CHEMARENA: EVALUATING TASK-SPECIFIC REASONING REQUIREMENTS IN MULTIMODAL CHEMICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

We present CHEMARENA, a task-aware benchmark for evaluating observable reasoning requirements in multimodal chemical reasoning. It contains 1,033 instances spanning five representative chemical task families, organizing three observable reasoning requirements: Evidence Utilization, Constraint Adherence, and Cross-Turn State Maintenance. To evaluate these task-specific requirements, CHEMARENA introduces a requirement-aware evaluation protocol that integrates task-specific outcome verification, reference-guided process alignment, and context-enriched reasoning assessment. Experiments on 16 representative MLLMs demonstrate that models with similar task-level performance can exhibit substantially different reasoning behaviors across chemical scenarios. CHEMARENA provides a systematic framework for analyzing multimodal chemical reasoning beyond accuracy-based evaluation and enables fine-grained understanding of model capabilities and limitations.

1 INTRODUCTION

Recent advances in Multimodal Large Language Models (MLLMs) have demonstrated substantial potential in chemistry, enabling a wide range of applications such as chemical question answering, molecular analysis, and reaction understanding Li et al. (2025); Huang et al. (2025b); Zhao et al. (2025); Xu et al. (2026). To systematically assess the capabilities of MLLMs in these scenarios, a variety of chemical evaluation benchmarks have been developed, focusing on task-level objectives such as answer correctness, molecular editing success, and property prediction performance Guo et al. (2023); Huang et al. (2025a); Baranwal & Vyas (2026); Hao et al. (2026). However, as chemical reasoning tasks become increasingly complex, outcome-level metrics alone provide limited understanding of the observable reasoning behaviors underlying model predictions. A model may obtain the correct answer without utilizing decisive visual evidence, generate a chemically valid molecule while violating the requested structural modification, or fail to reach the correct conclusion despite satisfying several intermediate requirements. These cases suggest that correct task outcomes do not necessarily reflect reliable reasoning, and different failures may originate from distinct deficiencies in the reasoning process.

To move beyond outcome-level evaluation, recent studies have incorporated process-oriented signals into chemical reasoning assessment, including reasoning-trajectory verification Hao et al. (2026); Guo et al. (2026), intermediate-state checking Guo et al. (2026), and interactive assessment Han et al. (2026). These approaches provide valuable insights by examining whether intermediate reasoning behaviors are correct, consistent, or verifiable. However, existing process-oriented evaluations typically characterize reasoning quality from task-independent perspectives, while overlooking that different chemical scenarios require different reasoning behaviors to achieve reliable solutions. As illustrated in Figure 1, visual chemistry reasoning requires models to ground conclusions in relevant multimodal evidence; molecular editing and optimization require preserving chemical and structural constraints; and interactive reasoning requires maintaining chemically relevant states across turns. This heterogeneity raises a fundamental question: *how can multimodal chemical reasoning be evaluated according to the observable requirements that define successful reasoning in different chemical scenarios?*

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.

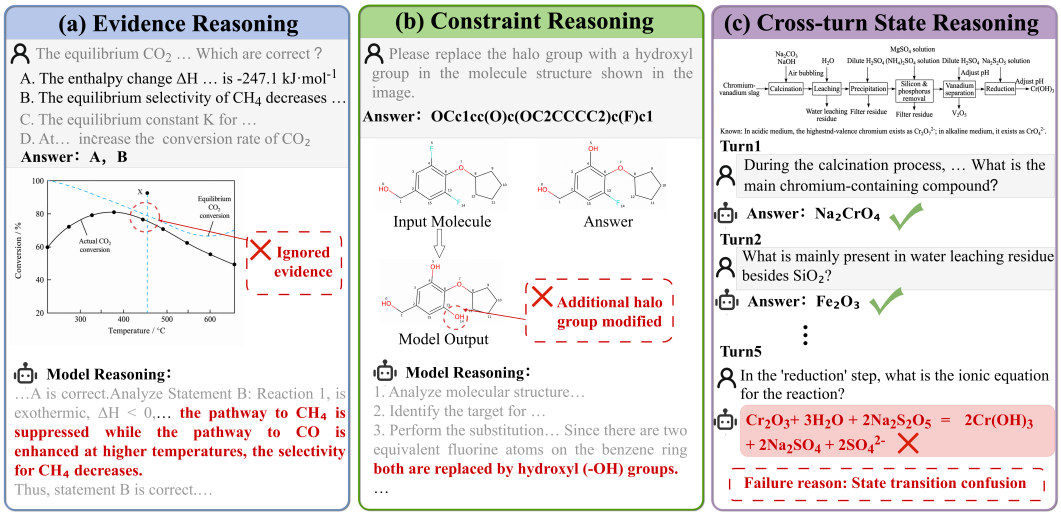


Figure 1: Motivation for requirement-aware evaluation in CHEMARENA. Although MLLMs may produce correct final answers, their generated reasoning traces can still fail to satisfy task-specific requirements. Three representative examples illustrate (a) insufficient Evidence Utilization in chemistry visual reasoning, (b) failure to Constraint Adherence in molecular editing, and (c) failure of Cross-Turn State Maintenance across turns in multi-turn reasoning.

Based on this perspective, we introduce CHEMARENA, a task-aware benchmark for evaluating multimodal chemical reasoning. CHEMARENA contains 1,033 instances spanning five representative chemical task families, including chemistry visual reasoning, molecular editing, molecular optimization, molecular property prediction, and multi-turn reasoning. Each task family is organized according to its dominant reasoning requirement and evaluated through a requirement-aware protocol that combines task-specific outcome verification, reference-guided process alignment, and context-enriched reasoning assessment. We evaluate 16 representative MLLMs on CHEMARENA and show that models with comparable task-level performance may exhibit substantially different reasoning behaviors across chemical scenarios. These results demonstrate that requirement-aware evaluation provides complementary insights beyond conventional accuracy-based comparison.

2 RELATED WORK

Chemical Evaluation Benchmarks. Chemistry-oriented benchmarks have expanded from text-based question answering to a broader spectrum of capabilities, including molecular generation, property prediction, and multimodal molecular understanding, etc. Guo et al. (2023); Fang et al. (2024); Mirza et al. (2024); Yu et al. (2024); Li et al. (2025); Xie et al. (2025); Zhou et al. (2025); Cui et al. (2025); Li et al. (2026); Runcie et al. (2025); Baranwal & Vyas (2026); Huang et al. (2025a). Although these benchmarks cover diverse chemical tasks and modalities, their evaluation criteria are primarily designed around task-level objectives, such as answer correctness, generation success, or prediction accuracy, providing limited support for analyzing observable reasoning behaviors across heterogeneous chemical scenarios.

Recent chemistry benchmarks have begun to incorporate process-level information through different forms of reasoning representation. One line of work focuses on trajectory-oriented reasoning representations and evaluation. ChemCoTBench Hao et al. (2026) formalizes chemical reasoning through step-wise modular operations, while SUPERChem Zhao et al. (2025) provides expert-authored solution paths and evaluates reasoning-path fidelity. Another line introduces structured intermediate states that can be explicitly verified. ChemCoTBench-V2 Guo et al. (2026) evaluates structured intermediate commitments using deterministic chemical rules and reference traces. An interaction-oriented direction is represented by MolQuest Han et al. (2026), where models iteratively acquire experimental evidence and update molecular hypotheses across multiple turns. Beyond textual reasoning processes, ChemVTS-Bench Huang et al. (2025b) investigates multimodal reasoning un-

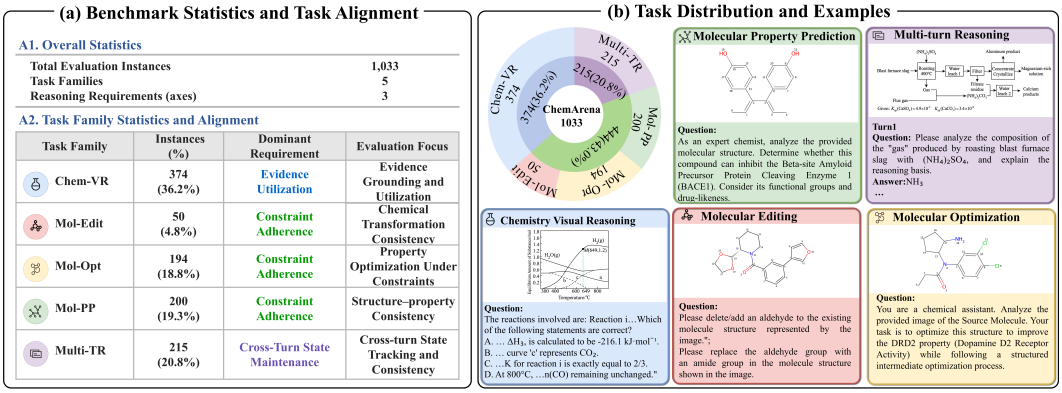


Figure 2: Overview of the CHEMARENA benchmark composition and task organization. **(a)** Benchmark statistics and the primary alignment of five task families with three observable reasoning requirements: Evidence Utilization, Constraint Adherence, and Cross-Turn State Maintenance. **(b)** Task distribution and representative instances from chemistry visual reasoning, molecular editing, molecular optimization, molecular property prediction, and multi-turn reasoning.

der different visual, visual-textual, and symbolic input configurations, while ReactBench Xu et al. (2026) evaluates spatial and topological reasoning in chemical reaction diagrams.

Despite the efforts for understanding chemical reasoning, existing evaluation representations remain closely tied to individual task settings and lack a unified perspective for comparing reasoning behaviors across heterogeneous chemical scenarios. CHEMARENA addresses this gap by introducing a requirement-oriented evaluation framework that organizes diverse chemical tasks according to observable reasoning requirements and evaluates them through complementary process-level signals.

Reasoning Process Evaluation and Verification. Reasoning process evaluation examines whether generated reasoning artifacts are grounded, consistent, and sufficient to support their conclusions, rather than relying exclusively on final-answer correctness Lanham et al. (2023); Turpin et al. (2023). Existing approaches include deterministic verifiers, process reward models, trajectory alignment, and LLM-as-Judge evaluation. Process reward models have been most successful in domains with structured outputs and readily checkable intermediate steps, whereas multimodal chemical reasoning involves heterogeneous visual evidence, domain-specific constraints, and partially open-ended explanatory steps Lightman et al. (2024); Zhang et al. (2025). MME-CoT Jiang et al. (2025) evaluates multimodal CoT quality, robustness, and efficiency, but does not provide chemistry-specific verification of Evidence Utilization, Constraint Adherence, or Cross-Turn State Maintenance. LLM-as-Judge methods provide flexible semantic assessment, yet their reliability can vary with judge selection, scoring design, and evaluation context Li et al. (2024); Lehr et al. (2025).

CHEMARENA therefore adopts a three-stage evaluation protocol comprising Task-Specific Outcome Verification, Reference-Guided Process Alignment, and Context-Enriched Reasoning Assessment. The protocol preserves Outcome Performance, Process Alignment, and Reasoning Quality as three complementary metrics, together with an explicit Failure Diagnosis as distinct but connected evaluation outputs, with the structured artifacts from earlier stages informing subsequent judgments.

3 CHEMARENA FRAMEWORK

3.1 OVERVIEW

CHEMARENA is a task-aware benchmark designed to diagnose observable failures in multimodal chemical reasoning. Rather than assessing task completion alone, it organizes evaluation around three observable reasoning requirements that serve as requirement axes: Evidence Utilization, Constraint Adherence, and Cross-Turn State Maintenance.

As summarized in Figure 2, CHEMARENA contains 1,033 evaluation instances spanning five task families: chemistry visual reasoning, molecular editing, molecular optimization, molecular property

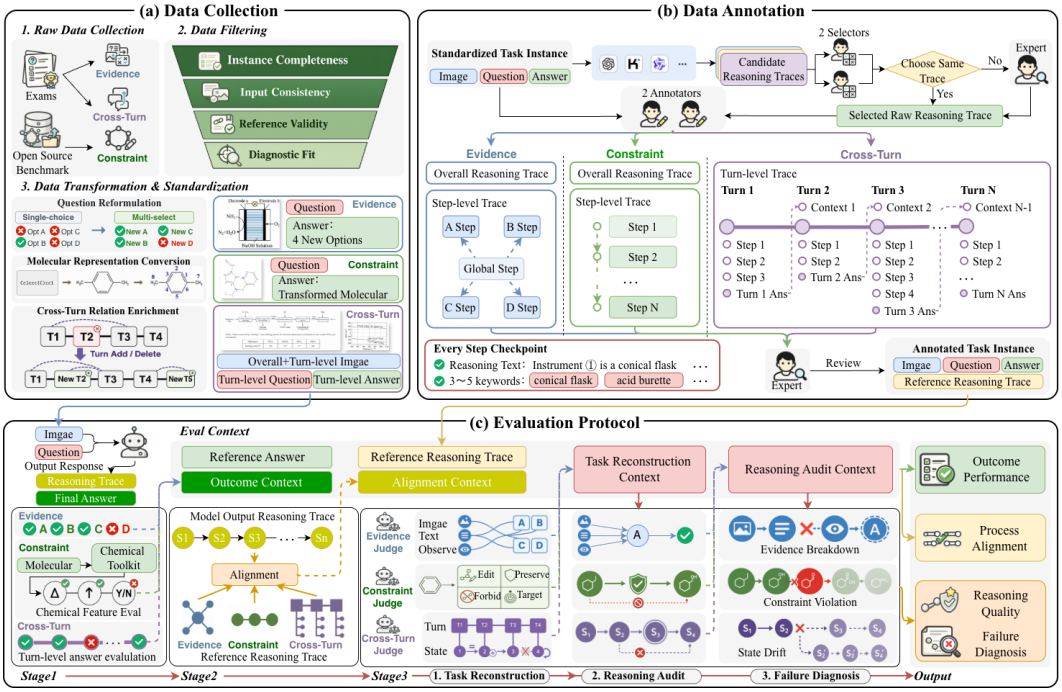


Figure 3: Overview of the CHEMARENA construction pipeline and its requirement-aware, three-stage evaluation framework. **(a) Data Collection** covers raw-data acquisition, quality filtering, and task transformation and standardization. **(b) Data Annotation** generates candidate reasoning traces, selects reference traces through independent review and expert adjudication, and converts them into requirement-specific structures with step-level checkpoints. **(c) Evaluation Protocol** proceeds through Stage I, Task-Specific Outcome Verification; Stage II, Reference-Guided Process Alignment; and Stage III, Context-Enriched Reasoning Assessment. Structured contexts are carried forward across stages, while Stage III further comprises Task Reconstruction, Reasoning Audit, and Failure Diagnosis. Together, the three stages produce Outcome Performance, Process Alignment, and Reasoning Quality, along with a Failure Diagnosis.

prediction, and multi-turn reasoning. Each task family is associated with the primary requirement axis that best captures its core evaluation objective, enabling a unified comparison of reasoning behaviors across heterogeneous chemical scenarios.

3.2 REQUIREMENT AXES AND TASK-FAMILY ALIGNMENT

CHEMARENA organizes its five task families according to the primary observable reasoning requirement emphasized by their task design and evaluation, rather than treating each family as an independent reasoning dimension. This mapping is primary rather than exclusive: a task may involve multiple requirements, while its assigned axis identifies the one it most directly evaluates. The three requirement axes are Evidence Utilization, Constraint Adherence, and Cross-turn State Maintenance.

Evidence Utilization. Evidence Utilization assesses whether a model grounds its conclusion in task-relevant multimodal evidence. Chemistry visual reasoning instantiates this axis by requiring models to integrate textual descriptions, options, molecular structures, and chemical diagrams, and connect decisive evidence to the final conclusion.

Constraint Adherence. Constraint Adherence assesses whether reasoning and outputs remain consistent with task-specified chemical conditions. Molecular editing tests intended transformations and structure preservation; molecular optimization tests property improvement under validity

and task constraints; and molecular property prediction tests consistency with relevant structural features and structure–property relationships.

Cross-Turn State Maintenance. Cross-turn State Maintenance assesses whether a model preserves and correctly updates chemically relevant information across interactions. Multi-turn reasoning instantiates this axis because later responses depend on prior conditions, intermediate conclusions, corrections, or updated chemical states.

Together, the five task families provide task-specific realizations of three shared reasoning requirements, supporting the analysis of observable reasoning behaviors across different chemical task families.

3.3 Data Construction and Annotation

ChemArena draws on two data sources: chemistry examination materials for Evidence Utilization and Cross-turn State Maintenance, and molecule-centered tasks adapted from ChemCoTBench Hao et al. (2026) for Constraint Adherence. Candidate instances are filtered for completeness, input consistency, reference validity, and diagnostic suitability. We then apply task-specific transformations, including question reformulation, molecular representation conversion with standardized atom indices, and cross-turn relation enrichment, producing standardized task instances with consistent inputs, questions, and reference answers.

For each instance, general-purpose and chemistry-oriented models repeatedly sample candidate reasoning traces, yielding approximately 15 candidates on average. Two chemistry-trained selectors independently perform a two-stage selection: they first identify a suitable trace within each model and then compare the model-level selections. Agreement yields a selected raw reasoning trace, while disagreements are resolved through expert adjudication. This procedure filters redundant, exploratory, or internally inconsistent generations before annotation.

