ChemEval2: A Diagnostic Benchmark for Evaluating Tool-Integrated Reasoning in Chemistry

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are transforming chemical research by enabling specialized reasoning models and tool-augmented agents. Yet existing benchmarks typically evaluate reasoning or tool invocation in isolation, overlooking the dynamic interaction between internal deliberation and external execution. We present **ChemEval2**, a multimodal benchmark for tool-augmented chemical reasoning, containing 2,410 samples spanning fundamental principles, spectroscopic analysis, and complex reaction understanding. Built from professional textbooks and experimental data through automated generation and rigorous human annotation, ChemEval2 introduces a *tool-augmented Chain-of-Thought* evaluation framework that embeds tool-use checkpoints at stages where unaided language reasoning becomes unreliable. This design evaluates whether models can identify information gaps, invoke appropriate tools, and integrate tool outputs into reasoning trajectories. Systematic evaluation of SOTA models reveals a prevalent process–outcome inconsistency: models often complete intermediate reasoning steps but fail to combine them with tool outputs to produce correct final answers. These findings indicate that the major bottleneck of current chemical AI lies in the orchestration of complex reasoning and tool use. We open-source ChemEval2 and an MCP-based suite of specialized chemical tools for reproducible[1].

1 Introduction

Large Language Models (LLMs) are reshaping scientific discovery by enabling both long-chain reasoning and tool-augmented problem solving (Hu et al., 2025; Achiam et al., 2023; Zhang et al., 2025b; Boiko et al., 2023; Yang et al., 2025b). Recent reasoning-centric models, such as OpenAI-o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025), have demonstrated the potential of reinforcement learning to internalize complex deliberation. In chemistry, models such as ChemDFM-R (Zhao et al., 2025b) and RETRO-R1 (Liu et al.) further improve chemical understanding by explicitly modeling reaction principles and mechanistic logic (Narayanan et al., 2025; Zhang et al., 2025a; Ferrag et al., 2025). Meanwhile, agentic frameworks such as ChemCrow (M. Bran et al., 2024), SciToolAgent (Ding et al., 2025), and CheMatAgent (Wu et al., 2025) extend LLMs beyond static knowledge by integrating specialized chemical tools for analytical calculation and multi-step planning (Ramos et al., 2025; Kim et al., 2025; Yu et al., 2025; Li et al., 2025c). These advances suggest that next-generation chemical AI will increasingly depend on the coordinated interaction between internal reasoning and external tool use.

To evaluate the specialized capabilities of chemical LLMs, several benchmarks have recently been proposed. ChemCoTBench (Li et al., 2025a), oMeBench (Xu et al., 2025), and SuperChem (Zhao et al., 2025a) focus on reasoning trajectories and logical consistency, while SciToolBench (Ma et al., 2024) and ChemToolBench (Wu et al., 2025) evaluate tool orchestration through annotated tool invocation chains. However, most existing benchmarks still assess reasoning and tool use in isolation. They therefore fail to capture the closed-loop interaction required by future chemical agents, where tool outputs provide feedback for refining intermediate reasoning, and clearer reasoning enables more precise tool selection and execution (Wei et al., 2026; Chai et al., 2025). As a result, current evaluations remain insufficient for measuring whether models can reliably and adaptively switch between internal deliberation and external execution in complex chemical problem solving.

To address this gap, we introduce **ChemEval2**, a diagnostic benchmark for tool-augmented chemical reasoning. ChemEval2 contains 2,410 high-

[1] https://anonymous.4open.science/r/ChemEval2-F211

quality instances across three dimensions: *Knowledge Reasoning*, *Reaction Reasoning*, and *Tool-Integrated Reasoning*. Its key design is a *tool-augmented Chain-of-Thought (CoT) checkpoint* mechanism, which evaluates not only final answers but also critical intermediate reasoning steps and tool-use decisions. By inserting tool checkpoints at stages where heuristic language reasoning becomes unreliable, ChemEval2 examines whether models can recognize their cognitive boundaries, invoke chemical tools, and integrate tool outputs back into the reasoning process. The benchmark also incorporates multimodal scenarios involving spectra and molecular structures, enabling fine-grained diagnosis of agentic chemical intelligence. These three dimensions form a progressive hierarchy: knowledge reasoning and reaction reasoning evaluate the foundational capabilities required for tool-integrated reasoning, while TIR serves as the central diagnostic target of ChemEval2.

Empirically, our evaluation reveals a pronounced reasoning–tool gap. Although current models possess substantial chemical knowledge, they often struggle to identify when tool assistance is necessary and to maintain consistency between tool outputs and long-chain reasoning. We further observe a process–outcome inconsistency: models may satisfy intermediate checkpoints but fail to reach correct final answers, especially in complex multimodal settings. These findings indicate that the major bottleneck of chemical AI is shifting from knowledge acquisition to the fine-grained orchestration of reasoning and tool execution.

The main contributions of this work are:

- **A diagnostic benchmark for tool-augmented chemical reasoning**. We introduce ChemEval2, containing 2,410 high-quality instances across knowledge reasoning, reaction reasoning, and tool-integrated reasoning, advancing evaluation beyond static chemical question answering.

- **A checkpoint-based evaluation mechanism**. We propose tool-augmented CoT checkpoints to assess whether models can identify reasoning bottlenecks, strategically invoke specialized tools, and integrate tool outputs into intermediate reasoning and final answers.

- **Standardized tools and fine-grained failure analysis**. We release an MCP-based suite of high-frequency chemical tools and provide a systematic diagnosis of current LLMs, revealing key limitations in multi-step reaction reasoning and tool-assisted reasoning.

2 Related Work

2.1 Large Language Models in Chemistry

Large language models (LLMs) have shown strong potential in scientific research (Chang et al., 2024; Zhang et al., 2026; Zhao et al., 2023). General-purpose models such as GPT-4o (Hurst et al., 2024) and Qwen (Yang et al., 2025a) demonstrate broad scientific capabilities but still face challenges in chemically constrained reasoning. To address this issue, domain-specific models such as ChemDFM (Zhao et al., 2024), LlaSMol (Yu et al., 2024), and ChemLLM (Zhang et al., 2024b) incorporate chemical structures and molecular knowledge to improve specialized understanding (Li et al., 2025b; Zhang et al., 2024a; Taylor et al., 2022). More recently, reasoning-oriented models such as ChemDFM-R (Zhao et al., 2025b) and RETRO-R1 (Liu et al.) further improve reaction prediction and synthesis planning through explicit long-chain reasoning (Narayanan et al., 2025; Zhang et al., 2025a; Ferrag et al., 2025). However, parametric knowledge alone remains insufficient for complex scientific decision making, motivating the development of tool-augmented chemical agents.

2.2 Chemical Agents and Tool Integration

Tool-augmented agents extend chemical LLMs by coupling model reasoning with external tools and computational environments (Ramos et al., 2025; Kim et al., 2025; Yu et al., 2025). Representative systems include ChemCrow (M. Bran et al., 2024), SciToolAgent (Ding et al., 2025), RETRO-R1 (Liu et al.), and CheMatAgent (Wu et al., 2025). These studies show that tool integration can mitigate hallucinations and expand the capability boundaries of LLMs (Tang et al., 2025; Li et al., 2025c). However, existing evaluations focus on final performance or tool invocation accuracy, with limited diagnosis of how tools interact with reasoning.

2.3 Evaluation of LLMs in Chemistry

Evaluation benchmarks are essential for measuring LLM progress. In chemistry, benchmarks such as ChemLLMbench (Guo et al., 2023), ChemEval (Huang et al., 2025), ChemBench (Mirza et al., 2025), and MaCBench (Alampara et al., 2025) evaluate chemical knowledge and multimodal understanding, while ChemCoTBench (Li et al.,

Table 1: Comparison between ChemEval2 and existing chemical LLM benchmarks.

Benchmark	QA	Reasoning	Tools	TIR
ChemLLMbench	✓	×	×	×
ChemBench	✓	×	×	×
MaCBench	✓	×	×	×
ChemCoTBench	✓	✓	×	×
oMeBench	✓	✓	×	×
SuperChem	✓	✓	×	×
SciToolBench	✓	×	✓	×
ChemToolBench	✓	×	✓	×
ChemEval	✓	×	×	×
ChemEval2 (Ours)	✓	✓	✓	✓

2025a), oMeBench (Xu et al., 2025), and SuperChem (Zhao et al., 2025a) focus on reasoning trajectories. Other benchmarks, including SciToolBench (Ma et al., 2024) and ChemToolBench (Wu et al., 2025), evaluate tool orchestration. Table 1 summarizes the capability coverage of ChemEval2 and representative existing chemical LLM benchmarks. However, existing benchmarks largely assess reasoning and tool use separately, leaving tool-integrated chemical reasoning underexplored. Building upon ChemEval, ChemEval2 addresses this gap through a unified diagnostic framework spanning knowledge reasoning, reaction reasoning, and tool-integrated reasoning (TIR). An expanded comparison with existing chemical reasoning and tool-use benchmarks is provided in Appendix D.

3 ChemEval2

ChemEval2 is designed to evaluate tool-augmented chemical reasoning, where models must coordinate internal reasoning with external tool use. Different from conventional benchmarks that focus on final-answer correctness or isolated tool invocation, ChemEval2 examines whether models can identify reasoning boundaries, invoke appropriate chemical tools, and integrate tool outputs into multi-step scientific reasoning.

