CHEMDIALOG: BENCHMARKING CHEMICAL REACTION UNDERSTANDING IN LLMS THROUGH DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the promise of large language models (LLMs) for chemical reaction understanding, current benchmarks rely heavily on question-answering (QA) for isolated tasks, such as molecule property prediction and stoichiometry. However, real-world chemistry applications require LLMs to engage in interactive dialogue and integrate information across related tasks. To address this gap, we introduce ChemDialog, a reaction-grounded dialogue benchmark for evaluating LLMs' ability to integrate chemical information across dialogue turns and related tasks. Specifically, we construct ChemDialog by organizing chemical reaction subtasks into coherent dialogue scenarios with LLM assistance, followed by extensive human review for quality control. We further develop a multidimensional evaluation framework that separately assesses task-specific chemical competence and dialogue-level interaction logic. Extensive experimental results indicate that mainstream LLMs still struggle with logical consistency and dynamic information integration in multi-turn chemical dialogues, and providing multi-turn, related task dialogue history significantly enhances model performance in synthesis tasks. We hope that ChemDialog and our findings encourage the research community to further advance the untapped potential of LLMs toward complex, real-world chemical interactions. The dataset and code will be released.

1 INTRODUCTION

Recent advances in artificial intelligence have expanded its application to specialized domains, including scientific research (Sun et al., 2024; Shojaee et al., 2025), healthcare (Jin et al., 2019; Liu et al., 2025a), and math (Liang et al., 2024; Cobbe et al., 2021). In chemistry, large language models (LLMs) are increasingly expected to support complex research tasks through multi-turn interaction. However, existing benchmarks for chemical LLMs predominantly evaluate isolated tasks in a single-turn question-answering (QA) format. Consequently, they provide limited assessment of models' ability to adapt during interaction and integrate chemical information across dialogue turns. This evaluation gap hinders the systematic development of LLMs for interactive chemistry applications.

Existing chemical benchmarks, *e.g.*, ChemLLMBench (Guo et al., 2023) and SuperChem (Zhao et al., 2025a), primarily consist of single-turn QA pairs, as seen in Fig. 1(a). Specifically, ChemLLMBench (Guo et al., 2023) comprises eight static chemical tasks, while SuperChem (Zhao et al., 2025a) advances this by integrating step-by-step reasoning. However, these benchmarks possess three critical limitations: ❶ **Limited evaluation of cross-task information integration**:they treat chemistry as static and one-shot problems, thus provide limited insight into whether models can integrate complementary chemical information from other chemical tasks to support coherent problem solving. ❷ **Insufficient evaluation of interaction logic**: they provide limited assessment of whether models can correctly interpret and respond to context-dependent interaction logic, such as error correction and reliability verification; ❸ **Lack of explicit multi-factual chemical evaluation**: they often rely on fixed answer-matching criteria, limiting their ability to assess diverse yet chemically valid model responses.

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.