Using both the standardized task instance and the selected raw trace, two annotators independently construct requirement-specific annotations. Evidence Utilization uses an overall trace with option-specific and global step-level checkpoints; Constraint Adherence uses an overall trace and an ordered constraint chain; and Cross-turn State Maintenance uses turn-level traces with explicit context and state dependencies. Each checkpoint records its reasoning text and 3–5 keywords, together with dependency links where required. The keywords support subsequent checkpoint alignment, while the overall or turn-level structure provides global context for semantic assessment. An expert finally reviews the annotation structure, chemical validity, and checkpoint consistency, producing an annotated task instance containing the standardized input, question, answer, and reference reasoning trace. The reference trace serves as an evaluation anchor rather than the only valid reasoning path.

3.4 Evaluation Protocol

As illustrated in Figure 3, ChemArena adopts a task-specific, three-stage evaluation protocol. The protocol progresses from Task-Specific Outcome Verification, through Reference-Guided Process Alignment, to Context-Enriched Reasoning Assessment. It evaluates both final answers and reasoning traces because Evidence Utilization, Constraint Adherence, and Cross-Turn State Maintenance require distinct notions of success beyond uniform answer matching. To distinguish task completion from reasoning reliability and localize errors, Stage I verifies task outcomes, Stage II aligns reasoning traces with task-critical checkpoints, and Stage III performs context-enriched semantic judgment and failure diagnosis. Each stage retains its structured intermediate artifacts in a shared evaluation context to inform subsequent stages.

Task-Specific Outcome Verification. Task specificity begins at the outcome level, where each task family is evaluated against its own success criteria. Evidence Utilization uses normalized option-set matching, whereas Constraint Adherence applies chemical checks tailored to molecular editing and preservation, constrained optimization, or property prediction. Cross-Turn State Maintenance selects a turn-level evaluator according to the answer type, such as multiple-choice selections, chemical expressions, or numerical values, before aggregating dialogue-level performance. Rather than reducing these evaluations to outcome scores, Stage I retains normalized outputs, parsing status, deterministic verification results, matched and unmatched answer components, and turn-wise

answer states in an *Outcome Context*. This context allows subsequent stages to interpret the reasoning in light of what the model actually produced, including cases in which a correct final answer lacks adequate reasoning support.

Reference-Guided Process Alignment. Stage II aligns the model's reasoning trace with a fixed set of task-critical checkpoints from human-verified reference annotations, using deterministic rules rather than the LLM Judge. Encoding accepted lexical variants, step types, prerequisites, and modality or history requirements, these checkpoints allow the matcher to locate candidate spans and record complete, partial, or missing coverage together with rule-verifiable ordering and dependencies. Task specificity is reflected in the alignment topology: an overall-plus-options structure for Evidence Utilization, a constraint chain for Constraint Adherence, and a turn-state graph for Cross-Turn State Maintenance. The reference trajectory provides evaluation anchors rather than a canonical reasoning script, leaving chemically valid alternative paths admissible. These matches yield rule-derived metrics such as checkpoint recall and order or dependency coverage, while their structured records form the *Alignment Context* for Stage III. This context localizes relevant reasoning evidence without treating lexical coverage as proof of chemical or causal correctness.

Context-Enriched Reasoning Assessment. This stage employs a human-calibrated LLM Judge to assess whether the model's reasoning is chemically valid and sufficient to support its final answer, and to identify the source of consequential failures. The Judge receives the original task instance, reference answer and process annotations, model response, and the *Outcome Context* and *Alignment Context* produced by the preceding stages. The evaluation proceeds through three logical passes: *Task Reconstruction*, *Reasoning Audit*, and *Failure Diagnosis*. The structured output of each pass is appended to the shared evaluation context and used to inform subsequent judgments.

1. Task Reconstruction. Task specificity first enters through the representation against which the reasoning trace will be judged. Evidence Utilization is represented as an evidence–option map connecting decisive multimodal observations to option-level conclusions; Constraint Adherence as a contract specifying the intended transformation, protected conditions, forbidden changes, and target properties; and Cross-Turn State Maintenance as an expected trajectory of state preservation, update, rejection, and correction. By making explicit what must be established, preserved, or updated, this pass converts heterogeneous task requirements into concrete targets for semantic evaluation. The resulting representations constitute the *Task Reconstruction Context*.

2. Reasoning Audit. The second pass evaluates the model's reasoning against the reconstructed task representation rather than its surface similarity to the reference trajectory. The audit is correspondingly specialized: evidence–option relations are examined for correct grounding and answer support, constraint contracts for valid transformations and constraint propagation, and state trajectories for appropriate history use and state updates. During this process, the *Outcome Context* exposes discrepancies between the final result and its reasoning support, while the *Alignment Context* directs the Judge to matched spans, missing checkpoints, and dependency cues. A dialogue of T turns expands the state audit into T sequential judgments, resulting in $1 + T + 1$ Judge calls across the three logical passes. The resulting judgments record step-level correctness, dependency validity, unsupported or contradictory claims, and state inconsistencies in the *Reasoning Audit Context*.

3. Failure Diagnosis. The final pass treats the audited trace as an error process and seeks the first consequential failure rather than merely aggregating local scores. The relevant failure unit follows the reconstructed task representation: a broken evidence path, a violated constraint or invalid structural transition, or a divergent dialogue state. Tracing the downstream effects of this failure supports shared semantic metrics such as step correctness, path fidelity, and answer support, together with axis-specific metrics and a human-readable diagnostic report. The final report preserves Outcome Performance, Process Alignment, and Reasoning Quality as distinct evaluation metrics, together with Failure Diagnosis. The Judge used in Stage III is selected through a separate human-calibration study and does not introduce an additional per-instance evaluation stage.

Table 1: Performance of 16 MLLMs on CHEMARENA across the three observable reasoning requirements. Outcome, Align., and Quality denote Outcome Performance, Process Alignment, and Reasoning Quality, respectively. Comp. denotes the composite score within each requirement, and Overall is the mean of the three requirement-level composite scores. Best, second-best, and third-best distinct results in each column are shown in **bold**, underlined, and *italic*, respectively. All scores are percentages.

Model	Evidence Utilization				Constraint Adherence				Cross-Turn State Maintenance				Overall
	Outcome	Align.	Quality	Comp.	Outcome	Align.	Quality	Comp.	Outcome	Align.	Quality	Comp.	
GPT-5.4-xHigh	82.9	84.3	*91.2*	86.2	*73.1*	**58.6**	**78.1**	**70.0**	53.7	*85.3*	80.8	73.3	**76.5**
Qwen3.5-397B	**88.1**	84.1	91.6	**88.0**	65.6	57.7	65.7	63.0	53.2	86.9	87.0	75.8	75.6
Kimi-K2.6	80.3	**86.9**	88.2	85.2	73.2	58.5	65.8	65.8	*56.4*	76.5	*82.5*	71.9	*74.3*
GPT-5.4-High	80.6	*85.0*	84.9	83.5	69.0	**58.6**	72.4	66.7	54.6	79.7	80.4	71.7	74.0
Gemini 3.1 Pro	82.8	83.4	86.5	84.3	63.2	52.3	66.2	60.6	**57.5**	**89.9**	80.9	**76.2**	73.7
Qwen3.5-122B	*86.7*	84.9	90.1	*87.3*	61.1	53.0	62.7	59.0	57.2	80.0	82.0	73.1	73.1
Qwen3.5-27B	85.4	84.2	89.3	86.3	58.4	51.9	58.0	56.1	54.1	83.6	**88.3**	*75.5*	72.6
GPT-5.4-Medium	76.0	84.8	81.5	80.8	69.0	*58.4*	*73.0*	*66.9*	50.8	81.3	77.6	70.0	72.5
GPT-5.4-Low	67.5	79.7	**92.4**	80.0	65.6	58.1	64.0	62.6	47.3	74.6	77.3	66.5	69.7
Gemini 3 Pro	87.5	86.2	89.4	87.7	**75.9**	54.5	73.2	67.9	31.4	48.8	73.0	51.3	69.0
GLM-5V	73.3	77.9	86.0	79.1	59.6	50.5	63.0	57.7	53.4	74.3	81.2	69.7	68.9
Claude Sonnet 4.6	65.8	77.5	76.2	73.2	66.8	46.2	63.8	59.0	51.0	75.6	64.5	63.7	65.3
Gemini 3 Flash	62.5	75.3	74.3	70.7	63.6	46.1	62.0	57.3	47.2	75.8	74.2	65.8	64.6
Intern-S1-Pro	74.2	79.9	91.6	82.0	55.7	46.5	53.5	51.9	44.7	71.6	62.0	59.5	64.5
Intern-S1	61.9	71.6	84.2	72.7	44.9	39.5	41.1	41.8	47.8	72.1	69.1	63.0	59.2
ChemVLM-8B	16.0	50.7	45.2	37.4	36.5	29.0	34.9	33.5	20.0	36.9	21.9	26.2	32.4

Table 2: Meta-evaluation of candidate judge models against human annotations. P and S denote Pearson and Spearman correlations; MAE and RMSE denote mean absolute error and root mean squared error; Pair and Tie denote pairwise and tie-half accuracy; MF1 denotes macro-F1; F1+ and F1- denote positive- and negative-class F1; and κ denotes weighted Cohen's kappa. Comp. is the composite agreement score used for judge selection. Best, second-best, and third-best distinct results in each column are shown in **bold**, underlined, and *italic*, respectively.

Judge Model	Correlation		Error		Preference		Classification			Agreement	Composite
	P ↑	S ↑	MAE ↓	RMSE ↓	Pair ↑	Tie ↑	MF1 ↑	F1+ ↑	F1- ↑	κ ↑	Comp. ↑
GPT-5.4-mini	0.952	0.762	**0.161**	**0.184**	0.821	0.821	**0.873**	**0.889**	**0.857**	**0.750**	**86.4**
Gemini 3 Pro	0.958	*0.714*	*0.251*	*0.314*	*0.786*	*0.786*	0.733	0.800	0.667	0.500	78.0
Gemini 3 Flash	**0.974**	**0.862**	0.410	0.448	**0.857**	**0.875**	*0.564*	*0.727*	*0.400*	*0.250*	*73.0*
Qwen3.6-Flash	0.964	0.695	0.366	0.421	0.750	0.768	*0.564*	*0.727*	*0.400*	*0.250*	69.2
GPT-5.4-nano	0.952	0.595	0.199	0.219	0.750	0.750	*0.564*	*0.727*	*0.400*	*0.250*	68.8
Qwen3.6-Plus	*0.960*	0.667	0.360	0.416	0.750	0.750	*0.564*	*0.727*	*0.400*	*0.250*	68.2

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate 16 representative MLLMs, including general-purpose and chemistry-oriented models, on the five CHEMARENA task families: chemistry visual reasoning, molecular editing, molecular optimization, molecular property prediction, and multi-turn reasoning. The evaluated models cover both closed-source and open-source systems with different model scales and reasoning configurations.

Following the three-stage evaluation protocol described in the Evaluation Protocol section, we report three complementary measurements: outcome performance from Task-Specific Outcome Verification, process alignment from Reference-Guided Process Alignment, and reasoning quality from Context-Enriched Reasoning Assessment. These measurements are analyzed across the three requirement axes defined in CHEMARENA: evidence utilization, constraint adherence, and cross-turn state maintenance. Further experimental details are provided in the supplementary document.

4.2 Overall Evaluation Results

Table 1 summarizes the performance of 16 MLLMs on ChemArena. The highest Overall score is only 76.5%, indicating substantial remaining headroom even for the strongest evaluated systems. Moreover, no model dominates all three requirement axes: Qwen3.5-397B obtains the best Evidence Utilization Composite Score (88.0%),GPT-5.4-xHigh leads Constraint Adherence (70.0%), and Gemini 3.1 Pro leads Cross-Turn State Maintenance (76.2%). Overall therefore reflects a model's balance across requirements rather than universal superiority.

Observation 1: Outcome performance does not fully characterize reasoning quality. Although outcome performance directly measures whether a model completes a task, it does not necessarily indicate whether the underlying reasoning satisfies the required criteria. For instance, in evidence utilization, GPT-5.4-Low achieves only 67.5% outcome performance, while obtaining 79.7% process alignment and 92.4% reasoning quality. This indicates that the model can generate evidence-supported reasoning behaviors even when the final answer remains imperfect. Conversely, in constraint reasoning, Gemini 3 Pro achieves 75.9% outcome performance but only 54.5% process alignment, suggesting that a chemically plausible output does not necessarily imply that all task-specific constraints are properly considered.

Observation 2: Similar Overall scores conceal distinct requirement profiles. Gemini 3 Pro and GLM-5V achieve nearly identical Overall scores (69.0% versus 68.9%), yet their axis-level profiles differ sharply. Gemini 3 Pro exceeds GLM-5V by 8.6 points on Evidence Utilization and 10.2 points on Constraint Adherence, but trails it by 18.4 points on Cross- Turn State Maintenance. Likewise, Qwen3.5-27B and GPT-5.4-Medium differ by only 0.1 points Overall (72.6% versus 72.5%); Qwen3.5-27B is 5.5 points higher on both Evidence Utilization and Cross-Turn State Maintenance , but 10.8 points lower on Constraint Adherence. Thus, models that appear interchangeable under an aggregate ranking may fail in fundamentally different chemical scenarios, making the requirement-level decomposition essential for capability diagnosis.

Observation 3: Improvements are not uniform across requirements. Within the evaluated GPT-5.4 configurations, Overall increases monotonically from 69.7% (Low) to 72.5% (Medium), 74.0% (High), and 76.5% (xHigh). From Low to xHigh, the three axis-level Composite Scores rise by 6.2, 7.4, and 6.8 points, respectively, suggesting that additional reasoning effort primarily improves the balance of the resulting capability profile. A different pattern appears across the evaluated Qwen3.5 scales: Overall rises from 72.6% at 27B to 75.6% at 397B, but the gain is concentrated in Constraint Adherence (+6.9 points), compared with only +1.7 points on Evidence Utilization and +0.3 points on Cross- Turn State Maintenance. These within-family comparisons are descriptive rather than causal, but they show that aggregate improvements can arise from different requirement-specific gains.

4.3 Analysis of Requirement-specific Reasoning

While the overall evaluation demonstrates that reasoning capabilities vary across models, different chemical requirements expose distinct patterns of reasoning behavior. We therefore analyze the three diagnostic axes separately to understand how current MLLMs satisfy or fail the requirements associated with different chemical scenarios.

Evidence Utilization. Evidence Utilization evaluates whether models can effectively identify and use relevant multimodal information during visual chemistry reasoning. The results show that current MLLMs generally achieve relatively strong outcome performance on this axis, but the agreement between final answers and evidence-grounded reasoning varies across models. For example, GPT-5.4-Low achieves 67.5% outcome performance, while obtaining 79.7% process alignment and 92.4% reasoning quality. This indicates that the model can often generate reasoning traces containing relevant visual evidence even when the final answer is not always correct. In contrast, a model may achieve a high outcome score through visual recognition shortcuts without explicitly connecting observations to conclusions. These results suggest that evidence utilization measures a capability beyond visual perception, requiring models to ground decisions on chemically meaningful multimodal evidence.

Constraint Adherence. Constraint Adherence covers molecular editing, molecular optimization, and property prediction tasks, where successful reasoning requires satisfying multiple chemical requirements simultaneously. Compared with evidence utilization, this axis exhibits larger gaps between outcome performance and process alignment. For instance, Gemini 3 Pro achieves 75.9% outcome performance but only 54.5% process alignment and 73.2% reasoning quality. This result suggests that generating chemically plausible outputs does not necessarily indicate that models have correctly considered all task-specific constraints. Such discrepancies are particularly important for molecular reasoning tasks, where preserving structural properties, satisfying optimization objectives, and maintaining chemical validity may involve different requirements that are not reflected by final-answer correctness alone.

Cross-turn State Maintenance. Cross-Turn State Maintenance evaluates whether models can preserve and update evolving information during multi-turn chemical interactions. This axis shows the largest separation between outcome performance and process-oriented evaluations. For example, Gemini 3.1 Pro obtains 57.5% outcome performance while achieving 89.9% process alignment and 80.9% reasoning quality. This indicates that the model can maintain many relevant intermediate states and historical information, but may still fail to correctly apply updated conditions when generating final responses. Therefore, multi-turn chemical reasoning requires not only context retention but also reliable state transition and information update mechanisms.

4.4 LLM-AS-JUDGE CALIBRATION

Since Stage III, Context-Enriched Reasoning Assessment, relies on LLM-as-Judge evaluation, judge reliability is critical for obtaining valid reasoning-quality measurements. We construct a human-annotated calibration subset sampled from the evaluation instances. Each model response is independently scored by eight human annotators with chemistry backgrounds according to predefined criteria, including reasoning support, evidence interpretation, chemical consistency, and overall process quality. The aggregated human annotations serve as the reference for evaluating candidate judge models.

Candidate judge models are calibrated against human annotations through a meta-evaluation of score consistency, preference agreement, and categorical agreement. Table 2 reports the detailed metrics and results. Based on the composite agreement score, GPT-5.4-mini is selected as the primary judge model for Stage III, Context-Enriched Reasoning Assessment. The judge receives the original task instance, reference answer and process annotations, model response, and the Outcome Context and Alignment Context produced by the preceding stages. It then performs three logical passes: Task Reconstruction, Reasoning Audit, and Failure Diagnosis.

5 CONCLUSION

We introduce CHEMARENA, a task-aware benchmark for evaluating observable reasoning requirements in multimodal chemical reasoning. CHEMARENA organizes evaluation around Evidence Utilization, Constraint Adherence, and Cross-Turn State Maintenance. Its three-stage protocol separates task-specific outcome verification, reference-guided process alignment, and context-enriched reasoning assessment, enabling Outcome Performance, Process Alignment, and Reasoning Quality, together with an explicit Failure Diagnosis. Experiments with 16 representative MLLMs show that similar outcome performance can mask substantially different reasoning behaviors and that model strengths vary across chemical reasoning requirements.