As shown in Figure 1, ChemEval2 contains 2,410 high-quality samples reconstructed from professional textbooks and authentic experimental records. The benchmark spans inorganic, organic, physical, analytical, and experimental chemistry, and covers three task dimensions: **Knowledge Reasoning**, **Reaction Reasoning**, and **Tool-Integrated Reasoning**. It also includes multimodal inputs such as spectra, molecular structures, charts, and experimental diagrams, reflecting practical chemical problem-solving scenarios.

3.1 Task Design

Knowledge Reasoning evaluates models' understanding of core fundamental chemical concepts and their ability to perform quantitative scientific inference from textual, symbolic, structural, and visual information. The task includes multiple-choice, fill-in-the-blank, computational, short-answer, and chart-analysis questions.

Reaction Reasoning focuses on dynamic chemical process understanding, including product prediction, condition recommendation, mechanistic analysis, and retrosynthetic feasibility assessment. Models are required to reason over reactants, products, catalysts, solvents, conditions, and plausible multi-step reaction pathways with transformations.

Tool-Integrated Reasoning evaluates whether models can strategically use external tools when internal reasoning becomes unreliable, such as in molecular property calculation, formula analysis, spectral interpretation, and reaction prediction. This task measures not only answer correctness, but also the ability to identify information gaps, execute appropriate tool calls, and incorporate tool outputs into coherent reasoning. Each ChemEval2 instance is represented by a structured schema consisting of a question, optional visual context, final answer, expert-annotated checkpoints, and tool traces. The formal instance schema and annotation protocol are provided in Appendix A.

3.2 Evaluation Methodology

ChemEval2 adopts a checkpoint-based evaluation framework to assess intermediate reasoning and final-answer correctness. For each problem, expert annotators define a set of weighted checkpoints:

$$C = \{c_1, c_2, \ldots, c_N\},$$

where each checkpoint represents a reasoning step or final conclusion. An LLM-based evaluator compares model predictions with reference checkpoint semantics and, for tool-related checkpoints, expert-annotated tool-use instructions.

Overall Accuracy (OA) measures the weighted proportion of checkpoints reached by the model:

$$OA = \frac{\sum_{c_i \in C} s_i \cdot w_i}{\sum_{c_i \in C} w_i}.$$

For tool-integrated tasks, a checkpoint is satisfied if the model either matches the required reasoning semantics or correctly executes the corresponding

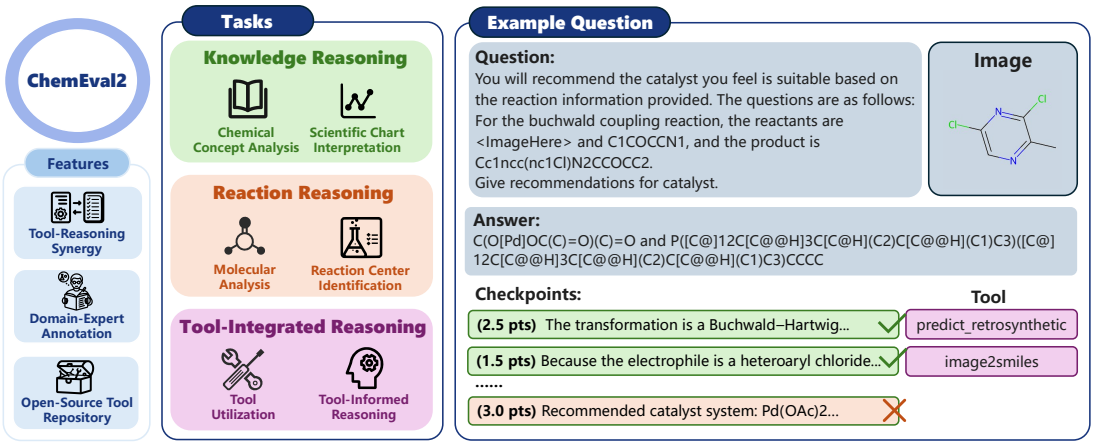

Figure 1: Overview of ChemEval2. The benchmark integrates three progressive reasoning dimensions with checkpoint-based evaluation for diagnosing chemical reasoning, multimodal understanding, and strategic tool use.

tool call. We define:

$$a_i = \text{Match}_{sem}(t_i, T_i), \quad b_i = \text{Match}_{tool}(m_i, M_i),$$

$$s_i = \mathbb{I}(a_i = 1 \vee b_i = 1),$$

where t_i and m_i denote the model-generated reasoning and tool call, T_i and M_i denote the expert-annotated reference semantics and tool instruction, and a_i and b_i indicate semantic and tool-call matching results, respectively.

Final-Answer Accuracy (FA) evaluates whether the final conclusion is consistent with the ground truth. Let $C_{\text{final}} \subseteq C$ denote the subset of final-answer checkpoints:

$$FA = \frac{\sum_{c_i \in C_{\text{final}}} s_i \cdot w_i}{\sum_{c_i \in C_{\text{final}}} w_i}.$$

To further diagnose tool-use behavior, we report two complementary metrics: **Tool Call Precision (TCP)** and **Tool Call Recall (TCR)**. TCP measures the proportion of model-generated tool calls that are required by expert-annotated checkpoints, while TCR measures the proportion of required tool-use checkpoints successfully covered by the model. We define:

$$r_j = \mathbb{I}[\exists c_i \in C_{\text{tool}}.\text{Match}(m_j, c_i)]$$

$$q_i = \mathbb{I}[\text{Match}(c_i)]\,.$$

$$TCP = \frac{\sum_{m_j \in M_{\text{pred}}} r_j}{|M_{\text{pred}}|}, \quad TCR = \frac{\sum_{c_i \in C_{\text{tool}}} q_i w_i}{\sum_{c_i \in C_{\text{tool}}} w_i}.$$

Here, M_{pred} denotes the set of model-generated tool calls, C_{tool} denotes the set of expert-annotated tool-use checkpoints, r_j indicates whether a predicted tool call matches any required valid checkpoint, and q_i indicates whether a required checkpoint is successfully covered. All evaluations are conducted under a standardized prompt setting and consistent review protocol. The full inference and automated evaluation prompts are provided in Appendix F to support fair, transparent, reproducible, and auditable evaluation.

3.3 Benchmark Construction Pipeline

The construction pipeline of ChemEval2 consists of four stages, as shown in Figure 2. First, raw materials are collected from approximately 600 chemistry textbooks and nearly 9,000 authentic experimental records from collaborative laboratories. To reduce data leakage, expert annotators perform semantic rewriting and structural reconstruction while preserving the underlying scientific logic.

Second, reconstructed samples are filtered through rigorous manual review and rule-based screening. This step removes semantically redundant items, ambiguous questions, samples without scientifically unique answers under reasonable scientific assumptions, and outdated content based on obsolete theories or revised experimental facts.

Third, experts construct tool-integrated reasoning instances through a human-in-the-loop process. Initial reasoning paths are generated and then revised by chemistry experts into gold-standard reasoning chains with logical checkpoints and scoring weights. For nodes where internal reasoning is insufficient or unreliable, experts annotate standardized tool-use instructions and verify them through both automated and manual checks.

Finally, ChemEval2 adopts a three-tier quality control protocol. Undergraduate annotators conduct initial checkpoint inspection, answer verification, and format correction. Graduate-level reviewers verify scientific consistency and solution correctness. Faculty experts perform final audits on

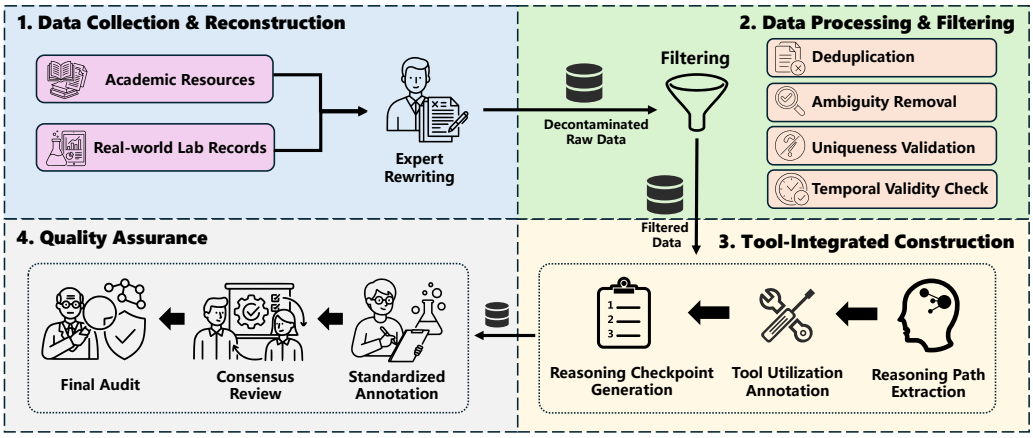

Figure 2: Construction pipeline of ChemEval2. The process includes data collection and reconstruction, data filtering, tool-integrated instance construction, and three-tier quality assurance.

high-difficulty samples and overall dataset quality.

3.4 Data Statistics

ChemEval2 contains 2,410 samples, including 897 Knowledge Reasoning samples, 541 Reaction Reasoning samples, and 972 Tool-Integrated Reasoning samples. The dataset is designed with a clear difficulty gradient, with samples categorized into middle-school, high-school, and undergraduate or research-level standards. To challenge current models, undergraduate and research-level samples constitute 54.2% of the dataset.

