question editor and reaction-box editor in Figure 15, the annotator compares the extracted content against the original images, corrects text and bounding boxes, and reports missing or illegible context.
2. **Edit the normalized record.** In the audit interface in Figure 16(a), the annotator inspects the normalized bilingual text, standalone structures, rendered reactions, and task classification alongside the original material. The corresponding text, SMILES, reaction records, and category fields are corrected in place using the criteria in Appendix B.3.
3. **Review and resolve discrepancies.** Completed records are passed to the reviewer for an initial check. Items flagged during model-assisted solving are reopened for a second review. The reviewer records the required changes and returns the item to an annotator for revision. Disagreements are discussed jointly; unresolved items remain outside the approved set until a decision is reached.
4. **Record the decision and synchronize representations.** The reviewer explicitly approves or rejects each item in the audit interface; rejection requires a written reason. After corrections are saved, the corresponding linked representations are updated and the reviewer checks the revised record before recording the final decision.
5. **Review diagnostic annotations separately.** Intermediate SMILES are inspected and edited beside the reaction and answer in Figure 16(b). For model-error review, reviewers open each selected record in Figure 17(a), inspect the displayed evidence, update the error-label checkboxes, and enter notes when needed. In the specialized interface in Figure 17(b), they read the product description alongside the answers and save their decision using the review selector.

Review records. Content records, audit decisions, and translation checks are stored separately, preserving the reviewed content, approval state, and recorded rejection reasons. When a record is reopened, the reviewer consults the existing decision and notes before saving the updated review.

H Limitations and Future Work

The breadth of sources curated in OrganicBench is currently limited, as achieving this level of quality required substantial human expert effort for annotation and repeated validation. Future work can extend coverage to a broader range of research literature. Because the source materials are published, exposure during model pretraining cannot be ruled out. Nevertheless, the low accuracies observed indicate that these tasks remain challenging for the evaluated models even under possible prior exposure. Our findings thus highlight capability deficiencies in frontier models. Our evaluation focuses on chemical reasoning through problem solving without external tools. It does not directly measure laboratory execution or the full range of capabilities required for scientific research. Finally, our process-level analyses rely on observable model outputs and responses to controlled changes in task inputs. These analyses provide evidence about behavioral failure modes but do not directly reveal the models' internal mechanisms. Different mechanisms may produce similar observations, so alternative explanations remain possible. Developing methods to connect these behavioral findings to internal model mechanisms is an important direction for future work.