Limitations and Future Work. The three diagnostic axes are evaluation-oriented abstractions rather than an exhaustive taxonomy of chemical reasoning capabilities, and the current task families do not cover all chemical modalities, laboratory settings, or forms of scientific reasoning. The framework also relies on structured reference annotations and LLM-based semantic judgments, which entail annotation costs and may introduce reference- or judge-dependent bias despite human verification and calibration. Future work will expand the coverage of chemical tasks and interaction settings, incorporate more diverse valid reasoning paths and stronger human evaluation, and improve the robustness, efficiency, and scalability of automated reasoning assessment.

AI USE STATEMENT

We used generative AI tools for language polishing, translation, code debugging, and presentation refinement. All AI-assisted outputs were manually reviewed and verified by the authors. Generative AI was not used to make benchmark annotation decisions, determine scientific claims, or draw the final conclusions of this work. Experimental results, evaluation procedures, and reported statistics were independently checked by the authors. We take full responsibility for the final content of this paper and all AI-assisted artifacts.

REFERENCES

Aaditya Baranwal and Shruti Vyas. Chempro: A progressive chemistry benchmark for large language models. *Artificial Intelligence Chemistry*, pp. 100118, 2026.

Yiming Cui, Xin Yao, Yuxuan Qin, Xin Li, Shijin Wang, and Guoping Hu. Evaluating large language models on multimodal chemistry olympiad exams. *Communications Chemistry*, 2025.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *ICLR*. OpenReview.net, 2024. URL https://openreview.net/pdf?id=Tlsdsb6l9n.

Hongyu Guo, Hao Li, He Cao, Gongbo Zhang, and Li Yuan. From answers to states: Verifiable process-level evaluation of chemical reasoning in large language models. *arXiv preprint arXiv:2606.03660*, 2026.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. What indeed can gpt models do in chemistry? a comprehensive benchmark on eight tasks, 2023.

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

Li Hao, 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. *Advances in Neural Information Processing Systems*, 38, 2026.

Yuqing Huang, Rongyang Zhang, Xuesong He, Xuyang Zhi, Hao Wang, Nuo Chen, Xin Li, Feiyang Xu, Deguang Liu, Huadong Liang, et al. Chemeval: A multi-level and fine-grained chemical capability evaluation for large language models. 2025a.

Zhiyuan Huang, Baichuan Yang, Zikun He, Yanhong Wu, Fang Hongyu, Zhenhe Liu, Lin Dongsheng, and Bing Su. Chemvts-bench: Evaluating visual-textual-symbolic reasoning of multimodal large language models in chemistry. *arXiv preprint arXiv:2511.17909*, 2025b.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Steven A Lehr, Mary Cipperman, and Mahzarin R Banaji. Extreme self-preference in language models. *arXiv preprint arXiv:2509.26464*, 2025.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

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.

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

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 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. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*, 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, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

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.

Qiang Xu, Shengyuan Bai, Yu Wang, He Cao, Leqing Chen, Yuanyuan Liu, Bin Feng, Zijing Liu, and Yu Li. Reactbench: A benchmark for topological reasoning in mllms on chemical reaction diagrams. *arXiv preprint arXiv:2604.15994*, 2026.

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.

Yao Zhang, Yu Wu, Haowei Zhang, Weiguo Li, Haokun Chen, Jingpei Wu, Guohao Li, Zhen Han, and Volker Tresp. Groundedprm: Tree-guided and fidelity-aware process reward modeling for step-level reasoning. *arXiv preprint arXiv:2510.14942*, 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*, 2025.

Yitong Zhou, Mingyue Cheng, Qingyang Mao, Yucong Luo, Qi Liu, Yupeng Li, Xiaohan Zhang, Deguang Liu, Xin Li, and Enhong Chen. Benchmarking multimodal llms on recognition and understanding over chemical tables. *arXiv preprint arXiv:2506.11375*, 2025.

A Extended Comparison with Existing Chemistry Benchmarks

Table 3 provides an extended comparison between ChemArena and representative chemistry benchmarks. The comparison considers input modalities, the availability of reference process information, explicit process-level evaluation, coverage of three requirement axes for observable failure diagnosis, and support for localized error analysis. It complements the discussion in the main paper by distinguishing broad task coverage from explicit evaluation of generated reasoning artifacts.

The entries summarize the evaluation functionality explicitly supported by each benchmark. A benchmark is marked as fully supported when the corresponding capability is an explicit part of its released task or evaluation design, partially supported when the capability is available only for a

Table 3: Comparison of chemistry benchmarks in terms of input modalities, process-level support, and requirement coverage. **Traj.** indicates the availability of reference trajectories or structured reasoning traces, and **Proc.** indicates explicit process-level evaluation. **Evid.**, **Constr.**, and **State** denote coverage of Evidence Utilization, Constraint Adherence, and Cross-Turn State Maintenance, respectively. **Error Loc.** indicates whether a benchmark supports localization of reasoning failures. ✓: supported; ●: partially supported; ✗: not supported.

			Input Modalities			Process-Level Support		Requirement Coverage			Error Analysis
Benchmark	**Year**	**Task Scope / Data Scale**	**Text**	**Image**	**Dialog.**	**Traj.**	**Proc.**	**Evid.**	**Constr.**	**State**	**Error Loc.**
ChemLLMBench Guo et al. (2023)	2023	8 tasks / ∼800 samples	✓	✗	✗	✗	✗	✗	✗	✗	✗
Mol-Instructions Fang et al. (2024)	2023	Multi-task instructions / 706K samples	✓	✗	✗	✗	✗	✗	✗	✗	✗
ChemBench Mirza et al. (2024)	2024	QA benchmark / 2.7K+ QA pairs	✓	✗	✗	✗	✗	✗	✗	✗	✗
LlaSMol / SMolInstruct Yu et al. (2024)	2024	Molecular instruction tasks / 3.3M samples	✓	✗	✗	✗	✗	✗	✗	✗	✗
ChemVLM Li et al. (2025)	2025	3 benchmarks / ChemOCR, MMCR-Bench, MMChemBench	✓	✓	✗	✗	✗	✗	✗	✗	✗
QCBench Xie et al. (2025)	2025	Quantitative reasoning benchmark / 350 problems	✓	✗	✗	✗	✗	✗	✗	✗	✗
ChemTable Zhou et al. (2025)	2025	Table QA benchmark / 1.3K+ tables, 9K+ QA pairs	✓	✓	✗	✗	●	✓	✗	✗	●
USNCO-V Cui et al. (2025)	2025	Exam benchmark / 473 multimodal MCQs	✓	✓	✗	✗	●	●	✗	✗	●
RxnBench Li et al. (2026)	2025	Reaction/document QA benchmark / 1,525 SF-QA, 540 FD-QA	✓	✓	✗	✗	●	●	●	✗	●
ChemIQ Runcie et al. (2025)	2026	Chemical intelligence benchmark / 816 questions	✓	✗	✗	✗	✗	✗	✗	✗	✗
ChemPro Baranwal & Vyas (2026)	2026	Progressive QA benchmark / 4.1K QA pairs	✓	✗	✗	✗	✗	✗	✗	✗	✗
ChemEval Huang et al. (2025a)	2026	62 tasks / 3.16K instances	✓	✓	✗	✗	✗	●	●	✗	✗
ChemCoTBench Hao et al. (2026)	2025	Modular reasoning tasks / nearly 20K CoT samples	✓	✗	✗	✓	●	✗	●	✗	●
SUPERChem Zhao et al. (2025)	2025	Multimodal reasoning benchmark / 500 problems, 3K+ checkpoints	✓	✓	✗	✓	✓	●	●	✗	●
ChemCoTBench-V2 Guo et al. (2026)	2026	18 tasks / 5,620 samples	✓	✗	✗	✓	✓	✗	✓	●	✓
MolQuest Han et al. (2026)	2026	Interactive elucidation benchmark / 530 cases	✓	●	✓	●	✓	●	●	●	✗
ChemVTS-Bench Huang et al. (2025b)	2025	VTS reasoning benchmark / 445 questions, 3 input modes	✓	✓	✗	✗	●	●	✗	✗	●
ReactBench Xu et al. (2026)	2026	Reaction diagram benchmark / 1,618 QA pairs	✓	✓	✗	✗	✗	●	●	✗	●
ChemArena	–	**5 task families / 1.0K+ instances**	✓	✓	✓	✓	✓	✓	✓	✓	✓

subset of tasks or through indirect analysis, and unsupported when it is not included in the reported benchmark protocol. These categories describe benchmark-level evaluation coverage rather than the general capabilities of the models evaluated on each benchmark.

In Table 3, **Traj.** denotes human-provided reference reasoning trajectories or structured step-wise rationales, and **Proc.** denotes an explicit process-level evaluation mechanism. **Evid.**, **Constr.**, and **State** denote evaluation coverage of Evidence Utilization, Constraint Adherence, and Cross-Turn State Maintenance, respectively. **Error Loc.** indicates whether the evaluation protocol can associate an observed failure with a specific reasoning step, evidence use, constraint violation, or intermediate state.

B Data Provenance and Source Details

This appendix documents the provenance, transformation, annotation, and evaluation procedures of ChemArena. The benchmark comprises 1,033 evaluation units across five task families. We use the term *evaluation unit* because, for the first four task families, each unit corresponds to a single-turn instance, whereas for multi-turn reasoning, each unit corresponds to an individually evaluated dialogue turn. This distinction is important: the multi-turn portion consists of 42 dialogues comprising 215 evaluated turns. Consistent with the main paper, the appendix organizes the benchmark and its evaluation around three requirement axes for observable failure diagnosis: *Evidence Utilization*, *Constraint Adherence*, and *Cross-Turn State Maintenance*. The five task families are chemistry visual reasoning, molecular editing, molecular optimization, molecular property prediction, and multi-turn reasoning.

B.1 Benchmark Composition

Table 4 summarizes the final benchmark composition. Each task family is assigned to a primary requirement axis according to the observable reasoning requirement it is designed to probe. This assignment guides task selection, annotation, evaluation, and reporting.

B.2 Molecule-Centered Tasks

Source and adaptation. The molecule-centered task families are adapted from the modular molecular reasoning setting introduced by ChemCoTBench Hao et al. (2026). ChemArena retains three task families: molecular editing, molecular optimization, and molecular property prediction. The adaptation changes their evaluation role. Rather than treating them only as independent

Table 4: Composition of CHEMARENA. Trajectory coverage counts instances with a human-reviewed reference rationale and structured step annotations. Multi-turn counts are turn-level evaluation units.

Requirement Axis	Task Family	Unit	Count	Traj.	Coverage	Input and Target
Evidence Utilization	Chemistry visual reasoning	Question	374	374	100.0%	Text, chemistry visual, and answer options; exact option set
Constraint Adherence	Molecular editing	Molecule	50	50	100.0%	Instruction and 2D molecule; edited molecular structure
	Molecular optimization	Molecule	194	194	100.0%	Optimization instruction and 2D molecule; optimized structure
	Molecular property prediction	Molecule	200	200	100.0%	Property question and 2D molecule; binary property label
Cross-Turn State Maintenance	Multi-turn reasoning	Turn	215	125	58.1%	Dialogue history and applicable visual context; turn-specific answer
Total			**1,033**	**943**	**91.3%**	

molecular tasks, CHEMARENA aligns their instructions, process fields, and diagnostic metrics with Constraint Adherence.

Two-dimensional molecular input. Each molecular record is paired with a two-dimensional molecular image. The image, rather than a raw SMILES string alone, is presented as the molecular input to evaluated MLLMs. The structured molecular representation is retained internally for answer verification. All records are packaged with the textual instruction, source structure, target or label, reference rationale, and task-specific constraint fields.

Molecular normalization and deterministic verification. For tasks requiring molecular outputs, CHEMARENA extracts SMILES strings from model responses and parses them using RDKit with its default sanitization settings. A prediction is considered valid only if `Chem.MolFromSmiles` successfully returns a molecular object. For the non-stereochemical comparisons reported in this work, parsed molecules are converted to canonical, non-isomeric SMILES using `Chem.MolToSmiles(..., canonical=True, isomericSmiles=False)` before comparison. The current evaluator does not implement a separate stereochemistry-preserving comparison branch; therefore, we do not claim dedicated evaluation of stereochemistry-sensitive outputs. CHEMARENA evaluates only SMILES outputs and neither requires nor evaluates SELFIES. Depending on the task, deterministic verification additionally includes Morgan-fingerprint-based Tanimoto similarity, exact matching and soft similarity of Bemis–Murcko scaffolds, as well as property improvement and constraint satisfaction. Property evaluation uses RDKit descriptors, ESOL calculations, or the corresponding TDC property oracle, depending on the task. These derived representations are computed exclusively within the evaluator and are not provided as inputs to the evaluated models.

Task-specific construction. For molecular editing, an instruction specifies an addition, deletion, or substitution. Structured fields record the requested operation, edit cardinality, structures that must be retained, and forbidden changes. For molecular optimization, a target property is optimized subject to molecular validity and structural-preservation conditions. For molecular property prediction, the model predicts a binary property and provides an explanation grounded in task-relevant structure–property factors. Table 5 gives the exact retained subtask counts.

B.3 CHEMISTRY VISUAL REASONING

Source scope. The evidence-oriented portion is curated from publicly available Chinese high-school chemistry examination and mock-examination materials. The collection includes Gaokao-

Table 5: Molecule-centered task composition.

Task Family	Subtask	Count
Mol-Edit	add	8
	delete	19
	sub	23
Mol-Opt	drd	34
	gsk	39
	jnk	51
	logp	30
	qed	40
Mol-PP	bace	50
	clintox	50
	hiv	50
	tox21	50
Total		**444**

style and school-level assessment items. Records retain local source identifiers and image filenames when available. Because some upstream materials were collected from redistributed examination compilations rather than a single canonical publisher, source-site metadata is not complete for every item. The release should therefore distinguish records with a recoverable original citation from records with only a local provenance identifier.

Question format. All 374 instances contain answer options and are evaluated as option sets, which accommodates both single-answer and multiple-answer questions. Visual content includes molecular and crystal structures, reaction-mechanism diagrams, experimental apparatuses, process flowcharts, equilibrium plots, and electrochemical cells. An item is retained only when the designated visual is task-relevant: removing or misreading it must change the defensible answer set or leave a task-critical ambiguity unresolved.

Image handling. Visuals are placed into a common image field and checked for legibility, completeness, and consistency with the text and options. Packaging preserves chemically meaningful labels, arrows, legends, axes, and spatial relations. We do not claim that every source image was redrawn. Items for which resolution, cropping, or corruption prevents reliable interpretation are excluded.

Evidence annotations. Annotations identify visual evidence anchors, text–visual correspondences, option–evidence mappings, distractors, and the evidence required to support the reference option set. These fields enable evaluation of visual evidence utilization, cross-modal alignment, evidence–answer grounding, distractor rejection, and option-conditioned reasoning.

B.4 Multi-Turn Reasoning

Construction. The multi-turn portion is constructed and reformulated from chemical process-flow and sequential problem-solving materials. It is not obtained by simply splitting arbitrary single-turn answers. Each dialogue is designed so that at least one later turn requires information introduced, selected, revised, or corrected earlier.

State-carrying elements. State fields include molecular or ionic identities, reaction conditions, selected evidence, intermediate products, material streams, separation results, numerical quantities, and conclusions that remain active in later turns. Annotations distinguish carried state, newly introduced information, updates, corrections, superseded assumptions, and references to prior turns. This supports direct checks of history usage, state consistency, state update, reference resolution, and state persistence.

Table 6: Subtask composition of chemistry visual reasoning.

Subtask	Count
Basic chemistry experiments	23
Organic structure and properties	41
Chemical kinetics and equilibrium	27
Ionic reactions	10
Atomic, molecular, and crystal structure and properties	18
Reaction mechanism	34
Novel reaction mechanism diagram	23
Electrolyte solution	25
Electrochemical cell analysis	28
Chemical equilibrium graph analysis	24
Crystal structure and calculations	35
Rate law	10
Novel electrochemical cell analysis	36
Simplified process flowchart	29
Equilibrium constant calculations and applications	11
Total	**374**

Table 7: Multi-turn answer formats. Multi-part answers may contain more than one inner answer type; inner-type counts are therefore reported separately.

Turn Answer Type	Count	Multi-Part Inner Type	Count
Multi-part	53	Equation	35
Free text	49	Free text	34
Chemical equation	33	Formula	23
Chemical formula	30	Numerical	12
Option	19	Option	5
Numerical value	15	Electron configuration	2
Text set	6	Ion symbol	3
Electron configuration	5		
Ion symbol	4		
Formula text	1		

Scale and structure. The set contains 42 dialogues and 215 evaluated turns. Dialogue lengths range from three to seven turns: 4 dialogues have three turns, 9 have four turns, 13 have five turns, 10 have six turns, and 6 have seven turns. Of the 215 turns, 125 contain reference trajectories and structured steps; the remaining 90 are answer-only turns. The broad content distribution is 146 turns on separation or purification, 62 turns on raw-material conversion or preparation, and 7 university-style turns.