The benchmark also exhibits substantial multimodal and chemical diversity. Multimodal samples include charts, spectra, molecular structures, and experimental diagrams, with some highly complex cases containing up to 12 images. The reaction-related samples cover a broad range of structural transformations, from fundamental reactions to complex coupling and cyclization processes. We further use the Synthetic Accessibility score to quantify molecular complexity, revealing a long-tail distribution with scores reaching 9.0. These statistics indicate that ChemEval2 covers both basic chemical reasoning and challenging realistic scenarios involving complex molecules, multimodal evidence, and tool-augmented reasoning. Details are provided in Appendix B.

4 Experiments

We use ChemEval2 to evaluate LLMs across chemical reasoning and tool integration. We first introduce the experimental setup, then analyze model performance on knowledge, reaction, and tool-integrated reasoning, followed by comparisons between specialized and general models.

4.1 Setup

We evaluate a diverse set of mainstream LLMs covering different training paradigms, openness levels, and modality capabilities, including **GPT-5.2** (Singh et al., 2025), **o4-mini** (OpenAI, 2025), **Gemini 3 Pro** (Google, 2025), **Claude Opus 4.5** (Anthropic, 2025), **LLaMA-3.1-8B** (Grattafiori et al., 2024), **DeepSeek-V3.2** (Liu et al., 2025), **DeepSeek-R1** (Guo et al., 2025), **Qwen3-235B-A22B** (Yang et al., 2025a), **Qwen3-VL-235B-A22B**, **ChemLLM** (Zhang et al., 2024b), **ChemVLM** (Li et al., 2025b), **ChemDFM-v1.5** (Zhao et al., 2025c), and **Intern-S1-mini** (Bai et al., 2025). For comparison, we also evaluate the corresponding base models: **InternLM2-chat-7B** (Cai et al., 2024), **InternLM2.5-7b-chat**, **LLaMA-3-8B**, and **Qwen3-8B**. Detailed model descriptions, deployment settings, decoding parameters, and evaluator configurations are provided in Appendix C.

4.2 Main Results

Tables 2 and 3 summarize model performance on text-only and multimodal tasks in ChemEval2. Overall, we observe a clear and consistent performance stratification across model families. Frontier proprietary models consistently outperform open-source and chemistry-specialized models, especially on tasks requiring multi-step reasoning and multimodal interpretation. Among them, Gemini 3 Pro, Claude Opus 4.5, and GPT-5.2 achieve the strongest overall robust performance.

A major finding is that reaction reasoning is substantially more challenging than knowledge reasoning. Nearly all models show a clear performance drop on reaction tasks, and this gap becomes more pronounced in multimodal settings involving

Table 2: Performance overview of text tasks on ChemEval2.

Model	Knowledge		Reaction		Tool-Integrated				Overall	
	OA	FA	OA	FA	OA	FA	TCP	TCR	OA	FA
Closed-source Models										
Claude Opus 4.5	81.59±0.42	77.98±0.55	74.52±0.61	64.16±0.82	77.20±0.48	74.22±0.73	15.96±0.68	8.14±0.47	77.78±0.39	72.60±0.51
Gemini 3 Pro	77.94±0.35	77.32±0.41	74.47±0.38	67.05±0.59	**80.65±0.32**	**79.19±0.45**	21.53±0.75	9.22±0.49	**78.16±0.28**	**75.32±0.36**
GPT-5.2	74.74±0.62	65.96±0.75	73.69±0.58	56.65±0.91	69.56±0.66	67.59±0.82	**54.71±1.46**	5.26±0.35	72.23±0.45	64.12±0.68
o4-mini	72.09±0.58	65.30±0.81	62.75±0.72	52.31±1.05	57.13±0.95	65.50±0.88	20.45±0.71	0.51±0.12	63.13±0.61	61.85±0.74
Open-source Models										
Qwen3-235B-A22B	70.89±0.51	71.88±0.63	70.52±0.55	58.67±0.89	69.59±0.71	66.39±0.85	7.14±0.41	0.23±0.08	70.23±0.48	65.93±0.62
Qwen3-8B	64.74±0.75	66.58±0.82	51.18±1.12	51.45±1.05	56.49±0.92	57.09±0.98	0.00±0.00	0.00±0.00	57.51±0.68	58.39±0.77
InternLM2.5-7b-chat	34.21±1.25	33.78±1.18	28.78±1.35	32.37±1.42	8.96±0.85	10.31±0.91	0.00±0.00	0.00±0.00	21.89±0.95	23.32±0.88
InternLM2-chat-7b	19.79±0.95	16.80±0.88	20.58±1.02	23.70±1.15	4.82±0.55	4.97±0.61	28.31±1.04	3.51±0.31	13.58±0.72	13.60±0.68
LLaMA-3.1-8B	25.86±0.85	23.25±0.92	27.96±1.10	32.95±1.25	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	15.33±0.65	15.92±0.71
LLaMA-3-8B	14.33±0.78	23.08±0.85	13.31±0.95	18.79±1.05	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	7.90±0.55	12.01±0.62
DeepSeek-R1	79.18±0.38	78.38±0.45	69.09±0.52	61.56±0.78	74.96±0.42	72.10±0.65	3.39±0.28	3.56±0.30	74.62±0.35	71.10±0.48
DeepSeek-V3.2	**84.33±0.31**	**83.42±0.39**	**75.03±0.42**	**69.65±0.55**	70.91±0.48	69.71±0.58	12.76±0.61	**17.08±0.74**	76.04±0.32	73.79±0.41
Specialized Models										
Intern-S1-mini	66.96±0.65	67.37±0.72	55.32±0.85	53.76±0.92	50.24±0.78	53.04±0.81	30.65±1.09	4.52±0.34	56.61±0.55	57.51±0.68
ChemVLM	32.77±1.15	29.89±1.22	29.80±1.05	32.37±1.35	8.34±0.62	11.05±0.75	36.69±1.28	9.28±0.52	21.48±0.85	22.48±0.92
ChemLLM	8.84±0.45	14.06±0.68	9.41±0.55	14.74±0.72	3.35±0.32	8.10±0.51	0.00±0.00	0.00±0.00	6.64±0.38	11.69±0.55
ChemDFM-v1.5	8.43±0.42	10.34±0.58	13.28±0.65	17.34±0.85	5.99±0.45	11.79±0.62	0.00±0.00	0.00±0.00	8.70±0.48	12.87±0.52

molecular structures, reaction schemes, spectra, or experimental apparatus. This suggests that reaction reasoning requires models to go beyond factual recall and reason over functional groups, reaction centers, electronic effects, and bond transformations under specific chemical conditions.

We also observe a consistent gap between OA and FA. In many cases, models correctly satisfy several intermediate checkpoints but fail to integrate them into a correct final answer. This process–outcome inconsistency indicates that current LLMs can often recognize local reasoning steps but remain weak at maintaining global logical coherence and propagating constraints across long reasoning chains. The phenomenon is especially evident in complex multimodal tasks, where visual interpretation, chemical reasoning, and answer synthesis must be jointly coordinated.

These results show that ChemEval2 does not merely test chemical knowledge memorization. Instead, it exposes the model's ability to maintain coherent reasoning across intermediate checkpoints, multimodal evidence, and tool-assisted steps.

4.3 Specialized Models vs. Base Models

To examine whether domain-adaptive training improves chemical reasoning, we compare chemistry-specialized models with their corresponding base models. As shown in Figure 3, domain specialization generally improves knowledge-oriented performance, especially on factual tasks, suggesting that domain-specific training helps models better recognize chemical terminology, molecular representations, and basic concepts.

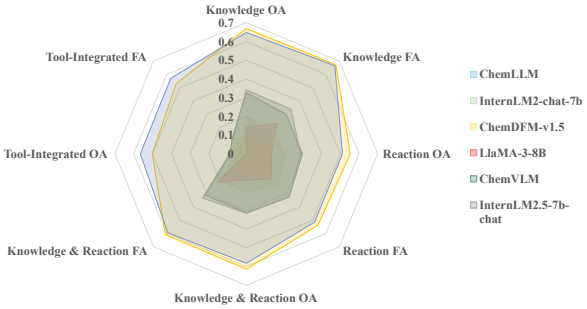

Figure 3: Comparison of performance between chemistry models and their base models on text tasks.

However, this advantage does not consistently transfer to complex reasoning tasks. In reaction reasoning, chemistry-specialized models often produce chemically plausible intermediate statements and avoid basic errors such as invalid valence assignments or non-existent molecular formulas, but their performance still lags behind frontier general-purpose models. This suggests that reaction-level reasoning depends not only on domain knowledge, but also on general reasoning ability, instruction following, and long-context integration.

The limitation is more evident in tool-integrated reasoning. Chemistry-specialized models show no consistent advantage over their base models and may even degrade in some settings. Since tool use requires identifying when assistance is needed, selecting tools, generating valid parameters, and integrating returned results, domain fine-tuning alone is insufficient to solve the orchestration problem required for tool-augmented chemical reasoning.

Table 3: Performance overview of Multimodal tasks on ChemEval2.