B.5 Selection and Exclusion Criteria

Candidate records are revised or removed when any of the following applies:

- the reference answer is non-unique without an explicit multi-answer specification;
- the text, options, image, or structured molecular representation are mutually inconsistent;
- the molecular structure is unparsable or violates the intended chemical specification;
- the proposed reasoning trajectory contains an unsupported chemical transition or an unresolved dependency;
- the answer requires external retrieval not supplied by the instance;
- the visual is too blurred, incomplete, or cropped to support a defensible answer;

- a dialogue turn can be answered without the history dependency it was intended to test; or
- the record does not provide an observable requirement aligned with its assigned requirement axis.

C Process Annotation and Quality Control

This section describes how task instances are converted into structured process references and how their quality is verified before evaluation. The annotation procedure preserves task-family-specific information while providing a consistent representation across the five task families. It covers the shared instance schema, candidate-trajectory generation and selection, requirement-specific checkpoint construction, annotation coverage, and human-in-the-loop verification.

The resulting reference rationales and structured task-critical checkpoints serve as observable evaluation anchors for Reference-Guided Process Alignment and provide task-specific context for Context-Enriched Reasoning Assessment. They are designed to characterize the evidence, constraints, and cross-turn states that a generated response should address, rather than to prescribe a unique reasoning path or make claims about a model's latent decision process.

C.1 Unified Instance and Annotation Schema

ChemArena normalizes its five task families into a shared conceptual representation:

$$\mathcal{I}_i = (m_i,\, q_i,\, c_i,\, v_i,\, h_i,\, y_i^*,\, p_i^*)\,, \tag{1}$$

where m_i contains task metadata, including the task family, subtask, primary observable requirement, and expected answer type; q_i is the task instruction; c_i is the textual context; v_i is the visual input; h_i is the dialogue history; y_i^* is the human-verified reference answer; and p_i^* contains the reference rationale and structured process annotations. Optional modalities are omitted when they are not applicable; ChemArena does not add synthetic placeholder content for a missing image or dialogue history.

The released JSON records instantiate this conceptual schema using task-local fields. Single-turn records contain task, subtask, idx, question, image, answer_type, gt_answer, cot, and key_annotation_steps; molecule-centered records additionally retain the normalized source representation in original_image_smiles. A multi-turn record contains overall_problem, global_image, and an ordered turns list. Each turn stores its query, optional turn-specific image, answer type, reference answer, reference rationale when available, and structured checkpoints. This serialization preserves the distinctions between task-level context, turn-level state, final outcomes, and process annotations while permitting a common evaluation interface.

ChemArena organizes annotations around three observable requirements: *Evidence Utilization*, *Constraint Adherence*, and *Cross-Turn State Maintenance*. They are operational requirements for evaluation rather than claims about a model's latent cognitive process. Alternative reasoning paths remain admissible when they satisfy the annotated evidence, chemical constraints, state dependencies, and answer support requirements.

C.2 Candidate-Trajectory Generation and Selection

Reference annotation begins only after the underlying task instance has been standardized and its reference answer has been checked. For each retained instance, general-purpose and chemistry-oriented models repeatedly generate candidate reasoning traces, producing approximately 15 candidates per instance on average. Candidate generation is used to expose plausible solution structures and common reasoning variations; a sampled trace is never accepted directly as a gold annotation.

Two chemistry-trained selectors independently conduct a two-stage screening procedure. First, each selector identifies the most suitable trace among the candidates produced by each model. Second, the selectors compare the model-level selections to identify a single raw trace for annotation. Agreement yields the selected raw trace, whereas disagreements are resolved through expert adjudication. The selectors reject traces that are redundant, excessively exploratory, internally inconsistent, chemically invalid, or incompatible with the verified answer or task requirements.

Table 8: Requirement-specific process annotations and diagnostic metrics in CHEMARENA. The structured process fields characterize observable reasoning requirements and provide context for Reference-Guided Process Alignment and Context-Enriched Reasoning Assessment.

Requirement Axis	Task Family and Topology	Structured Process Fields	Requirement-Specific Diagnostic Metrics
Evidence Utilization	Chemistry visual reasoning; an overall trajectory with global and option-specific checkpoints	Evidence anchors; relevant visual and textual cues; cross-modal entity relations; option–evidence mappings; distractors; answer-support requirements	Visual evidence utilization (Vis); cross-modal alignment (XAlign); evidence–answer grounding (EGround); distractor rejection (Dist); option-conditioned reasoning (OptR)
Constraint Adherence	Molecular editing, molecular optimization, and molecular property prediction; an overall trajectory with an ordered constraint chain	Edit targets; operation cardinality; preserved structures; forbidden changes; target properties; intermediate commitments; chemical dependencies	Structural validity (Valid); constraint satisfaction (Sat); scaffold preservation (Scaf); property–constraint compliance (Prop); chemical consistency (Chem)
Cross-Turn State Maintenance	Multi-turn reasoning; turn-level trajectories with a history-aware state graph	History references; carried state variables; updates; corrections; superseded information; final-turn consistency requirements	History usage (Hist); state consistency (StCons); state update (Upd); reference resolution (Ref); state persistence (Pers)

The selected raw trace serves only as an annotation aid. Two annotators then independently revise and structure the process reference using both the standardized task instance and the selected trace. This separation prevents model-generated wording or errors from being inherited uncritically by the benchmark.

C.3 REQUIREMENT-SPECIFIC PROCESS ANNOTATION

Each annotated evaluation unit contains a free-form reference rationale (cot) and structured task-critical checkpoints (key_annotation_steps). A checkpoint records at least a step index, semantic role, expected reasoning content, and evaluation keywords. Dependency or prerequisite links are included when the validity of a later step depends on an earlier commitment. The annotation guideline targets three to five core keywords per checkpoint; serialized records may retain additional accepted aliases when they are needed to cover chemically equivalent terminology.

The checkpoint topology is requirement-specific. Evidence Utilization uses an overall trace together with global and option-specific checkpoints. Constraint Adherence uses an overall trace and an ordered constraint chain. Cross-Turn State Maintenance uses turn-level traces connected through explicit history and state dependencies. Table 8 summarizes the corresponding fields and Stage III diagnostic metrics.

C.4 REFERENCE TRAJECTORIES

Role of the reference trajectory. A reference trajectory identifies task-critical checkpoints and their dependencies; it is not treated as the only valid surface realization of a solution. Chemically equivalent terminology, concise derivations, and alternative valid decompositions or step orderings are accepted when they satisfy the annotated requirements and prerequisite relations. ChemArena therefore evaluates observable support, validity, and consistency without claiming that a generated rationale reveals the model's latent decision process.

Reference-guided alignment. As specified in Stage II, a deterministic matcher aligns each applicable generated trace with the structured checkpoints. It uses accepted lexical variants, structured aliases, task-local matching rules, step types, and annotated prerequisite relations. At the conceptual

Table 9: Coverage of reference rationales and structured task-critical checkpoints in CHEMARENA. Traj. denotes the number of evaluation units containing both forms of process annotation. Multi-turn counts are based on turn-level evaluation units; the 215 evaluated turns originate from 42 dialogues.

Task Family	Units	Traj.	Coverage
Chemistry visual reasoning	374	374	100.0%
Molecular editing	50	50	100.0%
Molecular optimization	194	194	100.0%
Molecular property prediction	200	200	100.0%
Multi-turn reasoning	215	125	58.1%
Total	**1,033**	**943**	**91.3%**

level, checkpoint realization is recorded as *complete*, *partial*, or *missing*. In the evaluator output, the corresponding serialized labels are matched, partially_matched, and unmatched, which are mapped to 1, 0.5, and 0, respectively, when computing coverage-based summaries. These values quantify whether an annotated checkpoint is observably realized; they do not by themselves establish that the realization is chemically correct. The matcher additionally records rule-verifiable ordering, dependency satisfaction, and links between a realized checkpoint and the answer component it supports.

Coverage labels and checkpoint-local verification records are summarized by *process recall*, *rule-verifiable step correctness*, *path fidelity*, and *answer support*. Path fidelity combines ordering and dependency evidence:

$$S_{\text{path}} = 0.40S_{\text{order}} + 0.30S_{\text{dependency}} + 0.30S_{\text{prerequisite}}. \tag{2}$$

Step correctness at this stage is limited to checkpoint-local conditions that can be verified from structured rules and mappings. Lexical overlap or checkpoint coverage is not treated as proof of chemical correctness. Open-ended semantic validity and sufficiency are assessed by the human-calibrated LLM Judge in Stage III, while formally checkable molecular commitments, including molecular validity, edit satisfaction, and scaffold preservation, remain subject to deterministic chemical verification.

Alignment context. For every aligned response, ChemArena retains the matched, partially matched, and missing checkpoints; their localized response spans; verified order and dependency relations; and answer-support links. These artifacts form the *Alignment Context* supplied to Stage III. Separating these artifacts from the Stage III judgment makes it possible to distinguish missing reference support from a semantically incorrect or chemically invalid realization.

C.5 ANNOTATION COVERAGE

All 818 single-turn evaluation units contain both a reference rationale and structured checkpoints: 374 chemistry-exam instances, 50 molecular-editing instances, 194 molecular-optimization instances, and 200 molecular-property prediction instances. The multi-turn subset contains 42 dialogues and 215 turn-level evaluation units; 125 turns contain both trajectory fields. Accordingly, the reference-guided subset comprises 943 of the 1,033 evaluation units, giving an overall trajectory coverage rate of 91.3%.

The 90 multi-turn units without reference trajectories are excluded from the denominators of Stage II alignment metrics rather than assigned zero scores. They remain part of Stage I outcome evaluation and may receive applicable Stage III diagnostics when the required task and state context is available. All reported coverage and aggregate denominators therefore distinguish unannotated units from observed failures.

C.6 HUMAN-IN-THE-LOOP VERIFICATION

CHEMARENA uses a multi-stage verification procedure involving chemistry-trained annotators, including master's and doctoral students in chemistry-related disciplines at the authors' institution. The procedure covers source-instance review, candidate-trace selection, independent structured annotation, cross-checking, and expert adjudication when necessary.

Instance and answer review. Before trajectory annotation, annotators verify source completeness, reference-answer correctness, input consistency, visual–textual alignment, answer-type specification, and alignment with the intended observable requirement. Molecule-centered instances additionally undergo checks of the source structure, normalized representation, atom indexing, and consistency with the stated edit, scaffold, or property requirement. RDKit-based validity and structure checks assist this review where applicable but do not replace chemical expert judgment.

Independent trajectory construction. Following the two-selector candidate screening described in the Reference Trajectory Annotation and Quality Control section,, two annotators independently construct or revise the reference rationale and structured checkpoints. They verify that each checkpoint captures a necessary observable requirement, that its keywords and accepted variants identify the intended content, and that its dependency links do not exclude chemically valid alternative paths.

Cross-check and expert adjudication. The two independently produced annotations are then cross-checked for chemical soundness, checkpoint completeness, dependency consistency, and compatibility with the reference answer. Disagreements involving chemical interpretation, nomenclature, checkpoint granularity, prerequisite relations, or primary requirement-axis assignment are resolved through discussion and, when necessary, adjudicated by a professor or associate professor with relevant chemistry expertise. The adjudicator resolves the disputed annotations and verifies their chemical validity and internal consistency.

Revision and exclusion. An instance is revised, repackaged, or removed if review identifies an ambiguous or non-unique answer, an under-specified instruction, inconsistent text and image content, an invalid molecular structure, an unreadable visual, an unsupported chemical dependency, or insufficient alignment with its intended requirement. If the selected raw trace contains redundant exploration, unsupported claims, internal contradictions, or chemical errors, it is replaced as an annotation aid or corrected during reference construction; such content is not directly inherited by the final annotation.

Scope of reported quality-control statistics. The current project artifacts preserve exact instance counts, trajectory availability, reference rationales, and structured checkpoint fields, which support reproducible reporting of Table 9. However, early annotation and adjudication logs were not consolidated under a common record-level schema across the different checkpoint and state annotation fields. We therefore do not report a pooled inter-annotator agreement or adjudication rate, since these statistics cannot be reliably reconstructed from the available records. This reporting limitation does not imply perfect agreement and is separate from the human-scored development subset used to calibrate the Stage III Judge.

D Three-Stage Evaluation Protocol

ChemArena separates three complementary evaluation outputs: *Outcome Performance*, *Process Alignment*, and *Reasoning Quality*. They are produced by **Stage I: Task-Specific Outcome Verification**, **Stage II: Reference-Guided Process Alignment**, and **Stage III: Context-Enriched Reasoning Assessment**, respectively. Stage III additionally produces a structured *Failure Diagnosis*. The scalar outputs are retained separately so that final task success is not conflated with reference-trajectory coverage or the semantic quality of an observable reasoning artifact.

Terminology and score convention. The reported requirement axes directly operationalize *Evidence Utilization*, *Constraint Adherence*, and *Cross-Turn State Maintenance*. Unless otherwise stated, primitive metric values are normalized to $[0, 1]$ and reported on a $[0, 100]$ scale. Table 10 summarizes the role and output of each stage.

D.1 Stage I: Task-Specific Outcome Verification

Stage I evaluates all 1,033 evaluation units using task-specific outcome criteria: 374 chemistry-exam instances, 50 molecular-editing instances, 194 molecular-optimization instances, 200 molecular-property-prediction instances, and 215 evaluated turns from 42 multi-turn dialogues. Stage I uses

Table 10: Separation of responsibilities in the CHEMARENA evaluation protocol. Stage III consumes the structured contexts produced by Stages I and II but does not overwrite their scores.

Stage	Primary Mechanism	Evaluated Object	Retained Output
I. Task-Specific Outcome Verification	Deterministic parsing, answer comparison, and chemistry-tool verification	Final task output under task-family-specific success criteria	Outcome Performance and Outcome Context
II. Reference-Guided Process Alignment	Deterministic matching against fixed, human-verified checkpoints	Coverage, rule-verifiable realization, ordering, and dependencies of required process checkpoints	Process Alignment and Alignment Context
III. Context-Enriched Reasoning Assessment	Human-calibrated LLM Judge informed by Stages I and II	Semantic quality of Evidence Utilization, Constraint Adherence, or Cross-Turn State Maintenance	Reasoning Quality and Failure Diagnosis

Table 11: Meta-evaluation of candidate judge models against human annotations. P and S denote Pearson and Spearman correlations, respectively; MAE and RMSE denote mean absolute error and root mean squared error; Pair and Tie denote pairwise accuracy and tie-half accuracy; MF1 denotes approximate macro-F1; F1+ and F1- denote positive- and negative-class F1 scores; and κ denotes Cohen's kappa. Comp. is the composite agreement score used to select the primary judge model. Within each column, the best, second-best, and third-best distinct scores are indicated by **boldface**, underlining, and *italics*, respectively.

Judge Model	Correlation		Error		Preference		Classification			Agreement	Composite
	P ↑	S ↑	MAE ↓	RMSE ↓	Pair ↑	Tie ↑	MF1 ↑	F1+ ↑	F1- ↑	κ ↑	Comp. ↑
GPT-5.4-mini	0.952	0.762	**0.161**	**0.184**	0.821	0.821	**0.873**	**0.889**	**0.857**	**0.750**	**86.4**
Gemini 3 Pro	0.958	*0.714*	*0.251*	*0.314*	*0.786*	*0.786*	0.733	0.800	0.667	0.500	78.0
Gemini 3 Flash	**0.974**	**0.862**	0.410	0.448	**0.857**	**0.875**	*0.564*	*0.727*	*0.400*	*0.250*	*73.0*
Qwen3.6-Flash	0.964	0.695	0.366	0.421	0.750	0.768	*0.564*	*0.727*	*0.400*	*0.250*	69.2
GPT-5.4-nano	0.952	0.595	0.199	0.219	0.750	0.750	*0.564*	*0.727*	*0.400*	*0.250*	68.8
Qwen3.6-Plus	*0.960*	0.667	0.360	0.416	0.750	0.750	*0.564*	*0.727*	*0.400*	*0.250*	68.2

only deterministic parsing, rule-based comparison, and chemistry-tool verification; it does not use an LLM Judge.

Molecular normalization and deterministic chemical checks. For molecule-centered tasks, a generated SMILES is considered parseable only when `Chem.MolFromSmiles` succeeds under RDKit sanitization. Unless stereochemistry is task-critical, the parsed molecule is canonicalized using `Chem.MolToSmiles(..., canonical=True, isomericSmiles=False)` before comparison; task-critical stereochemical information is retained when applicable. ChemArena neither requires nor evaluates SELFIES outputs. Evaluator-internal chemical signals include Morgan fingerprints (radius 2, 1,024 bits) with Tanimoto similarity, Bemis–Murcko scaffolds, maximum-common-substructure-based soft scaffold comparison, and task-specific property oracles. These derived representations are used only for evaluation and are not supplied to the evaluated model as additional input.

Chemistry visual reasoning. The Chemistry visual reasoning family contains multi-answer questions. Let P_i and G_i denote the predicted and reference option sets for instance i. Exact option-set correctness and strict option recall are

$$s_{\mathrm{set},i} = \mathbb{I}[P_i = G_i], \tag{3}$$

$$s_{\mathrm{rec},i} = \begin{cases} |P_i \cap G_i|/|G_i|, & P_i \subseteq G_i,\ P_i \neq, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Table 12: Detailed results for Evidence Utilization. Outcome reports the Stage-I outcome score. Align. reports the Stage-II aggregate score together with process recall (Rec.), rule-verifiable step correctness (Step), path fidelity (Path), and answer support (Supp.). Quality reports the Stage-III aggregate score together with visual evidence utilization (Vis), cross-modal alignment (XAlign), evidence–answer grounding (EGround), distractor rejection (Dist), and option-conditioned reasoning (OptR). Comp. denotes the axis-level composite score. All values are percentages.

Model	Outcome	Align.					Quality						Comp.
	Score	Score	Rec.	Step	Path	Supp.	Score	Vis	XAlign	EGround	Dist	OptR	Score
GPT-5.4-xHigh	82.9	84.3	86.8	82.2	88.0	79.8	91.2	89.4	90.5	79.8	97.5	98.8	86.2
Qwen3.5-397B	88.1	84.1	83.9	79.6	90.4	83.0	91.6	90.5	90.1	83.0	95.9	98.3	88.0
Kimi-K2.6	80.3	86.9	92.7	84.7	90.0	79.1	88.2	85.4	85.0	79.1	93.8	97.7	85.2
GPT-5.4-High	80.6	85.0	88.4	82.4	88.5	80.1	84.9	79.7	78.4	80.1	87.5	98.6	83.5
Gemini 3.1 Pro	82.8	83.4	88.1	81.0	86.8	76.8	86.5	79.9	80.2	76.8	98.6	96.9	84.3
Qwen3.5-122B	86.7	84.9	88.0	82.9	87.0	81.6	90.1	89.2	88.8	81.6	94.4	96.7	87.3
Qwen3.5-27B	85.4	84.2	86.4	80.6	89.1	80.5	89.3	87.7	87.4	80.5	93.4	97.6	86.3
GPT-5.4-Medium	76.0	84.8	90.9	83.3	88.6	74.7	81.5	87.4	82.0	74.7	76.9	86.6	80.8
GPT-5.4-Low	67.5	79.7	86.6	78.9	83.7	67.5	92.4	100.0	100.0	67.5	100.0	94.4	80.0
Gemini 3 Pro	87.5	86.2	87.1	83.2	91.8	82.5	89.4	86.5	86.6	82.5	94.8	96.8	87.7
GLM-5V	73.3	77.9	85.2	73.9	83.0	68.3	86.0	87.7	86.6	68.3	90.2	97.4	79.1
Claude Sonnet 4.6	65.8	77.5	88.1	75.9	82.0	61.2	76.2	74.1	73.9	61.2	78.9	92.9	73.2
Gemini 3 Flash	62.5	75.3	86.5	73.3	78.2	60.7	74.3	65.6	64.8	60.7	87.2	93.1	70.7
Intern-S1-Pro	74.2	79.9	83.9	74.6	87.1	74.0	91.6	93.2	93.4	74.0	98.4	99.0	82.0
Intern-S1	61.9	71.6	82.3	65.2	78.5	59.4	84.2	84.1	84.1	59.4	96.3	97.3	72.7
ChemVLM-8B	16.0	50.7	76.8	39.1	59.7	24.2	45.2	30.9	25.7	24.2	60.9	84.2	37.4

Table 13: Detailed results for Constraint Adherence. Outcome reports the Stage-I outcome score. Align. reports the Stage-II aggregate score together with process recall (Rec.), rule-verifiable step correctness (Step), path fidelity (Path), and answer support (Supp.). Quality reports the Stage-III aggregate score together with structural validity (Valid), constraint satisfaction (Sat), scaffold preservation (Scaf), property–constraint compliance (Prop), and chemical consistency (Chem). Comp. denotes the axis-level composite score. All values are percentages.

Model	Outcome	Align.					Quality						Comp.
	Score	Score	Rec.	Step	Path	Supp.	Score	Valid	Sat	Scaf	Prop	Chem	Score
GPT-5.4-xHigh	73.1	58.6	64.5	58.2	69.5	38.4	78.1	98.0	71.1	69.0	74.6	77.9	70.0
Qwen3.5-397B	65.6	57.7	63.1	53.8	72.4	38.2	65.7	92.9	40.6	68.4	67.5	59.2	63.0
Kimi-K2.6	73.2	58.5	67.2	58.9	67.9	35.1	65.8	97.2	37.5	70.6	72.3	51.6	65.8
GPT-5.4-High	69.0	58.6	63.4	58.0	69.5	39.8	72.4	96.2	59.9	67.2	70.6	68.2	66.7
Gemini 3.1 Pro	63.2	52.3	61.6	52.9	61.5	28.1	66.2	95.2	50.5	66.8	59.4	58.9	60.6
Qwen3.5-122B	61.1	53.0	63.7	51.1	66.6	25.7	62.7	91.1	40.4	63.7	61.4	56.7	59.0
Qwen3.5-27B	58.4	51.9	65.9	50.2	64.4	21.2	58.0	87.6	37.2	55.2	55.6	54.2	56.1
GPT-5.4-Medium	69.0	58.4	63.6	57.5	69.5	39.5	73.0	92.6	71.9	52.0	71.6	77.0	66.9
GPT-5.4-Low	65.6	58.1	61.0	56.7	71.4	39.7	64.0	89.8	53.2	47.9	66.5	62.6	62.6
Gemini 3 Pro	75.9	54.5	57.9	52.0	66.7	38.8	73.2	99.0	52.6	78.0	77.9	58.7	67.9
GLM-5V	59.6	50.5	62.3	46.7	63.4	25.2	63.0	92.4	47.8	59.8	58.1	57.0	57.7
Claude Sonnet 4.6	66.8	46.2	56.8	44.9	57.3	21.2	63.8	95.4	49.0	51.5	67.3	55.7	59.0
Gemini 3 Flash	63.6	46.1	54.5	45.2	58.1	22.2	62.0	91.9	44.1	60.3	62.4	51.4	57.3
Intern-S1-Pro	55.7	46.5	55.4	45.0	57.9	23.4	53.5	83.8	42.6	36.3	55.6	49.2	51.9
Intern-S1	44.9	39.5	50.5	37.0	51.9	14.1	41.1	67.8	37.3	12.1	44.4	43.8	41.8
ChemVLM-8B	36.5	29.0	34.8	25.0	46.3	5.9	34.9	75.4	13.8	20.6	38.1	26.5	33.5

Thus, omitting a correct option can receive partial recall, whereas selecting any incorrect option sets strict recall to zero. The outcome score is

$$S_{\mathrm{out},i}^{\mathrm{exam}} = 100\,(0.80 s_{\mathrm{set},i} + 0.20 s_{\mathrm{rec},i})\,. \tag{5}$$

Table 14: Detailed results for Cross-Turn State Maintenance. Outcome reports the Stage-I outcome score. Align. reports the Stage-II aggregate score together with process recall (Rec.), rule-verifiable step correctness (Step), path fidelity (Path), and answer support (Supp.). Quality reports the Stage-III aggregate score together with history usage (Hist), state consistency (StCons), state update (Upd), reference resolution (Ref), and state persistence (Pers). Comp. denotes the axis-level composite score. All values are percentages.

Model	Outcome	Align.					Quality						Comp.
	Score	Score	Rec.	Step	Path	Supp.	Score	Hist	StCons	Upd	Ref	Pers	Score
GPT-5.4-xHigh	53.7	85.3	87.9	81.8	89.2	82.5	80.8	68.8	72.5	85.4	87.5	89.6	73.3
Qwen3.5-397B	53.2	86.9	88.1	80.4	94.0	86.5	87.0	67.6	69.6	100.0	97.8	100.0	75.8
Kimi-K2.6	56.4	76.5	72.6	68.8	86.4	80.8	82.5	61.5	65.4	92.9	92.9	100.0	71.9
GPT-5.4-High	54.6	79.7	74.5	71.9	90.8	84.0	80.4	67.0	71.0	86.1	88.9	88.9	71.7
Gemini 3.1 Pro	57.5	89.9	94.7	91.8	88.5	82.9	80.9	65.1	66.5	86.4	90.9	95.5	76.2
Qwen3.5-122B	57.2	80.0	81.9	77.3	85.8	74.2	82.0	65.1	68.2	92.5	86.8	97.5	73.1
Qwen3.5-27B	54.1	83.6	85.1	80.5	87.2	82.1	88.3	71.4	72.3	100.0	97.8	100.0	75.5
GPT-5.4-Medium	50.8	81.3	82.6	76.3	89.5	76.9	77.6	67.3	68.0	83.3	83.3	86.1	70.0
GPT-5.4-Low	47.3	74.6	78.2	70.5	81.2	68.1	77.3	59.0	60.1	89.5	88.2	89.5	66.5
Gemini 3 Pro	31.4	48.8	43.3	41.9	64.6	46.4	73.0	57.7	60.9	82.1	82.1	82.1	51.3
GLM-5V	53.4	74.3	64.1	61.8	91.9	83.8	81.2	64.7	66.2	90.0	90.0	95.0	69.7
Claude Sonnet 4.6	51.0	75.6	75.9	70.1	84.0	73.0	64.5	61.5	60.1	66.7	56.2	77.8	63.7
Gemini 3 Flash	47.2	75.8	76.1	67.0	86.9	75.0	74.2	55.4	54.8	87.5	85.7	87.5	65.8
Intern-S1-Pro	44.7	71.6	73.7	65.1	82.0	65.9	62.0	54.5	55.7	64.3	65.0	70.5	59.5
Intern-S1	47.8	72.1	77.8	65.9	81.7	62.1	69.1	56.8	57.6	73.8	71.1	86.4	63.0
ChemVLM-8B	20.0	36.9	49.4	28.2	49.4	18.5	21.9	26.9	24.1	16.7	16.7	25.0	26.2

Table 15: Performance of 16 MLLMs across ChemArena's three requirement axes. Outcome, Align., and Quality denote Outcome Performance, Process Alignment, and Reasoning Quality, respectively. Comp. is their weighted composite within each requirement axis, and Overall is the mean of the three axis-level composite scores. All values are percentages.

Model	Evidence Utilization				Constraint Adherence				Cross-Turn State Maintenance				Overall
	Outcome	Align.	Quality	Comp.	Outcome	Align.	Quality	Comp.	Outcome	Align.	Quality	Comp.	
GPT-5.4-xHigh	82.9	84.3	91.2	86.2	73.1	**58.6**	**78.1**	**70.0**	53.7	85.3	80.8	73.3	**76.5**
Qwen3.5-397B	**88.1**	84.1	91.6	**88.0**	65.6	57.7	65.7	63.0	53.2	86.9	87.0	75.8	75.6
Kimi-K2.6	80.3	**86.9**	88.2	85.2	73.2	58.5	65.8	65.8	56.4	76.5	82.5	71.9	74.3
GPT-5.4-High	80.6	85.0	84.9	83.5	69.0	**58.6**	72.4	66.7	54.6	79.7	80.4	71.7	74.0
Gemini 3.1 Pro	82.8	83.4	86.5	84.3	63.2	52.3	66.2	60.6	**57.5**	**89.9**	80.9	**76.2**	73.7
Qwen3.5-122B	86.7	84.9	90.1	87.3	61.1	53.0	62.7	59.0	57.2	80.0	82.0	73.1	73.1
Qwen3.5-27B	85.4	84.2	89.3	86.3	58.4	51.9	58.0	56.1	54.1	83.6	**88.3**	75.5	72.6
GPT-5.4-Medium	76.0	84.8	81.5	80.8	69.0	58.4	73.0	66.9	50.8	81.3	77.6	70.0	72.5
GPT-5.4-Low	67.5	79.7	**92.4**	80.0	65.6	58.1	64.0	62.6	47.3	74.6	77.3	66.5	69.7
Gemini 3 Pro	87.5	86.2	89.4	87.7	**75.9**	54.5	73.2	67.9	31.4	48.8	73.0	51.3	69.0
GLM-5V	73.3	77.9	86.0	79.1	59.6	50.5	63.0	57.7	53.4	74.3	81.2	69.7	68.9
Claude Sonnet 4.6	65.8	77.5	76.2	73.2	66.8	46.2	63.8	59.0	51.0	75.6	64.5	63.7	65.3
Gemini 3 Flash	62.5	75.3	74.3	70.7	63.6	46.1	62.0	57.3	47.2	75.8	74.2	65.8	64.6
Intern-S1-Pro	74.2	79.9	91.6	82.0	55.7	46.5	53.5	51.9	44.7	71.6	62.0	59.5	64.5
Intern-S1	61.9	71.6	84.2	72.7	44.9	39.5	41.1	41.8	47.8	72.1	69.1	63.0	59.2
ChemVLM-8B	16.0	50.7	45.2	37.4	36.5	29.0	34.9	33.5	20.0	36.9	21.9	26.2	32.4

Molecular editing. Generated SMILES are parsed and canonicalized with RDKit. The requested-edit score s_{edit} verifies the annotated functional-group transition: addition requires one new target group, deletion requires removal of one target group, and substitution requires the specified removal and addition simultaneously. Let s_{scaf} denote soft Bemis–Murcko scaffold preservation and s_{sim} denote Morgan-fingerprint Tanimoto similarity to the source molecule. The outcome is

$$S_{\text{out}}^{\text{edit}} = 100\,(0.75 s_{\text{edit}} + 0.15 s_{\text{scaf}} + 0.10 s_{\text{sim}})\,. \tag{6}$$

An invalid or unparsable molecular output receives zero on validity-dependent components. Requirements not exhausted by this local outcome check, such as edit cardinality and preservation of non-target substituents, remain available to the constraint-oriented process analysis.

Molecular optimization. Let $f_p(\cdot)$ be the property oracle for optimization objective p and $\Delta_p = f_p(\hat{m}) - f_p(m_{\text{src}})$. The implementation records $s_{\text{success}} = \mathbb{I}[\Delta_p > 0]$ and $s_{\text{best}} = \mathbb{I}[\Delta_p \geq \tau_p]$, corresponding to the implementation fields `success_rate` and `best_rate`, respectively. Here, $\tau_p = 0.3$ for QED, DRD2, JNK3, and GSK3B, and $\tau_p = 0.5$ for LogP. The other components are soft scaffold preservation s_{scaf}, RDKit parsing success $s_{\text{valid_smiles}}$, and availability of a valid property-oracle score $s_{\text{valid_score}}$:

$$S_{\text{out}}^{\text{opt}} = 100\big(0.40 s_{\text{success}} + 0.25 s_{\text{best}} + 0.20 s_{\text{scaf}} + 0.10 s_{\text{valid_smiles}} + 0.05 s_{\text{valid_score}}\big). \tag{7}$$

Property improvement and scaffold preservation remain separate signals; positive improvement alone is not interpreted as satisfaction of every structural constraint.

Molecular property prediction. Predictions are normalized to the task's binary label space and compared with the reference label:

$$S_{\text{out},i}^{\text{mpp}} = 100\,\mathbb{I}[\hat{y}_i = y_i]. \tag{8}$$

Missing or unrecognized labels are treated as incorrect predictions.

Multi-turn reasoning. Each turn is evaluated by a deterministic parser selected according to the annotated answer type. Yes/no, single-option, and numerical answers use exact matching. Option-set and text-set answers combine exact-set correctness and strict recall with weights $0.80/0.20$. Chemical formulas, ion symbols, formula text, and electron configurations combine exact matching and parse validity with weights $0.95/0.05$; chemical equations use exact matching and equation-component F1 with weights $0.90/0.10$. Name-text and free-text answers combine exact match, token F1, character similarity, and valid-prediction status with weights $0.75/0.15/0.05/0.05$. Multi-part answers are evaluated component-wise with the parser specified for each part, and the component scores are averaged.

The Cross-turn State Maintenance outcome is the mean over the 215 evaluated turns:

$$S_{\text{out}}^{S} = \frac{1}{215} \sum_{i=1}^{215} S_{\text{out},i}^{\text{turn}}. \tag{9}$$

Final-turn accuracy, follow-up-turn accuracy, valid-prediction rate, and dialogue completion are retained as auxiliary statistics. Dialogue completion does not replace turn-level correctness: a dialogue is counted as complete only when every evaluated turn attains a score of at least 0.999.

Axis-level outcome aggregation. Stage-I scores follow the task-to-axis organization:

$$S_{\text{out}}^{E} = \frac{1}{374} \sum_{i \in \mathcal{D}_{\text{exam}}} S_{\text{out},i}^{\text{exam}}, \tag{10}$$

$$S_{\text{out}}^{C} = \frac{1}{444} \sum_{i \in \mathcal{D}_{\text{edit}} \cup \mathcal{D}_{\text{opt}} \cup \mathcal{D}_{\text{mpp}}} S_{\text{out},i}, \tag{11}$$

$$S_{\text{out}}^{S} = \frac{1}{215} \sum_{i \in \mathcal{D}_{\text{turn}}} S_{\text{out},i}^{\text{turn}}. \tag{12}$$

Constraint Adherence is therefore instance-weighted across the three molecule-centered families rather than computed as an unweighted mean of their task-family averages.

Outcome Context. In addition to the scalar score, Stage I retains an auditable *Outcome Context*: the raw and normalized prediction, selected parser, parsing status, matched and unmatched answer components, molecular canonicalization results, applicable structural or property checks, and turn-wise answer states. This context allows later stages to interpret the reasoning in light of what the model produced. Stages II and III cannot modify the deterministic Stage-I outcome.

D.2 Stage II: Reference-Guided Process Alignment

Stage II aligns a generated reasoning trace with a fixed set of task-critical checkpoints from the human-verified reference annotation. It uses *deterministic matching rules rather than the LLM Judge*. Each checkpoint specifies accepted lexical variants or structured aliases, its step type, prerequisites, and any applicable visual-evidence or dialogue-history requirement. The matcher locates candidate response spans and records complete, partial, or missing realization together with rule-verifiable ordering and dependency relations.

The alignment topology follows the task family. Chemistry visual reasoning uses an overall-plus-options structure; molecule-centered tasks use a constraint or structure–property dependency chain; and multi-turn reasoning uses a turn-state graph. Reference trajectories provide evaluation anchors rather than a canonical surface script. Alternative terminology, concise derivations, and alternative orderings are admissible when they satisfy the annotated requirements and dependencies.