Model	Knowledge		Reaction		Tool-Integrated				Overall	
	OA	FA	OA	FA	OA	FA	TCP	TCR	OA	FA
Closed-source Models										
Claude Opus 4.5	79.85±0.45	75.65±0.52	48.02±0.95	21.13±0.88	64.32±0.65	53.49±0.72	0.00±0.00	0.00±0.00	68.59±0.58	58.02±0.69
Gemini 3 Pro	**85.43±0.35**	**84.92±0.38**	**55.91±0.55**	**38.66±0.62**	**77.65±0.41**	**68.45±0.48**	1.29±0.19	1.29±0.19	**77.47±0.32**	**70.84±0.45**
GPT-5.2	80.37±0.48	76.18±0.55	43.69±0.85	19.33±0.75	58.54±0.68	49.91±0.72	**30.49±1.12**	**27.23±1.05**	65.92±0.52	56.62±0.65
o4-mini	76.22±0.52	75.17±0.58	31.92±0.92	15.46±0.85	50.36±0.75	48.10±0.78	22.81±0.86	0.84±0.15	58.96±0.62	54.82±0.68
Open-source Models										
Qwen3-VL-235B-A22B	50.90±0.85	11.87±0.65	40.40±0.95	5.15±0.35	59.01±0.72	49.91±0.78	0.00±0.00	0.00±0.00	52.14±0.68	24.97±0.75
Chemistry Models										
ChemVLM	24.39±0.88	21.79±0.92	3.82±0.45	7.61±0.55	4.98±0.52	6.81±0.58	13.75±0.64	1.43±0.21	13.60±0.65	13.75±0.68
Intern-S1-mini	69.07±0.62	66.88±0.68	31.84±0.85	16.24±0.72	34.88±0.82	33.96±0.85	2.82±0.27	0.59±0.12	49.90±0.65	45.89±0.72

4.4 Tool-Integrated Reasoning Analysis

Tool-integrated reasoning is the central diagnostic dimension of ChemEval2. Figures 4 and 5 show that effective tool use remains a major weakness across current LLMs. Most models achieve low tool-call scores, particularly in tasks requiring autonomous planning, precise API parameter generation, and multi-step tool execution. This indicates that current models still struggle to translate internal reasoning needs into executable tool actions.

Nevertheless, successful tool use can substantially improve final-answer correctness. In cases where models correctly invoke tools and obtain valid outputs, FA is generally higher than in cases without effective tool use. This demonstrates the value of external tools for correcting uncertain parametric knowledge, supporting precise computation, and improving reaction prediction. However, tool outputs do not automatically improve the reasoning process. Some models obtain useful tool results but fail to integrate them coherently, leading to correct intermediate operations but incorrect final conclusions. This again reflects the process–outcome inconsistency revealed by ChemEval2.

To further analyze tool behavior, we decompose tool use into TCP and TCR. The results reveal different failure modes: some models rarely invoke tools when required, whereas others attempt tool calls but generate chemically invalid or irrelevant parameters. Thus, low tool-use performance reflects not only optional tool skipping, but also weaknesses in tool interaction and execution.

We also conduct a four-quadrant analysis by crossing tool correctness with final-answer correctness. The results reveal three representative patterns. First, SOTA collaborative models, such as GPT-5.2, Claude, DeepSeek-V3.2, and Gemini 3 Pro, frequently benefit from tool interaction in complex tasks, suggesting that tool use is important for reliably approaching the performance ceiling. Second, reasoning-centric models such as DeepSeek-R1 and o4-mini often solve some tasks through internal reasoning without matching the expected tool path, confirming that ChemEval2 does not enforce a single expected tool-dependent solution. Third, smaller or weaker models perform poorly in both tool-supported and tool-free pathways, indicating simultaneous deficiencies in chemical reasoning and instruction following.

These findings show that ChemEval2 can distinguish between tool-dependent reasoning, internal-knowledge reasoning, and general capability failure. Therefore, the benchmark does not reward static tool matching; instead, it provides a diagnostic view of how models select, execute, and integrate tools during chemical problem solving. Additional TCP/TCR evidence and four-quadrant distributions are reported in Appendix E.

5 Discussion

We first examine the reliability of ChemEval2 before discussing its broader implications, since its conclusions depend on statistically robust model comparisons and consistent checkpoint-based scoring of reasoning and tool-use behavior.

5.1 Evaluation Reliability and Statistical Robustness

We conduct two complementary analyses to validate ChemEval2. First, paired Wilcoxon signed-rank tests on representative model comparisons show that the observed performance differences are not artifacts of overlapping error bars. Selected results are reported in Appendix E.1. Frontier models significantly outperform smaller base models, confirming the effect of model scale and

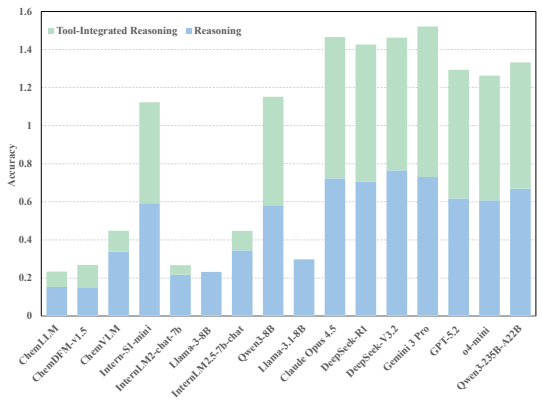

Figure 4: Impact of tool call on model reasoning performance in text tasks.

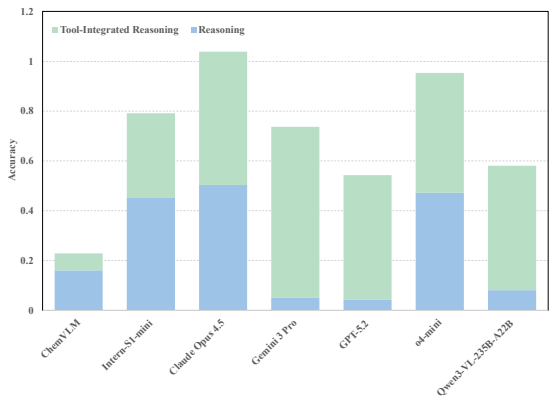

Figure 5: Impact of tool call on model reasoning performance in multimodal tasks.

general reasoning capability, whereas chemistry-specialized models do not consistently outperform their base models in tool-integrated reasoning. Model pairs with nearly identical scores show non-significant differences, indicating that the test serves as a robustness check rather than amplifying small numerical gaps.

Second, we assess whether automated checkpoint scoring aligns with expert judgment. On a randomly sampled validation set covering text-only and multimodal tasks, 240 instances and 1,424 atomic checkpoints are independently reviewed by chemistry experts. The automated evaluator achieves 93.12% agreement with human judgment, with Cohen's Kappa of 0.8107 and Pearson and Spearman correlations of 0.8108. These results suggest that the evaluator captures both checkpoint decisions and expert scoring trends.

Together, these analyses support ChemEval2 as a stable and reliable evaluation protocol. Importantly, the checkpoint design does not require models to reproduce a single rigid reasoning chain. It focuses on essential reasoning nodes and allows semantically correct internal reasoning to satisfy a checkpoint even when the expected tool is not called, thereby reducing bias against alternative valid paths while preserving tool-use diagnosis.

5.2 Standardized Tool Suite

To provide a reproducible environment for evaluating tool-augmented chemical agents, ChemEval2 constructs and releases an MCP-based tool suite covering high-frequency chemical operations, including molecular perception, representation conversion, property calculation, composition analysis, functional group and stereochemistry recognition, forward reaction prediction, and retrosynthetic reasoning. The MCP interface standardizes tool invocation, reducing implementation variability and enabling fair comparison across models. More importantly, it reflects a core requirement of autonomous scientific agents: models must translate reasoning needs into executable actions. Thus, tool interaction and API alignment are treated as essential capabilities rather than formatting details. Our analysis further shows that correct tool execution alone is insufficient. Some models can call tools properly but fail to integrate the outputs into chemical reasoning, highlighting the need for stronger tool-result interpretation, reasoning-state updating, and long-horizon planning.

6 Conclusion

In this work, we introduced ChemEval2, a framework that improves the evaluation of chemical artificial intelligence by focusing on the integration of reasoning and tool usage. While previous benchmarks focused on checking internal reasoning steps, ChemEval2 addresses a more difficult challenge: evaluating how well models can use professional tools to improve their reasoning. Our research findings showed a clear problem where even the best models struggle to keep their internal logic consistent with the results from external tools, especially in difficult tasks. By providing the ChemEval2 benchmark and a standard set of chemical tools based on the MCP, we provide a solid foundation for building the next generation of reliable tool-augmented chemical agents. These tools will help models move beyond just knowing chemical facts to actively performing reproducible, autonomous chemical research with the accuracy and reliability needed for real scientific discovery.

Limitations

ChemEval2 has several limitations. First, although the benchmark covers a broad range of chemical reasoning, reaction understanding, and tool-integrated tasks, the overall scale and the granularity of certain specialized subfields remain constrained by the high cost of expert annotation and the limited availability of high-quality experimental data. Second, the current MCP-based tool suite focuses on ten representative high-frequency tools and does not yet encompass the full spectrum of resources required for autonomous chemical research, such as laboratory automation platforms. Third, while the benchmark includes multimodal inputs such as spectra, molecular structures, reaction schemes, and experimental diagrams, these representations capture only selected aspects of real laboratory environments and cannot fully reflect the complexity of experimental workflows and instrument interactions. Finally, the checkpoint-based evaluation framework relies on expert-defined reasoning nodes and tool-use annotations, which may not exhaust all valid solution paths despite being designed to allow semantically equivalent reasoning. Future work will expand dataset coverage, enrich the standardized tool ecosystem, and develop more flexible evaluation strategies for modeling diverse reasoning trajectories and long-horizon scientific decision making.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, NM Anoop Krishnan, and Kevin Maik Jablonka. 2025. Probing the limitations of multimodal language models for chemistry and materials research. *Nature computational science*, 5(10):952–961.

Anthropic. 2025. Claude opus 4.5 system card. Technical report, Anthropic PBC.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, and 1 others. 2025. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*.