The structured alignment record is summarized by four dimensions: *Process Recall*, *Step Correctness*, *Path Fidelity*, and *Answer Support*. Process Recall measures coverage of required checkpoints. Step Correctness is restricted to checkpoint realizations whose correctness can be verified by the fixed matching specification and task-local rules; broader semantic validity remains a Stage-III question. Path Fidelity evaluates rule-verifiable ordering and dependencies, and Answer Support measures whether the matched checkpoints establish the annotated support required for the final response. Their weighted summary is

$$S_{\text{align}} = 0.25S_{\text{recall}} + 0.30S_{\text{step}} + 0.25S_{\text{path}} + 0.20S_{\text{support}}. \tag{13}$$

Path Fidelity is computed as

$$S_{\text{path}} = 0.40S_{\text{order}} + 0.30S_{\text{dep}} + 0.30S_{\text{prereq}}. \tag{14}$$

Matched and missing checkpoints, localized spans, and verified relations are retained as the *Alignment Context* for Stage III. These records localize relevant reasoning evidence; lexical or structural coverage alone is not treated as proof of chemical correctness, causal faithfulness, or the model's latent decision process.

Stage II is computed only for units with applicable process annotations. All 818 single-turn records and 125 of the 215 multi-turn evaluation units contain reference trajectories, yielding 943 annotated units in total (91.3%). Unannotated turns are excluded from the relevant denominator rather than assigned a zero alignment score.

Implementation note. Some task-local artifacts retain a legacy convenience field named `final_cot_score` from an earlier generic process summary. This field is not the reported Process Alignment score and is not used in the requirement-axis composite. All supplementary alignment results are recomputed from the four displayed components according to Eq. 13.

D.3 Stage III: Context-Enriched Reasoning Assessment

Stage III evaluates semantic properties that cannot be exhaustively reduced to answer matching or deterministic checkpoint rules. The human-calibrated LLM Judge receives the original task instance, reference answer and process annotations, model response, *Outcome Context*, and *Alignment Context*. Formally verifiable commitments remain the responsibility of deterministic tools. For example, RDKit parsing, molecular validity, requested-edit completion, scaffold checks, and computable property constraints are supplied to the Judge as context rather than replaced by its opinion.

The assessment proceeds through three logical passes: *Task Reconstruction*, *Reasoning Audit*, and *Failure Diagnosis*. The output of each pass is appended to the shared evaluation context and informs the next pass.

Task Reconstruction. Evidence Utilization is reconstructed as an evidence–option map linking decisive visual or textual observations to option-level conclusions. Constraint Adherence is reconstructed as a contract containing the intended transformation, protected structures, forbidden changes, and target-property requirements. Cross-Turn State Maintenance is reconstructed as an expected trajectory of state preservation, update, rejection, and correction. The result is retained as the *Task Reconstruction Context*.

Reasoning Audit. The Judge evaluates the model response against the reconstructed task representation rather than rewarding surface similarity to the reference. The requirement-specific diagnostic metrics are:

- **Evidence Utilization:** Visual Evidence Utilization, Cross-Modal Alignment, Evidence–Answer Grounding, Distractor Rejection, and Option-Conditioned Reasoning.
- **Constraint Adherence:** Structural Validity, Constraint Satisfaction, Scaffold Preservation, Property-Constraint Compliance, and Chemical Consistency.
- **Cross-Turn State Maintenance:** History Usage, State Consistency, State Update, Reference Resolution, and State Persistence.

Explicit corrections, when required by a dialogue, are evaluated within State Update and State Persistence rather than reported as a separate scalar metric. For a dialogue containing T evaluated turns, the state audit expands to T sequential judgments, yielding $1+T+1$ Judge calls across Task Reconstruction, Reasoning Audit, and Failure Diagnosis. The resulting step-level judgments, dependency assessments, unsupported claims, contradictions, and state inconsistencies form the *Reasoning Audit Context*.

Failure Diagnosis. The final pass seeks the earliest consequential *observable* failure: a broken evidence path, a violated constraint or invalid structural transition, or a divergent dialogue state. It records the unmet requirement, its location, and its downstream effects. When localization is ambiguous, the report preserves that uncertainty. Failure Diagnosis is a human-readable explanatory output, not an additional scalar score and not a claim about the latent causal mechanism of a model decision.

Judge Calibration. The primary Judge is selected through meta-evaluation against human annotations on a calibration subset sampled from the evaluation instances. As detailed in the Judge Calibration and Reliability Section, candidate judges are compared using multiple agreement metrics covering score consistency, preference agreement, and categorical agreement. GPT-5.4-mini achieves the highest composite agreement score and is therefore used as the primary Judge for the reported Stage-III results.

Reasoning Quality aggregation. For each requirement axis a, diagnostic metric m is first averaged over the units to which it applies:

$$\bar{S}_m^{(a)} = \frac{1}{|\mathcal{D}_{a,m}|} \sum_{i \in \mathcal{D}_{a,m}} S_{i,m}, \tag{15}$$

where $\mathcal{D}_{a,m}$ is the applicable evaluation subset. The reported Reasoning Quality score is the arithmetic mean across the requirement-specific metric set $\mathcal{M}_a$:

$$S_{\text{quality}}^{(a)} = \frac{1}{|\mathcal{M}_a|} \sum_{m \in \mathcal{M}_a} \bar{S}_m^{(a)}. \tag{16}$$

Records marked `not_required` are excluded from the corresponding metric denominator. Missing Judge outputs are not converted to zero; they are rerun when possible and otherwise remain missing until the evaluation is complete.

D.4 Hierarchical Aggregation

The three stage outputs remain visible as *Outcome*, *Align.*, and *Quality* in the main results. For compact requirement-axis comparison, they are combined using

$$S_{\text{comp}}^{(a)} = 0.33 S_{\text{out}}^{(a)} + 0.33 S_{\text{align}}^{(a)} + 0.34 S_{\text{quality}}^{(a)}. \tag{17}$$

The near-equal allocation gives comparable importance to task completion, reference-guided alignment, and requirement-specific semantic assessment while summing exactly to one. Each component is first aggregated over its applicable units; unavailable fields are not silently imputed.

The reported Overall score is the macro-average of the three requirement-axis composite scores:

$$S_{\text{overall}} = \frac{S_{\text{comp}}^{(\text{Evidence})} + S_{\text{comp}}^{(\text{Constraint})} + S_{\text{comp}}^{(\text{State})}}{3}. \tag{18}$$

This hierarchy prevents metric-count bias: a requirement axis does not receive greater weight merely because it contains more task families, evaluation units, or diagnostic metrics. The component and requirement-specific results remain available alongside Overall so that aggregate ranking does not replace ChemArena's intended fine-grained diagnosis.

E Judge Calibration and Reliability

Formally verifiable chemical properties are evaluated independently of the LLM judge through deterministic verification procedures. The LLM judge is only applied to context-enriched reasoning assessment, where evaluation requires interpreting evidence utilization, reasoning consistency, and the relationship between generated responses and task-specific reference information. Therefore, selecting a reliable judge model is important to reduce additional evaluation bias.

E.1 Human-Annotated Calibration Subset

Before full-scale evaluation, we perform judge calibration on a human-annotated development subset that is excluded from the final benchmark reporting. The calibration subset covers representative chemical task families, including chemistry visual reasoning, molecular editing, molecular optimization, molecular property prediction, and multi-turn chemical reasoning.

Each calibration instance contains the complete evaluation context, including the original task input, chemical figures when available, reference answer, reference reasoning information, and the response generated by the evaluated model.

Eight annotators with chemistry backgrounds independently evaluate each response according to four aspects: reasoning support, evidence interpretation, chemical consistency, and overall process quality. Each response is assigned an ordinal score ranging from 1 to 5.

For each response instance i, the aggregated human score is computed as:

$$h_i = \frac{1}{M} \sum_{m=1}^{M} h_i^{(m)}, \tag{19}$$

where $M = 8$ denotes the number of human annotators and $h_i^{(m)}$ represents the score assigned by annotator m for instance i.

The averaged human score is used as the reference signal for continuous agreement evaluation. For binary discrimination evaluation, the averaged score is converted into a binary quality label using a threshold of 3.5:

$$y_i = \begin{cases} 1, & h_i \geq 3.5, \\ 0, & h_i < 3.5. \end{cases}$$

A label of 1 represents a relatively high-quality response, whereas 0 represents an insufficient response.

E.2 Candidate Judge Models

We compare multiple multimodal judge models under the same evaluation setting. Each candidate judge receives identical task information, including the input question, associated chemical images, reference information, and the evaluated model response.

For a candidate judge model J, the predicted scores are denoted as:

$$J = [j_1, j_2, \ldots, j_N],$$

where N is the number of calibration instances and j_i is the score assigned by the judge model.

The reliability of each candidate judge is evaluated from four perspectives: continuous score agreement, ranking consistency, scoring error, and binary quality discrimination.

E.3 Continuous Score Agreement

Pearson correlation measures the linear consistency between judge scores and human scores:

$$\rho_P = \frac{\sum_i (j_i - \bar{j})(h_i - \bar{h})}{\sqrt{\sum_i (j_i - \bar{j})^2}\sqrt{\sum_i (h_i - \bar{h})^2}}. \tag{20}$$

To evaluate ranking consistency, Spearman correlation is computed as Pearson correlation between the ranked judge scores and ranked human scores:

$$\rho_S = \rho_P(r(J), r(H)),$$

where $r(\cdot)$ denotes the ranking operation.

This formulation naturally handles tied ordinal scores commonly produced by 1–5 human annotations.

E.4 Scoring Error

The absolute deviation between judge scores and human scores is measured by mean absolute error (MAE):

$$MAE = \frac{1}{N}\sum_i |j_i - h_i|, \tag{21}$$

and root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_i (j_i - h_i)^2}. \tag{22}$$

Lower MAE and RMSE indicate closer agreement with human scoring.

E.5 Ranking Consistency

Since benchmark analysis often requires comparing model quality rather than only predicting absolute scores, we additionally evaluate pairwise ranking consistency.

For each response pair (i, k) with different human scores ($h_i \neq h_k$), the predicted preference is compared with the human preference:

$$sign(j_i - j_k) \quad \text{vs.} \quad sign(h_i - h_k).$$

Pairwise accuracy is defined as:

$$\text{PairAcc} = \frac{\sum_{i<k,\ h_i \neq h_k} \mathbb{I}[\text{sign}(j_i - j_k) = \text{sign}(h_i - h_k)]}{|\{(i,k) : i < k,\ h_i \neq h_k\}|}. \tag{23}$$

To additionally consider judge predictions with identical scores, we report tie-half accuracy. A correct preference receives a score of 1, a predicted tie receives 0.5, and an incorrect preference receives 0:

$$TieAcc = \frac{N_{correct} + 0.5N_{tie}}{N_{valid}}, \tag{24}$$

where N_{valid} denotes the number of evaluated response pairs.

E.6 Binary Quality Discrimination

To evaluate whether a judge can distinguish high-quality and low-quality responses, judge scores are converted into binary predictions using the same 3.5 threshold:

$$\hat{y}_i = \begin{cases} 1, & j_i \geq 3.5, \\ 0, & j_i < 3.5. \end{cases}$$

Precision and recall are computed for both positive and negative classes:

$$Precision = \frac{TP}{TP + FP}, \tag{25}$$

$$Recall = \frac{TP}{TP + FN}. \tag{26}$$

The class-specific F1 scores are:

$$F1^{+} = \frac{2Precision^{+}Recall^{+}}{Precision^{+} + Recall^{+}}, \tag{27}$$

$$F1^{-} = \frac{2Precision^{-}Recall^{-}}{Precision^{-} + Recall^{-}}. \tag{28}$$

The macro-F1 score is calculated as:

$$MF1 = \frac{F1^{+} + F1^{-}}{2}. \tag{29}$$

E.7 Agreement Beyond Chance

Beyond classification accuracy, we further measure agreement between judge predictions and human-derived labels using Cohen's kappa:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{30}$$

where p_o denotes the observed agreement between judge and human labels, and p_e denotes the expected agreement by chance.

E.8 Composite Judge Selection Score

Since individual metrics capture complementary aspects of judge reliability, we combine them into a unified composite score. All positively oriented metrics, including Pearson correlation (P), Spearman correlation (S), pairwise accuracy (Pair), tie-aware accuracy (Tie), macro-F1 (MF1), positive- and negative-class F1 ($F1^{+}$ and $F1^{-}$), and Cohen's κ, are already reported on a common $[0, 1]$ scale and are therefore used directly.

MAE and RMSE are negatively oriented and are converted to higher-is-better scores. Because the human rating scale ranges from 1 to 5, the maximum possible rating difference is four. We therefore define

$$z_{\mathrm{MAE}} = 1 - \frac{\mathrm{MAE}}{4}, \qquad z_{\mathrm{RMSE}} = 1 - \frac{\mathrm{RMSE}}{4}.$$

Let

$$\mathcal{M}^{+} = \{P, S, \mathrm{Pair}, \mathrm{Tie}, \mathrm{MF1}, F1^{+}, F1^{-}, \kappa\}.$$

The final composite score is computed by equally weighting the ten direction-aligned metrics:

$$\mathrm{Comp} = \frac{100}{10}\left(\sum_{m \in \mathcal{M}^{+}} m + z_{\mathrm{MAE}} + z_{\mathrm{RMSE}}\right). \tag{31}$$

A higher composite score indicates stronger overall consistency with human evaluation. The candidate with the highest composite score is selected as the primary judge model for full-scale evaluation, while an auxiliary judge is retained for sensitivity analysis.

E.9 Final Judge Selection

Table 11 summarizes the meta-evaluation results. Although Gemini-3-flash achieves the strongest ranking consistency, reflected by the highest Pearson, Spearman, Pair Accuracy, and Tie-half Accuracy, it exhibits larger scoring deviations and weaker binary discrimination capability. In contrast, GPT-5.4-mini achieves the lowest MAE and RMSE, the highest MF1, and the strongest agreement in binary classification. It also obtains the highest composite score among all candidate judges.

Therefore, GPT-5.4-mini is selected as the primary judge model for Context-Enriched Reasoning Assessment. Gemini-3-flash is retained as an auxiliary judge for ranking sensitivity analysis to verify that benchmark conclusions are not dependent on a single evaluator.

F Additional Model Results

F.1 Full Process and Diagnostic Breakdown

Tables 12, 13, and 14 report the individual measurements used to construct Process Alignment and Reasoning Quality for Evidence Utilization, Constraint Adherence, and Cross-Turn State Maintenance, respectively. In each table, Out. denotes Outcome Performance from Stage I. Rec., Step, Path, and Supp. denote the four reference-guided dimensions summarized as Process Alignment in Stage II: process recall, rule-verifiable step correctness, path fidelity, and answer support. The remaining columns report the requirement-specific diagnostic metrics used to derive Stage III Reasoning Quality with the calibrated GPT-5.4-mini Judge.

A score of 100 on an individual diagnostic metric indicates that every applicable evaluation case for that metric is satisfied. It does not imply perfect performance on the corresponding requirement axis, since each metric captures only one aspect of the evaluated reasoning requirement. Requirement-axis conclusions therefore rely on the hierarchical composite rather than any single process or diagnostic submetric.

F.2 Requirement-level Performance Profiles

To further examine whether outcome-level performance sufficiently characterizes model behavior, we visualize the relationship between Outcome Performance and process-level evaluation signals for each requirement axis. Specifically, we compare Outcome Performance with Process Alignment and Reasoning Quality, respectively. These profiles provide a diagnostic perspective by revealing whether successful task outcomes are accompanied by reliable observable reasoning behaviors.

For each requirement axis, the profile space is divided according to the mean values of Outcome Performance and the corresponding process metric. The upper-right region represents models that achieve both strong task outcomes and reliable process behaviors, indicating that the final answers are supported by requirement-consistent reasoning. In contrast, models in the lower-right region achieve relatively high outcome performance but weaker process signals, revealing cases where correct answers may not be supported by sufficient evidence utilization, constraint adherence, or cross-turn state maintenance. Such cases are difficult to identify through outcome-only evaluation.

Models in the upper-left region exhibit relatively reliable process behaviors but limited task success, suggesting that satisfying intermediate reasoning requirements does not always guarantee successful

task completion. The lower-left region corresponds to models with limitations in both outcome performance and observable reasoning behaviors.

The three profiles further illustrate the diagnostic role of ChemArena's requirement axes. The Evidence Utilization profile examines whether models can effectively connect relevant multimodal evidence with final answers, rather than relying on unsupported recognition or shortcuts. The Constraint Adherence profile reveals whether chemically valid outputs are obtained through consistent preservation of task-specific chemical requirements. The Cross-Turn State Maintenance profile evaluates whether models can preserve and update chemically relevant states across interactions while producing correct responses.

By jointly analyzing outcome and process-level signals, these profiles separate models with similar task success but different observable failure patterns. Therefore, they complement the aggregate ranking by providing requirement-specific evidence for diagnosing where and why models fail under different chemical reasoning scenarios. Figures 9, 10, and 11 present the requirement-level performance profiles for Evidence Utilization, Constraint Adherence, and Cross-Turn State Maintenance, respectively.

F.3 Axis-Level Aggregation

Table 15 reports all three evaluation stages, the axis-level composite and the overall score from Eq. 18. These values provide a compact audit trail from the three-stage protocol to the headline summary. The best result in each component column is bolded.

G Qualitative Diagnostic Analysis

G.1 Evidence Utilization Failure Cases

Although multimodal models can achieve high answer accuracy, they may fail to effectively utilize available visual or textual evidence during reasoning. Such failures cannot be identified through answer-only evaluation, because the final prediction may remain correct despite missing or weak evidence support. Figure 4 presents a representative example where the model reaches the correct answer but fails to establish the necessary connection between visual observations and the final decision.

Figure 5 further illustrates another type of Evidence Utilization failure, where the model recognizes relevant information from the input but does not incorporate it into subsequent reasoning steps. These cases demonstrate that ChemArena can distinguish between successful answer generation and reliable evidence utilization, revealing failures that would otherwise remain hidden under conventional accuracy-based evaluation.

G.2 Constraint Adherence Failure Cases

Final molecular validity does not necessarily indicate successful adherence to task-specific constraints. A model may generate chemically valid outputs while violating required transformations, preserving incorrect structures, or failing to satisfy optimization objectives.

Figure 6 shows a representative Constraint Adherence failure. Although the generated molecule appears chemically reasonable, the model violates the explicit transformation constraints defined by the task. This example highlights that ChemArena evaluates not only whether a valid molecular output is produced, but also whether the required chemical constraints are consistently maintained throughout the generation process.

G.3 Cross-Turn State Maintenance Failure Cases

Multi-turn chemical reasoning requires models to preserve and update previously established information across dialogue turns. However, a model may generate locally plausible responses while discarding important intermediate states or previously introduced chemical conditions.

Figure 7 presents a representative Cross-Turn State Maintenance failure. The model correctly interprets parts of the current question but fails to carry forward the chemical state established in earlier turns, resulting in an inconsistent final reasoning trajectory. This case demonstrates the importance of evaluating state maintenance beyond single-turn correctness.

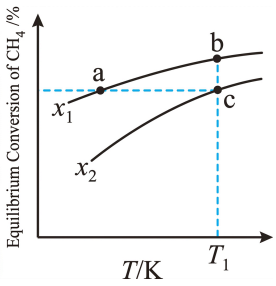
G.4 Reasoning Effort Analysis

We further investigate how reasoning effort influences observable reasoning behaviors under the ChemArena evaluation framework. Unlike conventional accuracy-based analysis, ChemArena distinguishes final answer correctness from process reliability and reasoning quality.

Figure 8 presents a qualitative comparison of different reasoning-strength configurations of GPT models on the same chemical reasoning instance. The low-effort model produces an incorrect prediction due to early chemical misinterpretation, while stronger reasoning configurations progressively recover relevant structural evidence and improve reasoning consistency. However, even the strongest configuration may contain localized reasoning errors that do not affect the final answer, demonstrating the value of fine-grained diagnostic evaluation beyond outcome accuracy.

ID: Evidence-229

MLLM：Kimi-K2.6



Question： mol of CH_4 and a certain amount of H_2O are introduced into a constant-volume sealed container, where the reaction $CH_4(g) + H_2O(g) \rightleftharpoons CO(g) + 3H_2(g)$ occurs. The equilibrium conversion of CH_4 as a function of temperature for different feed ratios x (x = $n(CH_4)/n(H_2O)$) is shown in the figure.Which of the following statements are correct?
A. The equilibrium concentration of $H_2O(g)$ at point a is greater than at point c.
B. The forward reaction rate at point b is greater than at point c.
C. The reaction is spontaneous at any temperature.
D. The average molar mass of the gas mixture at point b is greater than at point a.

Reference Answer： A,B

Model Answer： **A,B** ✓

Reference Trace

Option A
"The upper curve corresponds to more water, so $x_1 < x_2$... **equal conversion means equal water consumption...** $c(H_2O)_a > c(H_2O)_c$. A is correct."

Option B
"At b and c, temperature is identical... $x_1 < x_2$ means **more initial H_2O at b... higher reactant concentration gives** $v_b > v_c$. B is correct."

Global Step
"Conversion increases with temperature... the forward reaction is endothermic... b, c share T_1, while a, c share the same conversion."

Option C
"$\Delta H > 0$ and $\Delta S > 0$... $\Delta G < 0$ **only above a sufficiently high temperature**. C is incorrect."

Option D
"From a to b, conversion and total gas moles increase... total mass is conserved... $\bar{M}_b < \bar{M}_a$. D is incorrect..."

Model Response Trace

Option A ✓
"a and c have the same conversion... a initially contains more water... $c(H_2O)_a > c(H_2O)_c$. A is correct."

Option B ✗
"At the same temperature, K is identical... **larger** $[CO][H_2]^3$ **implies larger** $[CH_4][H_2O]$... **therefore** $v_b > v_c$. B is correct."

Global Step ✓
"More H_2O increases CH_4 conversion... therefore $x_1 < x_2$."

Option C ✓
"Conversion increases with temperature... $\Delta H > 0, \Delta S > 0$... **spontaneous only at sufficiently high temperature**. C is incorrect."

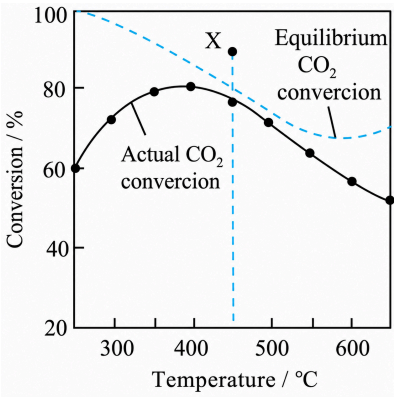
Option D ✓
"Higher conversion increases total gas moles... $\bar{M}_b < \bar{M}_a$. D is incorrect"

Failure Diagnose：

The final answer is correct, but Option B contains an unsupported kinetic leap:
the model uses the equilibrium expression to infer reactant concentrations and then directly concludes $v_b > v_c$, implicitly treating the overall equilibrium reaction as a rate law.

Figure 4: A representative failure case of **Evidence Utilization**. The model obtains the correct final answer but fails to incorporate decisive visual evidence from the provided chemical diagram. Although the response identifies relevant chemical concepts, the reasoning process does not explicitly connect the observed visual information with the final option selection, leading to an unsupported conclusion. This case demonstrates that outcome correctness alone does not guarantee effective utilization of multimodal evidence.

ID: Evidence-019 *MLLM Kimi-K2.6*



Question The main reactions in the CO_2 hydrogenation to CH_4 process are:$CO_2(g) + 4H_2(g) \rightarrow CH_4(g) + 2H_2O(g)$, $\Delta H = -164.7$ kJ·mol^{-1};$CO_2(g) + H_2(g) \rightarrow CO(g) + H_2O(g)$, $\Delta H = +41.2$ kJ·mol^{-1}.In a closed container at 1.01×10^5 Pa with initial $n(CO_2):n(H_2) = 1:4$, the equilibrium CO_2 conversion and the actual CO_2 conversion versus temperature are shown in the figure. Which of the following statements are correct

A. The enthalpy change ΔH for the reaction $2CO(g) + 2H_2(g) = CO_2(g) + CH_4(g)$ is -247.1 kJ·mol^{-1}.

B. The equilibrium selectivity of CH_4 decreases as the temperature increases, due to the exothermic nature of methanation and endothermic nature of the reverse water-gas shift reaction.

C. The equilibrium constant K for the side reaction $CO_2(g) + H_2(g) \rightleftharpoons CO(g) + H_2O(g)$ decreases as the temperature increases.

D. At 450°C, increasing the system pressure or increasing the initial molar ratio $n(H_2)/n(CO_2)$ can both increase the equilibrium conversion rate of CO_2.",

Reference Answer A,B,D

Model Answer **A,B,D** ✓

Reference Trace

Option A
… The target reaction is $2CO + 2H_2 \rightarrow CO_2 + CH_4$. **This can be derived as Reaction 1 - 2 * Reaction 2. ΔH = -164.7 - 2(41.2) = -164.7 - 82.4 = -247.1 kJ/mol.** The statement matches the calculation.",

Option B
The figure shows that equilibrium CO_2 conversion … higher temperature favors CO formation and suppresses CH_4 formation. Therefore,CH_4 selectivity decreases as temperature increases.

Global Step
Analysis of the Target Image reveals a plot of CO_2 conversion versus temperature, distinguishing between actual conversion (solid line, …Therefore, A, B, and D are the correct statements.

Option C:
The side reaction is $CO_2(g) + H_2(g) \rightleftharpoons CO(g) + H_2O(g)$ with $\Delta H = +41.2$ kJ/mol…The statement claims K decreases,…Thus, the statement is incorrect.",

Option D:
The main reaction $CO_2 + 4H_2 \rightleftharpoons CH_4 + 2H_2O$ … raise the equilibrium conversion rate, potentially reaching the level of point X. The statement is correct.",

Model Response Trace

Option A: ✓
…The target reaction is $2CO + 2H_2 \rightarrow CO_2 + CH_4$. … ΔH = -164.7 - 2(41.2) = -164.7 - 82.4 = -247.1 kJ/mol. The statement matches the calculation.",

Option B: ✗
Selectivity is defined as $n(CH_4)$ / $n(CO_2$ reacted). …**Both factors contribute to a decrease in CH_4 selectivity as temperature rises.** The statement is correct.",

Global Step ✓
…increasing the H_2/CO_2 ratio drives the reaction forward. Both actions increase equilibrium conversion, supporting Option D. Therefore, A, B, and D are the correct statements.

Option C: ✓
…Since the reaction is endothermic ($\Delta H > 0$), increasing the temperature will shift the equilibrium to the right,… Thus, the statement is incorrect.",

Option D: ✓
…Increasing the ratio of H_2/CO_2 (adding reactant H_2) also shifts equilibrium to the right, increasing CO_2 conversion. …The statement is correct.",

Failure Diagnose

The final answer is correct, but Option B contains an unsupported kinetic inference:
the model uses equilibrium composition as a proxy for reaction rate and fails to incorporate the visual evidence from the temperature–conversion diagram

Figure 5: Another example of **Evidence Utilization** failure. The model recognizes part of the visual information but fails to use the identified evidence in the subsequent reasoning chain. The missing evidence-to-answer connection results in an unsupported inference, despite producing a plausible final response. This example highlights the ability of ChemArena to identify failures where models perceive relevant information but do not effectively utilize it for decision making.

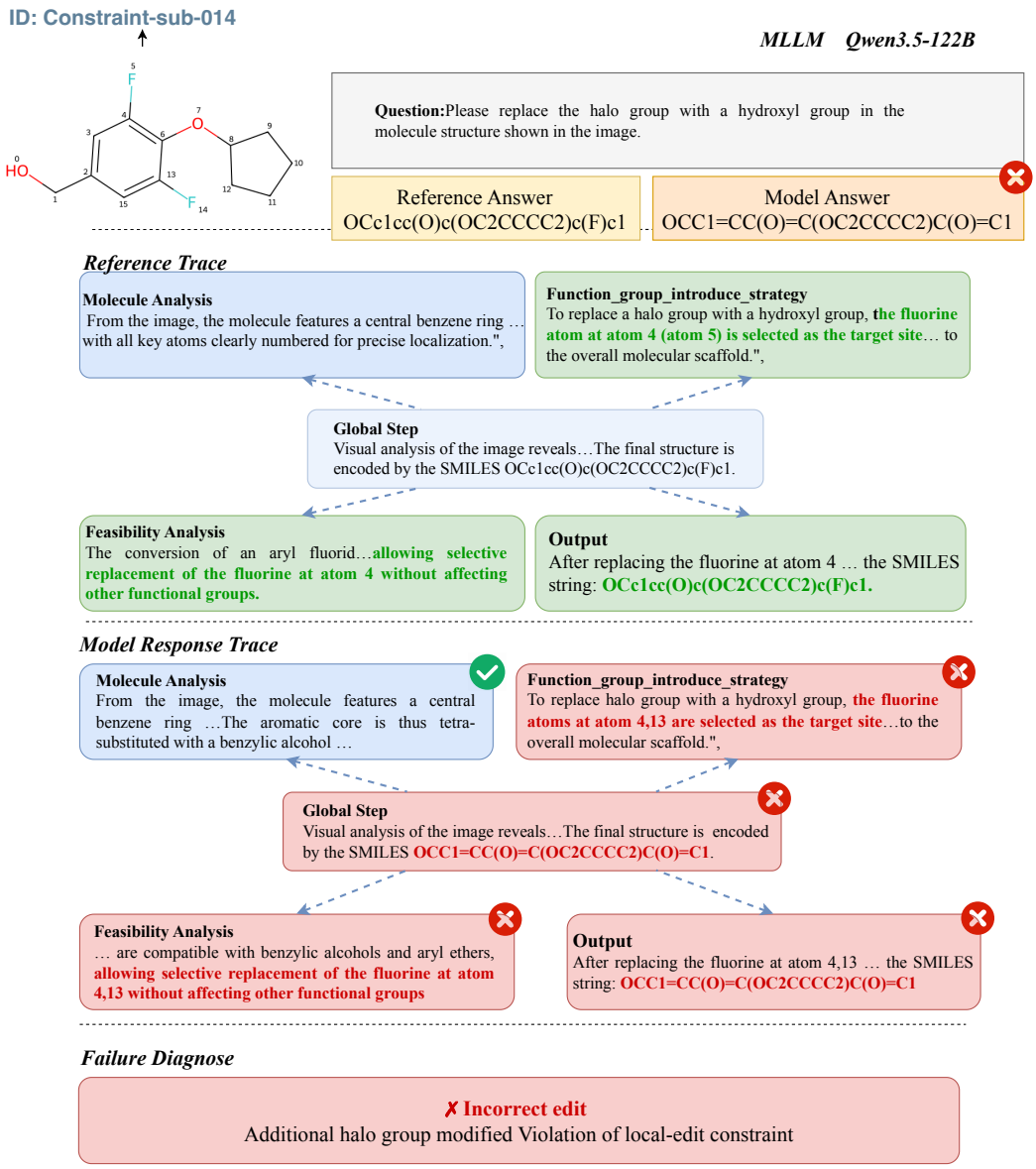


Figure 6: A representative failure case of **Constraint Adherence**. The model generates a chemically plausible molecular output but violates task-specific structural constraints defined by the instruction. The failure is localized to the incorrect transformation or constraint propagation step rather than the final molecular validity alone. This case illustrates that valid molecular structures do not necessarily imply successful constraint satisfaction.

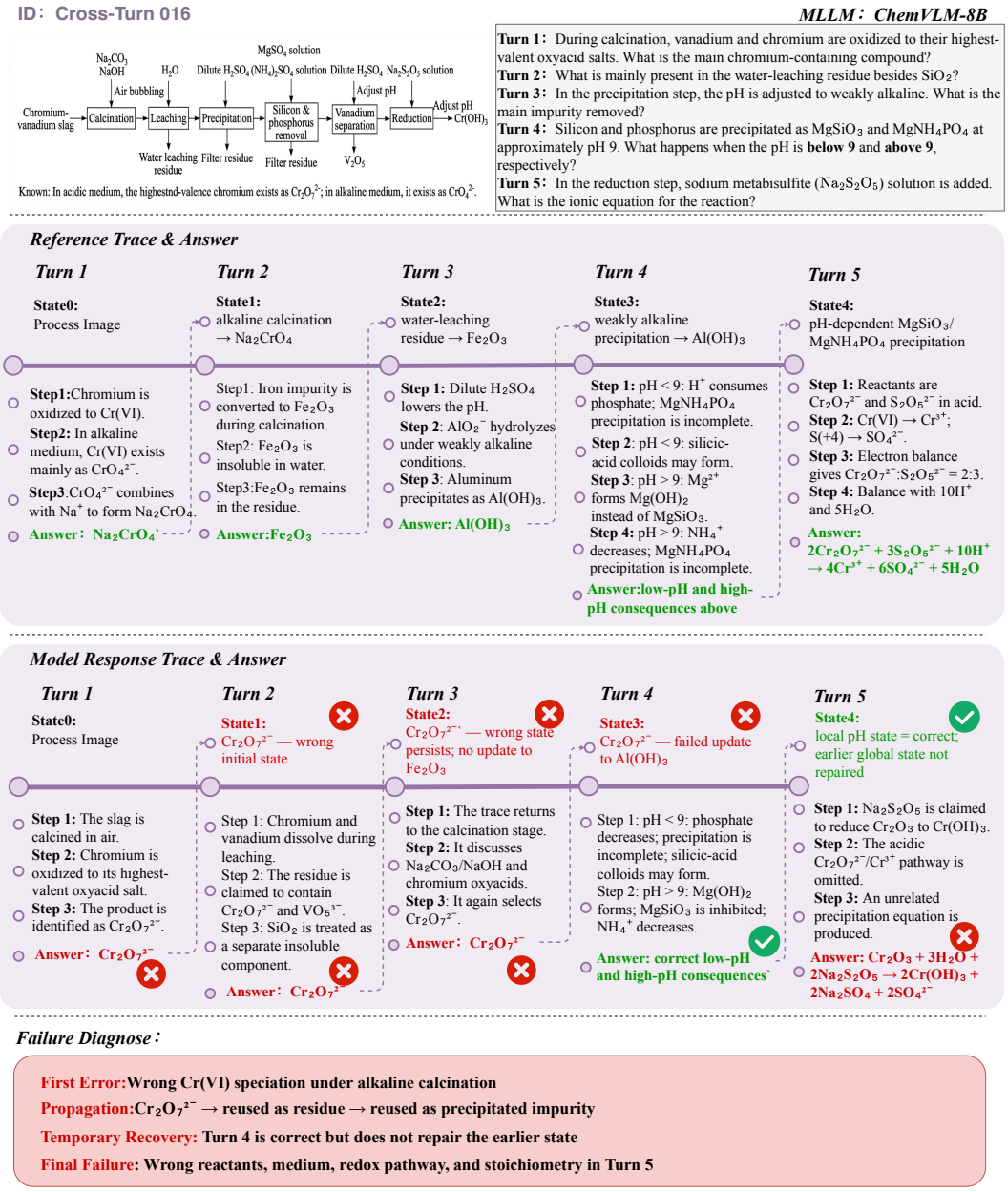


Figure 7: A representative failure case of **Cross-Turn State Maintenance**. The model produces a chemically plausible response in the final turn but fails to preserve the chemical state established in previous dialogue turns. The error occurs when previously introduced species, conditions, or intermediate commitments are discarded or replaced by an unsupported state, causing inconsistency in the final reasoning process. This case demonstrates the importance of evaluating state maintenance beyond single-turn answer correctness.