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

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and 1 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Jingyi Chai, Shuo Tang, Rui Ye, Yuwen Du, Xinyu Zhu, Mengcheng Zhou, Yanfeng Wang, Yuzhi Zhang, Linfeng Zhang, Siheng Chen, and 1 others. 2025. Scimaster: Towards general-purpose scientific ai agents, part i. x-master as foundation: Can we lead on humanity's last exam? *arXiv preprint arXiv:2507.05241*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Keyan Ding, Jing Yu, Junjie Huang, Yuchen Yang, Qiang Zhang, and Huajun Chen. 2025. Scitoolagent: a knowledge-graph-driven scientific agent for multitool integration. *Nature Computational Science*, 5(10):962–972.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. 2025. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*.

Google. 2025. Gemini 3 pro model card. Technical report, Google DeepMind.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and 1 others. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Ming Hu, Chenglong Ma, Wei Li, Wanghan Xu, Jiamin Wu, Jucheng Hu, Tianbin Li, Guohang Zhuang, Jiaqi Liu, Yingzhou Lu, and 1 others. 2025. A survey of scientific large language models: From data foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*.

Yuqing Huang, Rongyang Zhang, Xuesong He, Xuyang Zhi, Hao Wang, Nuo Chen, Xin Li, Feiyang Xu, Deguang Liu, Huadong Liang, and 1 others. 2025.

Chemeval: A multi-level and fine-grained chemical capability evaluation for large language models.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Hyomin Kim, Yunhui Jang, and Sungsoo Ahn. 2025. Mt-mol: Multi agent system with tool-based reasoning for molecular optimization. *arXiv preprint arXiv:2505.20820*.

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

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, and 1 others. 2025b. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 415–423.

Zhucong Li, Jin Xiao, Bowei Zhang, Zhijian Zhou, Qianyu He, Fenglei Cao, Jiaqing Liang, and Yuan Qi. 2025c. Chemhts: Hierarchical tool stacking for enhancing chemical agents. *arXiv preprint arXiv:2502.14327*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Wei Liu, Jiangtao Feng, Hongli Yu, Yuxuan Song, Yuqiang Li, Shufei Zhang, Lei Bai, Wei-Ying Ma, and Hao Zhou. Retro-r1: Llm-based agentic retrosynthesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, and 1 others. 2024. Sciagent: Tool-augmented language models for scientific reasoning. *arXiv preprint arXiv:2402.11451*.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, and 1 others. 2025. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, pages 1–8.

Siddharth M Narayanan, James D Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Samuel G Rodriques, and Andrew D White. 2025. Training a scientific reasoning model for chemistry. *arXiv preprint arXiv:2506.17238*.

OpenAI. 2025. Openai o3 and o4-mini system card.

Mayk Caldas Ramos, Christopher J Collison, and Andrew D White. 2025. A review of large language models and autonomous agents in chemistry. *Chemical science*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, and 1 others. 2025. Chemagent: Self-updating memories in large language models improves chemical reasoning. In *The Thirteenth International Conference on Learning Representations*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Tianxin Wei, Ting-Wei Li, Zhining Liu, Xuying Ning, Ze Yang, Jiaru Zou, Zhichen Zeng, Ruizhong Qiu, Xiao Lin, Dongqi Fu, and 1 others. 2026. Agentic reasoning for large language models. *arXiv preprint arXiv:2601.12538*.

Mengsong Wu, YaFei Wang, Yidong Ming, Yuqi An, Yuwei Wan, Wenliang Chen, Binbin Lin, Yuqiang Li, Tong Xie, and Dongzhan Zhou. 2025. Chemagent: Enhancing llms for chemistry and materials science through tree-search based tool learning. *arXiv preprint arXiv:2506.07551*.

Ruiling Xu, Yifan Zhang, Qingyun Wang, Carl Edwards, and Heng Ji. 2025. omebench: Towards robust benchmarking of llms in organic mechanism elucidation and reasoning. *arXiv preprint arXiv:2510.07731*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Cheng Yang, Jiaxuan Lu, Haiyuan Wan, Junchi Yu, and Feiwei Qin. 2025b. From what to why: A multi-agent system for evidence-based chemical reaction condition reasoning. *arXiv preprint arXiv:2509.23768*.

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

Botao Yu, Frazier N Baker, Ziru Chen, Garrett Herb, Boyu Gou, Daniel Adu-Ampratwum, Xia Ning, and Huan Sun. 2025. Tooling or not tooling? the impact of tools on language agents for chemistry problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7620–7640.

Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024a. Sciinstruct: a self-reflective instruction annotated dataset for training scientific language models. *Advances in Neural Information Processing Systems*, 37:1443–1473.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, and 1 others. 2024b. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and 1 others. 2026. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36.

Situo Zhang, Hanqi Li, Lu Chen, Zihan Zhao, Xuanze Lin, Zichen Zhu, Bo Chen, Xin Chen, and Kai Yu. 2025a. Reasoning-driven retrosynthesis prediction with large language models via reinforcement learning. *arXiv preprint arXiv:2507.17448*.

Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, and 1 others. 2025b. Artificial intelligence for science in quantum, atomistic, and continuum systems. *Foundations and Trends® in Machine Learning*, 18(4):385–849.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zehua Zhao, Zhixian Huang, Junren Li, Siyu Lin, Junting Zhou, Fengqi Cao, Kun Zhou, Rui Ge, Tingting Long, Yuexiang Zhu, and 1 others. 2025a. Superchem: A multimodal reasoning benchmark in chemistry. *arXiv preprint arXiv:2512.01274*.

Zihan Zhao, Bo Chen, Ziping Wan, Lu Chen, Xuanze Lin, Shiyang Yu, Situo Zhang, Da Ma, Zichen Zhu, Danyang Zhang, and 1 others. 2025b. Chemdfm-r: A chemical reasoning llm enhanced with atomized chemical knowledge. *arXiv preprint arXiv:2507.21990*.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, and 1 others. 2025c. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4).

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, and 1 others. 2024. Chemdfm: A large language foundation model for chemistry. In *Neurips 2024 Workshop Foundation Models for Science: Progress, Opportunities, and Challenges*.

A Annotation Schema and Evaluation Protocol

This section clarifies the annotation schema and evaluation protocol of ChemEval2. We first define the structure of a benchmark instance, then describe how reasoning checkpoints and tool-use annotations are designed to support flexible yet diagnostic evaluation.

A.1 Formal Definition of an Instance

A ChemEval2 instance is formally defined as a diagnostically structured five-tuple:

$$I = (Q, V, A, C, T),$$

where the components are defined as follows:

- Q **(Question):** a task description formulated in natural language, covering open-ended, context-dependent chemical problems such as reaction prediction, mechanism analysis, condition recommendation, property prediction, or quantitative reasoning.
- V **(Visual Context):** a set of task-relevant, multi-source visual inputs related to the question, such as molecular structures, spectra, reaction schemes, charts, or experimental diagrams. This field is empty for text-only tasks.
- A **(Final Answer):** a structured, machine-readable reference answer, typically represented in normalized forms such as SMILES strings, catalyst combinations, numerical values, or concise textual conclusions.
- C **(Checkpoints):** a set of expert-annotated reasoning nodes. Each checkpoint contains a semantic description and a score weight, representing a critical intermediate reasoning step or final conclusion.

- T **(Tooltrace):** annotated tool-call trajectories for tool-supported reasoning steps, including the associated checkpoint IDs, tool names, and standardized input parameters.

This schema makes the benchmark output more than a simple question-answer pair: each instance explicitly encodes both the reasoning structure and the tool-interaction structure required for fine-grained diagnostic evaluation.

A.2 Checkpoint Design and Valid Reasoning Paths

ChemEval2 does not require a model to reproduce a single complete reasoning chain. During annotation, experts identify decisive logical nodes that are shared by valid solution paths, rather than enforcing every intermediate statement in one reference derivation. As a result, a model can satisfy a checkpoint through semantically equivalent reasoning even when its wording or local derivation differs from the reference.

This design is important for chemistry, where the same conclusion may be reached through mechanistic reasoning, retrosynthetic analysis, structural analogy, or direct application of internal knowledge. To avoid unfairly penalizing reasonable step-skipping, checkpoints are weighted according to their importance. The evaluation therefore focuses on unavoidable key information and core chemical constraints, while allowing alternative valid trajectories and logical shortcuts.

A.3 Tool-Integrated Annotation Protocol

In tool-integrated construction, "replacing a reasoning node" means that experts identify a chemical subtask for which internal language reasoning is unreliable or insufficient and annotate it as a standardized executable tool-call node. Examples include molecular structure recognition, molecular representation conversion, physicochemical property calculation, database lookup, forward reaction prediction, and retrosynthetic analysis.

Formally, for a checkpoint c_i that corresponds to a tool-supported subtask, the reasoning node is associated with a tool-call specification:

$$c_i^{\text{reasoning}} \rightarrow c_i^{\text{tool}},$$

where c_i^{tool} specifies the tool name and structured input parameters. This substitution makes the reasoning process verifiable and reproducible while preserving the semantic role of the original checkpoint during evaluation.

The standardized tool suite was constructed by surveying high-frequency chemical operations and selecting a non-redundant core set for diagnostic evaluation. The suite includes domain-specific tools for reaction and structure analysis as well as widely used scientific resources such as PubChem-based retrieval and RDKit-based molecular processing. These resources are encapsulated into an MCP-based interface to ensure consistent invocation, comparable evaluation, and reproducible tool-augmented inference.