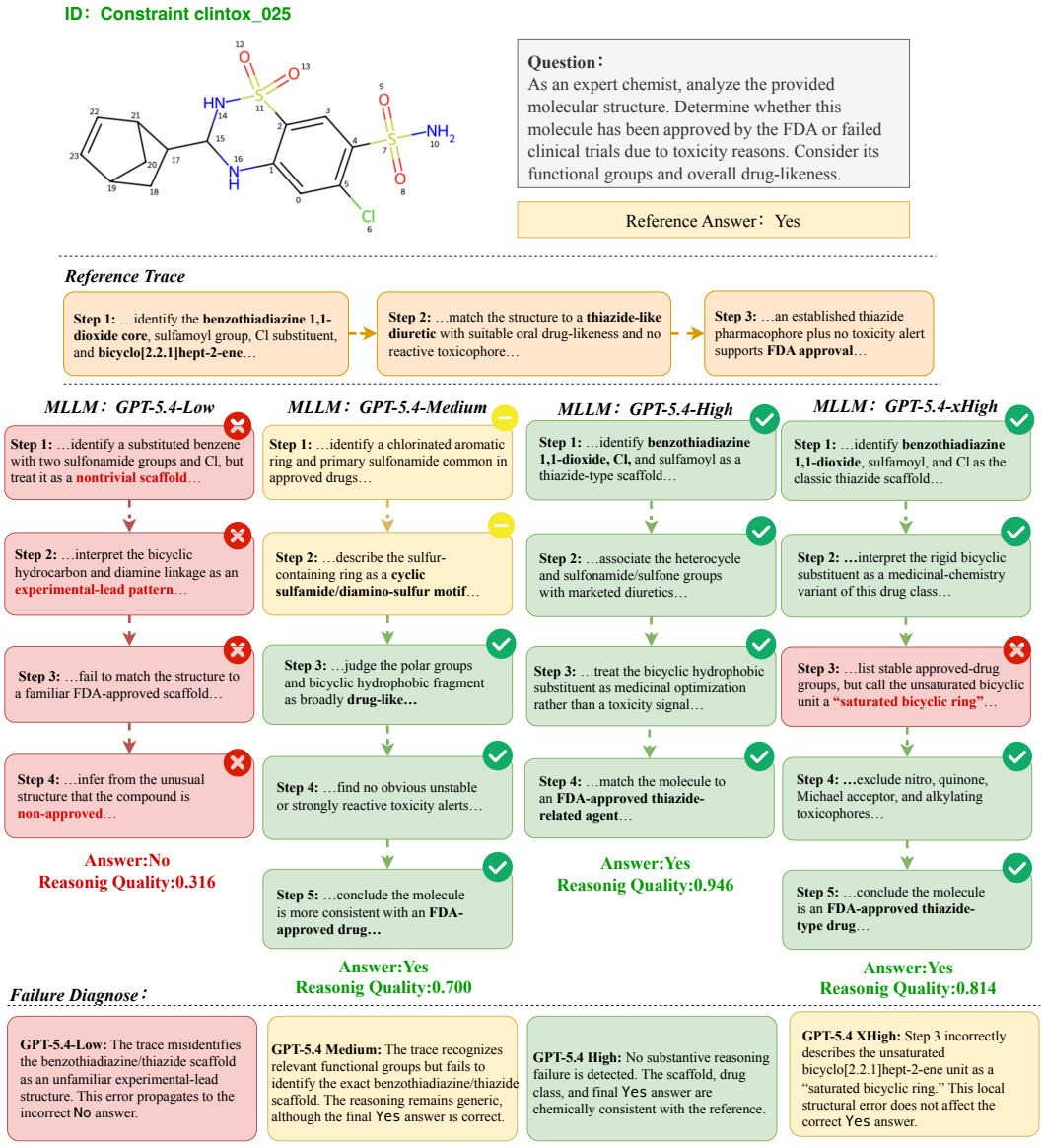


Figure 8: Qualitative analysis of reasoning effort using ChemArena diagnosis. We compare different reasoning-strength configurations of GPT models on the same chemical reasoning instance. Increasing reasoning effort generally helps recover relevant structural evidence and improves the completeness of observable reasoning behaviors. However, stronger reasoning configurations do not completely eliminate localized chemical errors, demonstrating that answer correctness alone cannot characterize reasoning reliability.

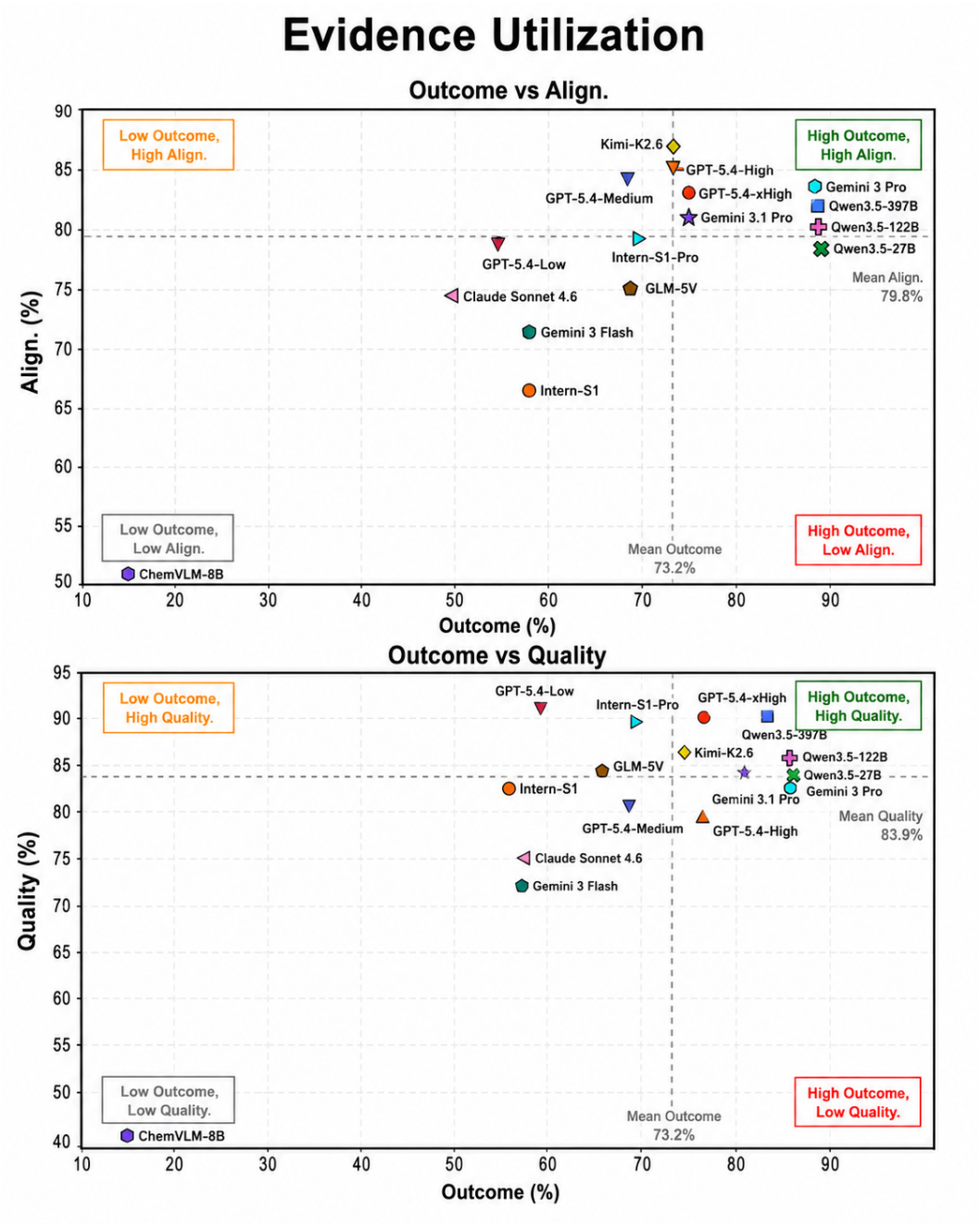


Figure 9: Performance profile of multimodal large language models on the **Evidence Utilization** requirement axis. The upper panel shows the relationship between Outcome Performance and Process Alignment, while the lower panel compares Outcome Performance and Reasoning Quality. Dashed lines indicate the corresponding mean values. Each quadrant highlights different combinations of task success and reasoning quality, revealing whether models achieve correct answers through reliable evidence utilization.

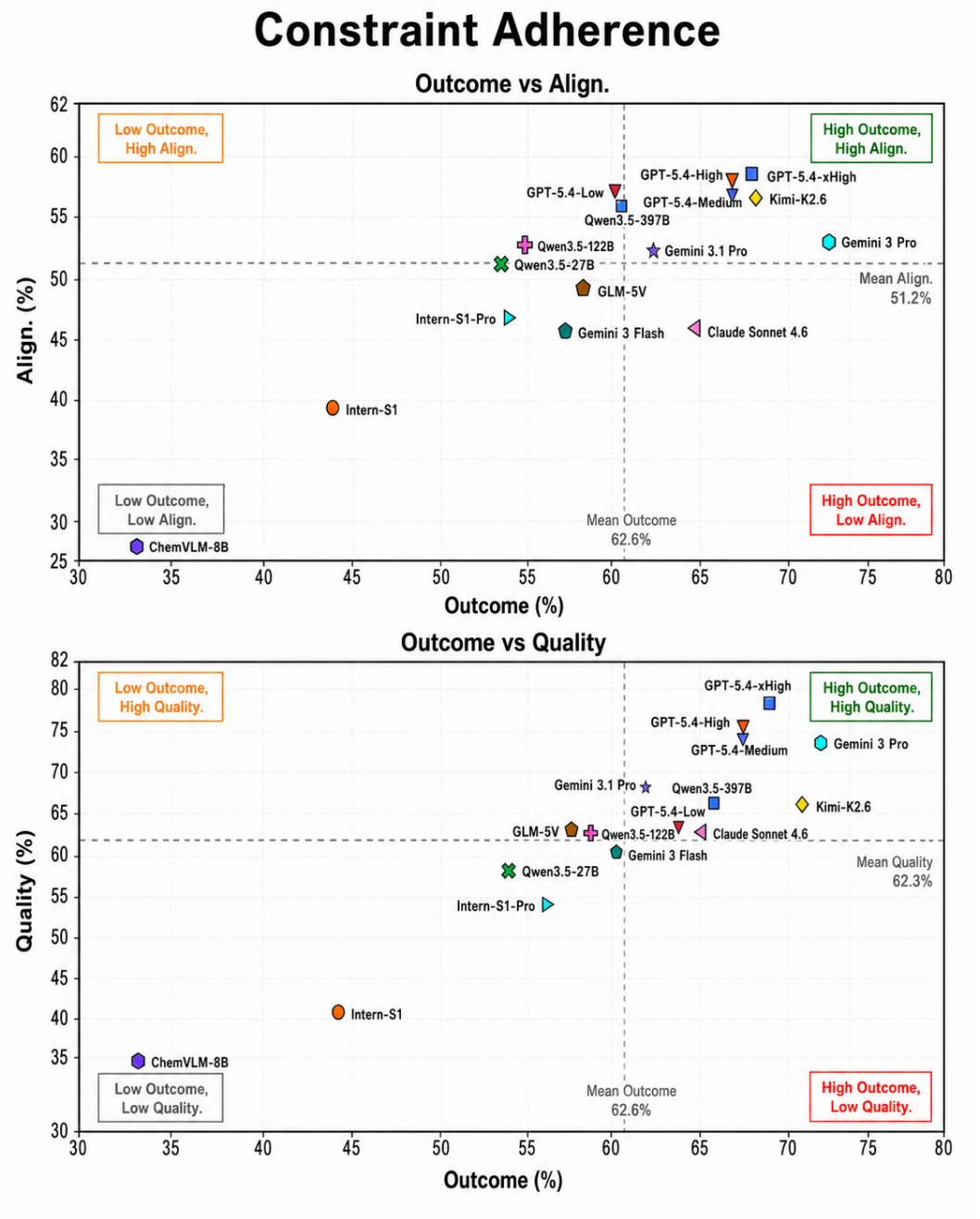


Figure 10: Performance profile of multimodal large language models on the **Constraint Adherence** requirement axis. The upper panel compares Outcome Performance with Process Alignment, and the lower panel compares Outcome Performance with Reasoning Quality. The visualization reveals differences between obtaining correct molecular outcomes and maintaining chemical, structural, and task-specific constraints throughout the generated solution process.

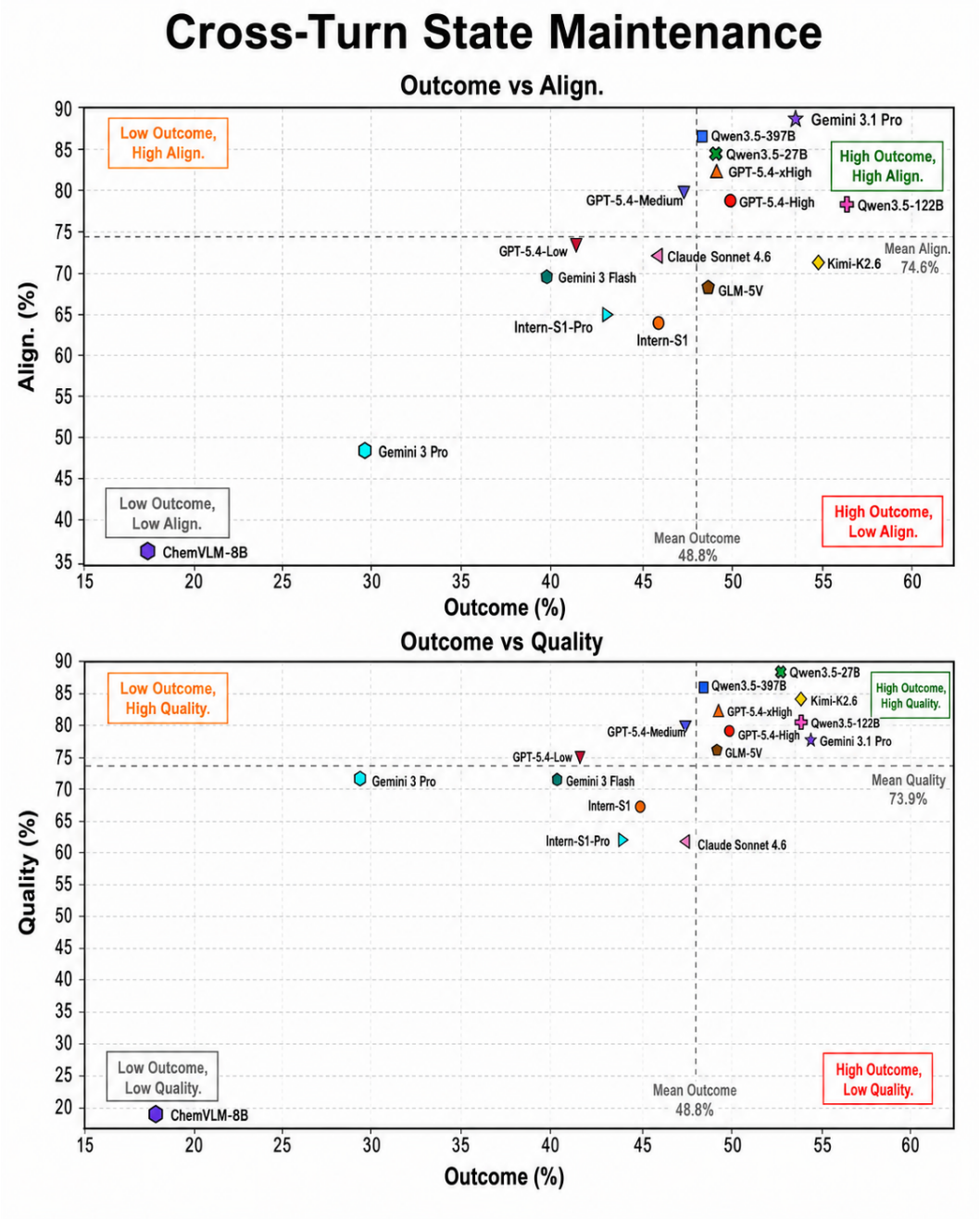


Figure 11: Performance profile of multimodal large language models on the **Cross-Turn State Maintenance** requirement axis. The upper panel presents Outcome Performance versus Process Alignment, and the lower panel presents Outcome Performance versus Reasoning Quality. The profiles characterize whether models can preserve and update previously established chemical states across multiple conversational turns while producing correct final answers.