The suite is intended to cover representative tool-use behaviors rather than the full infrastructure of autonomous laboratories. Thus, it supports controlled evaluation of tool planning, parameter generation, and result integration, while leaving broader resources such as laboratory automation platforms for future expansion.

A.4 Checkpoint and Tooltrace Statistics

Each instance contains approximately six structured checkpoints on average, indicating that the benchmark emphasizes multi-step reasoning rather than simple question answering. In tool-integrated tasks, each instance is associated with more than three tool calls on average, with a maximum of eleven tool calls, enabling evaluation of sequential multi-step tool interaction.

B Dataset Statistics and Coverage

ChemEval2 contains 2,410 high-quality samples designed to evaluate chemical reasoning and tool-integrated problem solving. The dataset covers diverse task types, modality compositions, difficulty levels, chemical reaction categories, molecular complexity profiles, visual contexts, and tool-call requirements. These design choices allow ChemEval2 to approximate real chemical research scenarios while addressing the limited coverage of existing chemical benchmarks.

B.1 Task and Modality Composition

ChemEval2 is structured around three core task categories comprising 897 samples for Knowledge Reasoning, 541 samples for Reaction Reasoning, and 972 samples for Tool-Integrated Reasoning. Across these tasks, we adopt a hierarchical modality design as illustrated in Figure 6. Even for high-level tasks such as reaction reasoning and tool-

ChemEval2 Instance with Tool-Trace Annotations


```
{
  "instance_id": "reaction_TIR_042",
  "question": "You will recommend the catalyst you feel is suitable based on the reaction
     information provided. The questions are as follows: For the buchwald coupling
    reaction, the reactants are <ImageHere> and C1COCCN1, and the product is Cc1ncc(nc1Cl
    )N2CCOCC2. Give recommendations for catalyst.",
  "images": ["reaction_images/reactant_1.png"],
  "reference_answer": "C(O[Pd]OC(C)=O)(C)=O and P([C@]12C[C@@H]3C[C@H](C2)C[C@@H](C1)C3)
    ([C@]12C[C@@H]3C[C@@H](C2)C[C@@H](C1)C3)CCCC",
  "checkpoints": [
    {
      "id": "cp_1",
      "content": "The transformation is a Buchwald--Hartwig cross-coupling reaction,
    requiring retrosynthetic analysis to determine intermediate viability.",
      "score": 2.5,
      "tool_call": {
        "tool_name": "predict_retrosynthetic",
        "arguments": {
          "product_smiles": "Cc1ncc(nc1Cl)N2CCOCC2"
        }
      }
    },
    {
      "id": "cp_2",
      "content": "Because the electrophile is a heteroaryl chloride, we must parse the
    structural features from the image to evaluate steric hindrance.",
      "score": 1.5,
      "tool_call": {
        "tool_name": "image2smiles",
        "arguments": {
          "image_path": "reaction_images/reactant_1.png"
        }
      }
    },
    {
      "id": "cp_final",
      "content": "Recommended catalyst system: Pd(OAc)2 as the palladium source and a
    bulky phosphine ligand.",
      "score": 3.0,
      "tool_call": null
    }
  ]
}
```


integrated reasoning, the dataset preserves substantial multimodal depth. In particular, at the high difficulty level, the number of multimodal samples reaches 519, which significantly exceeds the counts at the low and middle difficulty levels.

To facilitate a granular diagnosis of model performance, each primary task category is further stratified into diverse sub-task types as delineated in Figures 8, 9, 10. The Knowledge Reasoning task encompasses fundamental formats including multiple-choice questions, chart analysis tasks, fill-in-the-blank questions, and computational problems. The Reaction Reasoning task focuses on complex generative and discriminative challenges including reaction prediction, condition recommendation, and substrate recommendation. The Tool-Integrated Reasoning task synthesizes these formats to mirror the complexity of real-world problem-solving. This extensive typological diversity is engineered to rigorously evaluate the generalization of model reasoning capabilities across authentic chemical research workflows, ensuring that models are tested against specific cognitive demands rather than mere rote memorization.

B.2 Difficulty Stratification and Domain Coverage

To effectively differentiate the capability boundaries of diverse models, ChemEval2 explicitly

ChemEval2 Instance Schema (Formal Definition)

```
{
  "id": "string",
  "question": "string",
  "visual_context": {
    "images": ["string (image_path)"]
  },
  "final_answer": "string",
  "checkpoints": [
    {
      "id": "integer",
      "content": "string (natural language reasoning step)",
      "score": "float"
    }
  ],
  "tooltrace": [
    {
      "checkpoints_id": "integer",
      "tool_name": "string",
      "input": "object (structured parameters)"
    }
  ]
}
```

Figure 6: Distribution of task categories and modality compositions in ChemEval2.

Figure 7: Distribution of difficulty levels and task categories in ChemEval2.

stratifies problem difficulty by educational stage into a low difficulty level corresponding to middle school curriculum, a middle difficulty level aligned with high school standards, and a high difficulty level representing undergraduate and research contexts, as illustrated in Figure 7. Statistical analysis demonstrates that problems at the high difficulty level dominate the dataset by accounting for approximately 54.2% of all samples. This emphasis on advanced complexity is designed to surpass the performance ceiling of existing benchmarks and to rigorously stress-test models on expert-level tasks such as advanced chemical principle analysis, synthesis planning, and spectral interpretation.

Regarding domain coverage, the dataset maintains a broadly balanced distribution across major subfields comprising organic chemistry, inorganic chemistry, physical chemistry, spectroscopy-oriented analytical chemistry, and experimental chemistry. The substantial proportion of organic and physical chemistry reflects the critical demands of AI for Science applications, particularly within high-impact, computationally driven contexts of molecular drug discovery, cross-domain materials computation, and mechanistic analysis.

B.3 Chemical Reaction and Molecular Complexity

The reliability of the evaluation depends critically on the diversity of knowledge points. As illustrated in Figure 11, ChemEval2 encompasses knowledge points comprising 33.32% for general chemical knowledge and 21.29% for redox reactions. Furthermore, the benchmark emphasizes knowledge points central to synthetic chemistry by including

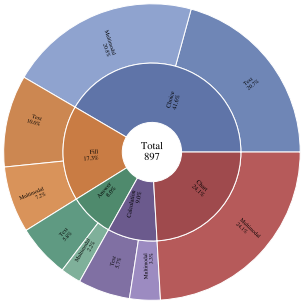

Figure 8: Distribution of task types and modality compositions in Knowledge Reasoning.

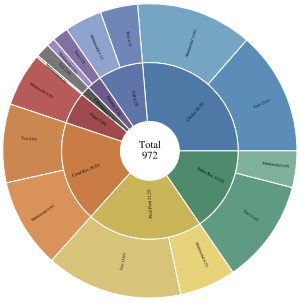

Figure 10: Distribution of task types and modality compositions in Tool-Integrated Reasoning.

Figure 9: Distribution of task types and modality compositions in Reaction Reasoning.

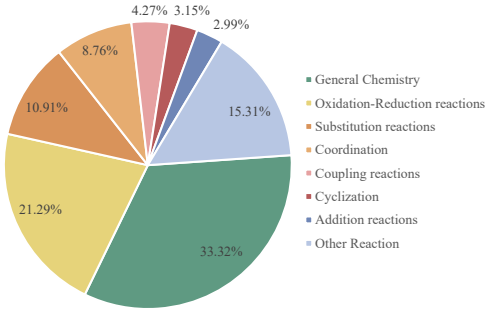

Figure 11: Distribution of chemical knowledge points in ChemEval2.

substitution reactions at 10.91%, acid–base and coordination reactions at 8.76%, coupling reactions at 4.27%, and cyclization reactions at 3.15%.

Additionally, the "Other" category constitutes 15.31% of the dataset and encompasses complex mechanisms including elimination, rearrangement, hydrolysis, and protecting group chemistry. This long-tail distribution ensures that evaluation extends beyond the mere memorization of common reaction templates to test the capacity of models to generalize across advanced mechanisms involving electron transfer and skeletal rearrangements.

Beyond reaction diversity, molecular complexity is another important dimension of chemical difficulty. To measure the molecular diversity and difficulty within ChemEval2, we extracted 200 chemical molecules from the dataset to analyze their Synthetic Accessibility score (SAScore) as illustrated in Figure 12. The distribution of scores follows an approximately normal pattern peaking within the range of 2.0 to 4.0 while exhibiting a pronounced long tail extending into the high-difficulty interval between 6.0 and 9.0.

This distribution confirms that ChemEval2 encompasses a diverse spectrum ranging from synthetically accessible molecules to structurally complex compounds such as natural products and challenging ligands. By spanning a realistic gradient from simple to highly complex molecular architectures, the dataset facilitates a systematic evaluation of model adaptability under conditions of varying practical synthesis difficulty.

B.4 Visual Context and Tool-Call Complexity

Regarding visual context in multimodal tasks, while single-image inputs constitute the majority, the dataset demonstrates a distinct long-tail distribution in visual complexity. As illustrated in Figure 13, a subset of tasks at the high difficulty level, particularly within reaction reasoning and tool-integrated reasoning, necessitates multi-image inputs ranging from 2 to 12 images.

These multi-image context scenarios impose stringent demands on the visual memory and information integration capabilities of models. For instance, analyzing a complex total synthesis pathway requires a model to jointly reason over starting materials, intermediate spectral data, and experimental apparatus diagrams. This configuration transcends conventional single-image question answering to accurately mirror the rigorous demands of real-world chemical reasoning.

The multimodal component of ChemEval2 is designed to reflect the information extraction and

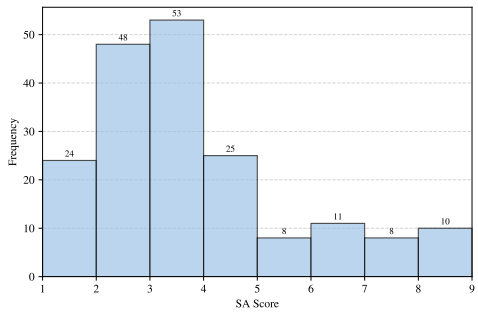

Figure 12: Distribution of SAScore in ChemEval2.

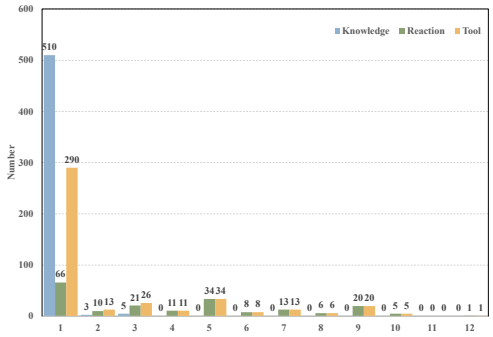

Figure 13: Distribution of image counts per sample in ChemEval2.

decision-making processes used by experts when solving chemical problems. Domain experts annotate key visual evidence from molecular structures, spectra, reaction schemes, charts, and experimental diagrams, and convert these focal points into reasoning checkpoints. This design captures but essential aspects of real laboratory interpretation, even though it cannot reproduce all details of experimental workflows and instrument interactions.

In addition to visual complexity, we analyze tool-call frequency to characterize the external-execution demands of ChemEval2. Each instance contains approximately six reasoning checkpoints on average, while tool-integrated tasks involve more than three tool calls per instance on average, with a maximum of eleven calls. This design enables ChemEval2 to evaluate multi-step tool interaction rather than isolated tool invocation. As illustrated in Figure 14, the data are categorized into pure text tasks and multimodal tasks. Regarding text-based tasks, the vast majority of samples require between 1 and 5 tool calls, with 2 calls being the most prevalent. This observation indicates that text tasks generally exhibit a lower dependency on external tools, suggesting that models can complete most text-based reasoning with limited external computation or knowledge retrieval.

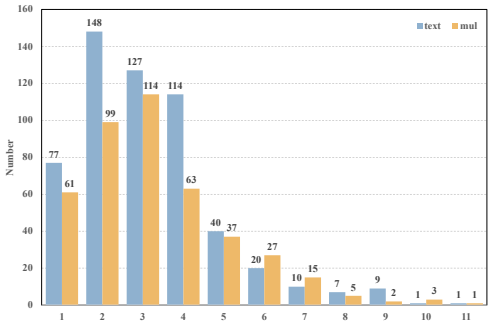

Figure 14: Distribution of tool call frequencies across text and multimodal tasks in ChemEval2.

Conversely, multimodal tasks demonstrate a significantly higher demand for tool usage characterized by a broader and more evenly distributed pattern. While the range of 2 to 4 tool calls remains dominant, a notably larger proportion of samples involve high-frequency tool calls compared to text tasks. This disparity implies that for complex visually dense scenarios such as parsing intricate reaction diagrams or integrating structured chemical information across multiple images, models must rely more extensively on external tools to support reasoning and information integration.

C Experimental Setup

C.1 Details of Evaluated Models

In this section, we provide a comprehensive introduction to the models evaluated in the **ChemEval2** benchmark. To ensure representativeness, we categorized the models into three distinct groups based on their accessibility and training objectives:

Frontier Proprietary Models. We select **GPT-5.2** (Singh et al., 2025), **o4-mini** (OpenAI, 2025), **Gemini 3 Pro** (Google, 2025), and **Claude Opus 4.5** (Anthropic, 2025) to serve as the upper-bound benchmarks for closed-source performance. These models demonstrate exceptional capabilities in handling complex scientific contexts and adhering to intricate instructions, providing a rigorous baseline for establishing the performance ceiling in chemical reasoning.

General-purpose Open-source Models. We include models from leading open-source families covering diverse architectures and scales. LLaMA-3.1-8B (Grattafiori et al., 2024) serves as a standard foundation for text-based tasks. DeepSeek-V3.2 (Liu et al., 2025) and the reasoning-optimized DeepSeek-R1 (Guo et al., 2025) represent the cutting edge of Mixture-of-Experts (MoE) and

reinforcement-learning-driven architectures. Furthermore, we utilize both the text-only Qwen3-235B-A22B (Yang et al., 2025a) and its multimodal variant Qwen3-VL-235B-A22B to investigate the scalability of vision-language integration.

Domain-specific Chemical Models. To evaluate the impact of specialized training, we assess a suite of chemistry-centric models including ChemLLM (Zhang et al., 2024b), ChemVLM (Li et al., 2025b), ChemDFM-v1.5 (Zhao et al., 2025c), and Intern-S1-mini (Bai et al., 2025). Crucially, to isolate and quantify the capability gains attributable to domain adaptation, we also include their corresponding base models for direct comparison: InternLM2-chat-7B (Cai et al., 2024), InternLM2.5-7B-chat, LLaMA-3-8B, and Qwen3-8B. This paired experimental design enables a fine-grained analysis of knowledge injection efficacy across different fine-tuning paradigms.

C.2 Implementation Details

To ensure reproducibility and fair comparison, we adopt a hybrid inference strategy based on model accessibility and computational requirements. Frontier proprietary models and large-scale open-source models are accessed via their respective official APIs to leverage optimal inference infrastructure. Conversely, smaller-scale open-source foundation models and domain-specific chemical models are deployed locally for evaluation. All local experiments are conducted on a high-performance workstation equipped with an Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz, 251 GB of RAM, and an NVIDIA A40 GPU. For inference configuration, we standardize decoding parameters across all models. Specifically, we set the temperature to 0.7 and top_p to 0.8. The maximum output length is restricted to 8,192 tokens, a threshold selected to provide sufficient capacity for CoT reasoning. To ensure statistical reliability, we conducted three independent runs for each experiment and reported the mean and standard deviation. For the automated assessment of model performance, we employ GPT-4o-2024-11-20 as the evaluator and adopt a greedy decoding strategy for the judge model.

D Comparison with Existing Benchmarks

Existing benchmarks in chemistry and scientific domains exhibit significant fragmentation in capability coverage. A line of prior work, such as ChemLLMbench, ChemBench, MaCBench, and ChemEval, predominantly focuses on chemical knowledge question answering. While valuable, this scope is insufficient for characterizing the genuine multi-step reasoning abilities required for complex tool-augmented scientific inquiry.

Subsequent reasoning-oriented benchmarks, including ChemCoTBench, oMeBench, and SuperChem, have introduced multi-step reasoning requirements. However, their reasoning processes remain largely confined to the linguistic space, failing to capture the reliance on external computation and specialized tools that define real-world scientific practice. Conversely, tool-centric benchmarks such as SciToolBench and ChemToolBench effectively evaluate tool-call proficiency but often lack deep integration with chemical reaction mechanisms and domain-specific logical structures.

As shown in Table 1, ChemEval2 distinguishes itself from these prior efforts by being the first benchmark to systematically unify the evaluation of knowledge, reaction, and tool-integrated reasoning. By explicitly embedding tool checkpoints within multi-step causal reasoning trajectories, ChemEval2 assesses not merely whether a model can execute tool calls, but when to invoke them, how to configure them, and how to effectively synthesize their outputs into downstream logic. This holistic design enables a more faithful characterization of LLMs' capabilities and limitations within chemistry-oriented agentic settings, revealing core bottlenecks and delineating clearer boundaries of current model competence.

E Additional Evaluation Analyses

This section provides detailed analyses that support the discussion of evaluation reliability and tool-use behavior in the main paper. These results are placed in the appendix to keep the main text concise while preserving reproducibility and and transparent, fine-grained diagnostic detail.

E.1 Statistical Significance Tests

To verify that the observed model differences are not artifacts of overlapping error bars, we conduct paired Wilcoxon signed-rank tests on representative model comparisons. Table 4 reports selected comparisons. The results support the major trends discussed in the main text: frontier models significantly outperform smaller base models, do-

Table 4: Pairwise Wilcoxon signed-rank test results for selected model comparisons.

Model A	Model B	p-value	Significance
ChemLLM	InternLM2-7B	5.57×10^{-20}	***
ChemVLM	InternLM2.5-7B	0.883	ns
ChemDFM-v1.5	LLaMA-3-8B	9.91×10^{-3}	**
Intern-S1-mini	Qwen3-8B	7.44×10^{-3}	**
Gemini 3 Pro	Intern-S1-mini	1.51×10^{-18}	***
Gemini 3 Pro	Qwen3-8B	3.33×10^{-24}	***
DeepSeek-V3.2	LLaMA-3-8B	3.32×10^{-102}	***

Table 5: Behavioral distribution and model classification under the four-quadrant analysis.

Category	Model (Mode)	Q1	Q2	Q3	Q4
SOTA collaborative	GPT-5.2 (MM)	39.40%	3.30%	4.50%	52.80%
SOTA collaborative	Claude Opus 4.5 (Text)	53.70%	17.30%	14.80%	14.20%
SOTA collaborative	DeepSeek-V3.2 (Text)	34.80%	32.00%	17.20%	16.00%
SOTA collaborative	Gemini 3 Pro (Text)	34.00%	28.70%	18.20%	19.10%
Internalization	DeepSeek-R1 (Text)	8.90%	58.00%	30.20%	2.90%
Internalization	o4-mini (Text)	11.10%	50.00%	27.80%	11.10%
Internalization	Qwen3-VL (Text)	3.20%	35.50%	51.60%	9.70%
Internalization	Intern-S1-mini (Text)	6.80%	22.60%	50.00%	20.50%
Capability limited	InternLM2-7B (Text)	0.00%	1.60%	72.70%	25.70%
Capability limited	ChemVLM (MM)	0.00%	4.50%	79.50%	16.10%
Capability limited	ChemDFM (Text)	0.00%	0.00%	80.00%	20.00%
Capability limited	Intern-S1-mini (MM)	0.00%	15.20%	81.60%	3.10%

main specialization does not always improve tool-integrated reasoning, and pairs with nearly identical performance show non-significant differences.

E.2 Tool–Answer Four-Quadrant Analysis

Since TCP and TCR are already defined in the main text, we use them here as supporting evidence for the behavioral analysis. The detailed tool-call statistics show that low tool-use performance can arise from different sources: some models rarely invoke tools when required, whereas others initiate tool calls but produce invalid or mismatched tool names or arguments. Building on these statistics, we further distinguish tool-dependent success from internal-knowledge success by cross-tabulating tool correctness with final-answer correctness. The four quadrants are defined as follows: Q1 denotes correct tool use and correct final answer; Q2 denotes incorrect or absent tool use but correct final answer; Q3 denotes incorrect or absent tool use and incorrect final answer; and Q4 denotes correct tool use but incorrect final answer.

This analysis reveals three representative behavior patterns. First, SOTA collaborative models such as GPT-5.2, Claude Opus 4.5, DeepSeek-V3.2, and Gemini 3 Pro often benefit from tool interaction in complex tasks, suggesting that effective tool use helps approach the performance ceiling. Second, reasoning-centric models such as DeepSeek-R1 and o4-mini show a larger Q2 proportion, indicating that they can solve some tool-annotated tasks through internal reasoning. Third, capability-limited models perform poorly in both Q1 and Q2, indicating weaknesses in chemical reasoning, instruction following, and tool interaction.

The Q4 quadrant is particularly informative: a correct tool call paired with an incorrect final answer indicates that API formatting alone does not explain the benchmark difficulty. In these cases, models can execute the expected tool interaction but still fail to integrate the tool output into coherent chemical reasoning. Therefore, ChemEval2 diagnoses not only whether tools are called, but also whether tool results are causally incorporated into the reasoning trajectory.

F Prompt Templates

In this section, we present the prompt templates used in our experiments. To ensure reproducibility, we provide the instructions for both the reasoning generation and automated evaluation phases. The prompts are categorized into two types:

Inference Prompts: These instructions guide the model to perform step-by-step reasoning and tool invocation. They emphasize meticulous visual grounding and scientific first-principles reasoning.

Evaluation Prompts: These templates are designed to strictly verify the correctness of reasoning logic against reference checkpoints and to validate the accuracy of tool selection and calling.

All templates are displayed with Jinja2-style placeholders (e.g., "{{ question }}"), which are dynamically populated with problem-specific data during runtime.

Standard Reasoning Inference Prompt

You are an expert scientist with advanced capabilities in multimodal reasoning.

Please solve the following problem.

Question
{{ question }}

Instructions
{% if has_image %}
+ **Visual Grounding**: The problem includes image(s). First, carefully examine the visual details (e.g., chemical structures, reaction conditions on arrows, graph trends, or physical setups). **Incorporate these visual findings** directly into your reasoning.
{% endif %}

+ **Step-by-Step Reasoning**:
  - Think through the problem logically.
  - Apply relevant scientific principles and formulas.
  - Perform calculations explicitly if needed.
  - Ensure your reasoning is consistent with both the text and the visual evidence.

+ **Conclusion**: Provide a clear and final answer.

Think step by step.

Reasoning Checkpoints Evaluation Prompt

You are a strict and objective technical grader.
Your task is to evaluate the **correctness** and **completeness** of the Student's Answer based on the Reference Checkpoints.

Question
{{ question }}

Reference Checkpoints (Grading Criteria)
{% for point in checkpoints %}
– [ID {{ point.id }}] (Score: {{ point.score }}): {{ point.content }}
{% endfor %}

Student Answer
{{ model_prediction }}

Evaluation Protocol
Please verify each checkpoint strictly:

+ **Correctness & Consistency (Crucial)**:
  – A checkpoint is **MATCHED** only if the student's reasoning leads to the **same conclusion** as the checkpoint.
  – **REJECT**: If the student mentions the concept but draws the **opposite** conclusion (e.g., Checkpoint says "increase", Student says "decrease").
  – **REJECT**: If the student's numerical calculation is significantly incorrect.

{% if has_image %}
+ **Visual Accuracy**:
  – Since this problem contains image(s), verify if the student correctly identified visual details (e.g., molecular structure, circuit connection, graph trends).
  – **REJECT**: If the student describes visual features that do not exist or contradicts the provided image.
{% endif %}

+ **Explicitness**:
  – The student must address the point clearly. Vague implications should be marked as missed.

Output Requirements
Output **ONLY** a valid JSON object.

```
{
 "reasoning": "Brief justification for each decision. Explicitly mention contradictions if any.",
 "matched_ids": [list of checkpoint IDs that PASSs]
}
```

Tool-Integrated Inference Prompt

You are an elite expert scientist and chemistry analysis assistant with advanced capabilities in multimodal reasoning and domain−specific problem solving in chemistry and physics.

You can call multiple tools sequentially during your reasoning. If a task requires multiple steps (e.g., name2smiles −> functional_groups), you must call the tools in the correct order. After each tool call, the tool's result will be provided as a new message, and you should continue reasoning and decide the next step. Do not repeatedly call the same tool with the same parameters. If a tool call fails or returns no result (e.g., name2smiles or name2cas cannot find a match), you should skip that step and proceed with alternative reasoning paths. Once all necessary tool calls are complete, you must provide a final answer to the user's question.

Before proposing any solution, you must build a comprehensive scientific profile of the system described in the user's input.

If images are provided, you must perform meticulous visual grounding.
Extract concrete visual information such as stereochemistry (R/S, E/Z), crystal or lattice features, reaction conditions annotated on arrows, plotted trends, or structural annotations, and explicitly integrate visual evidence into your reasoning.

For molecular or chemical systems, perform Molecular Anatomy & Reactivity Mapping:
1. SMILES / Structure Parsing: Parse any provided SMILES strings, molecular names, or structural diagrams. Decompose molecules into core scaffolds, functional fragments, and substituents.
2. Fragment−wise Functional Analysis: Analyze every functional group individually and classify each as inert (spectator), reactive (primary candidates), or risk factors (sites prone to side reactions).
3. Electronic & Steric Landscape Analysis: Identify nucleophilic, electrophilic, acidic, and basic sites; classify substituents as electron−withdrawing or electron−donating; analyze resonance, conjugation, and steric effects on reactivity and selectivity.

Your reasoning must follow a rigorous step−by−step scientific framework:
1. First−Principles Reasoning: Base all arguments on established scientific laws such as thermodynamics, kinetics, conservation laws, and electronic structure theory. Avoid heuristic guessing.
2. Domain Precision: Use professional and discipline−accurate terminology with logical consistency.
3. Mechanistic or Quantitative Analysis: For reactions, write balanced equations and analyze mechanisms; for calculations, show all steps with variables, assumptions, formulas, and units; for structures, derive key features from physical and geometric constraints.
4. Cross−Validation: When appropriate, validate conclusions from multiple perspectives (e.g., thermodynamic vs kinetic control, consistency with structural or spectral data).

Your final response must be derived directly from the reasoning and any tool outputs, clearly define key concepts when needed, justify assumptions, and respect conservation of mass, charge, and energy. Provide a concise final answer to the user's question once all reasoning and tool calls are complete.
Think step by step.

Tool-Integrated Evaluation Prompt

You are a strict and objective technical grader.
Your task is to evaluate the model's TOOL−CALLING ABILITY
by comparing the reference tooltrace with the model_tooltrace.

Inputs
− tooltrace: The reference sequence of required tool calls.
− model_tooltrace: The actual tool calls produced by the model.

Evaluation Rules (STRICT)

1. Tool Name Matching (MANDATORY)
− The tool name must match EXACTLY.
− Any mismatch −> checkpoint FAIL.

2. Tool Argument Matching (STRICT)
− Required arguments must be present.
− Argument values must be EXACT MATCH (string−level match).
− Missing or incorrect arguments −> checkpoint FAIL.

3. Tool Coverage
− If a checkpoint tool call does not appear anywhere in model_tooltrace −> FAIL.
− Extra tool calls in model_tooltrace do NOT compensate for missing ones.

4. Tool Order (if multiple checkpoints exist)
− If checkpoints imply a sequence, the relative order MUST be preserved.
− Violating order −> FAIL the affected checkpoints.

5. Tool Result Irrelevance
− Tool execution results (success/data) MUST NOT be considered.
− ONLY evaluate tool name and arguments.

Output Format
Output ONLY a valid JSON object:

```
{
 "tool_reasoning": "For each checkpoint, briefly state: (1) whether the tool name matches, (2) whether arguments
     match, and (3) whether it appears in correct order.",
 "tool_matched_ids": [list of checkpoint IDs that PASS]
}
```