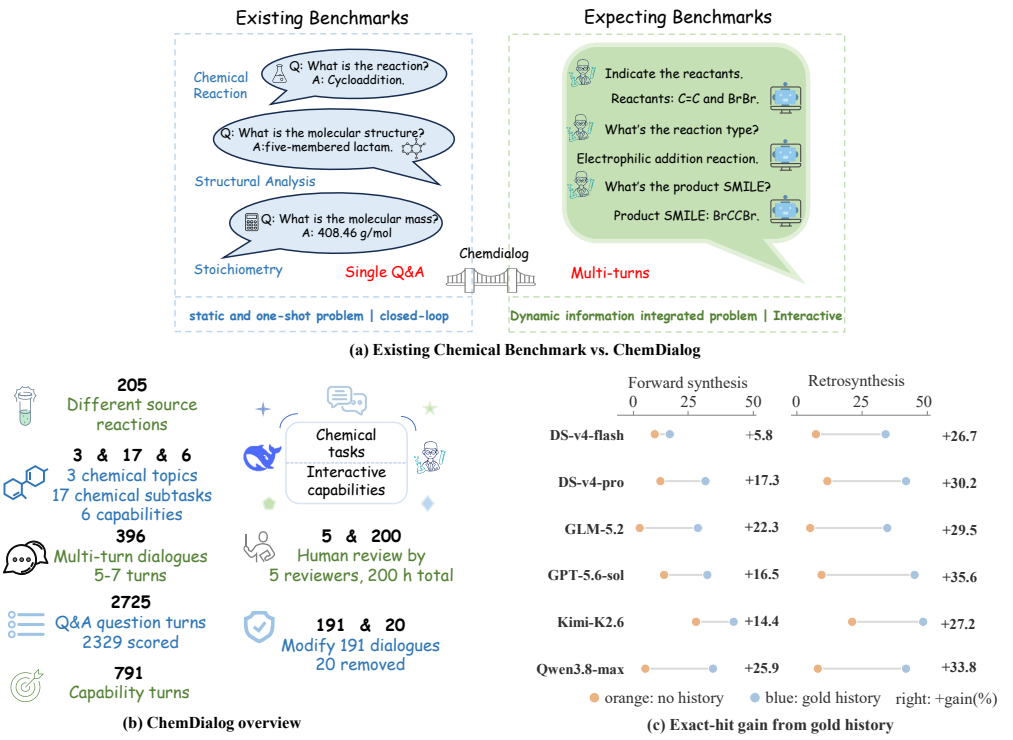


Figure 1: Overview of ChemDialog. (a) Existing benchmarks ask isolated single-turn questions on separate tasks. ChemDialog links questions on one reaction into a multi-turn dialogue. Later turns depend on answers from earlier turns. (b) Dataset statistics: source reactions, task topics, subtasks, capabilities, dialogues, turns, and human review. (c) Effect of gold history on the final structure question for six models (Table 2). Forward synthesis uses the exact product-SMILES hit. Retrosynthesis uses the exact reactant-set hit.

To address these limitations, we propose ChemDialog, a benchmark designed to evaluate chemical understanding in LLMs under complex interactive scenarios, as seen in Fig. 1(b). ChemDialog is constructed through a two-stage pipeline. First, we define a set of logical structures that organize reaction-related topics into interconnected chemical tasks. For each reaction, an LLM selects and arranges chemical tasks randomly derived from the logic structures to construct a coherent multi-turn dialogue. In these dialog, as shown in Fig. 1(a), information derived from earlier tasks serves as context for solving subsequent ones. This design addresses ❶ by enabling the evaluation of both basic chemical reaction understanding and the ability to dynamically integrate information throughout a dialog. Second, we define six types of interaction logic and prompt the LLM to adaptively determine which interaction pattern to incorporate and where to introduce it based on the evolving dialogue context. By explicitly incorporating interaction patterns, ChemDialog goes beyond static question answering and evaluates whether LLMs can respond effectively to diverse interaction dynamics and iterative feedback while maintaining chemical accuracy and conversational coherence. To ensure the quality of the generated dialogues, we conduct an extensive human review, as described in § A.7.

To address ❸, we develop a format-aware evaluation framework to accommodate the diverse response formats arising from multi-turn chemical dialogues. Specifically, we first constrain the expected response format for each task and adopt different evaluation strategies according to the output type. For structured outputs with deterministic answers, such as SMILES representations and numerical values, we employ *Exact Matching (EM)* to ensure rigorous and objective assessment. For open-ended textual responses, where semantically equivalent answers may exhibit substantial linguistic variation, we adopt an *LLM-as-a-Judge* approach to evaluate response correctness and quality. Rather than applying a universal judging criterion, we further design task-specific evaluation rubrics tailored to different chemical tasks and interaction logic, enabling the judge to account

for their distinct requirements. We also compare LLM judgments with human evaluation in § A.7, and observe strong agreement between them.

Extensive experiments on mainstream LLMs reveal two main findings. First, current models perform poorly on ChemDialog, demonstrating substantial limitations in both chemical-reaction understanding and interaction-dependent reasoning and leaving considerable room for improvement. Second, correct information from earlier dialogue turns is critical for solving subsequent chemical tasks, substantially improving performance on both forward synthesis and retrosynthesis, as shown in Fig. 1(c). However, the magnitude of these improvements varies considerably across models and between the two tasks, revealing important differences in how effectively models use cross-task information. This variation highlights the importance of evaluating cross-task information integration across both models and chemical tasks, as enabled by ChemDialog. The main contributions of this paper are summarized as follows:

- We present ChemDialog, a human-verified benchmark of 396 chemical reaction dialogues spanning 2,725 turns and 205 source reactions. By covering a wide range of chemical subtasks with intricate interactive logic, ChemDialog moves beyond traditional static benchmarks and enables the systematic evaluation of cross-task information integration and interaction-dependent reasoning in multi-turn chemical dialogues.
- We propose a multi-dimensional evaluation scheme that spans both chemical proficiency and interactive comprehension, utilizing specialized metrics to quantify how effectively models integrate domain knowledge with multi-turn conversational logic.
- We evaluate a range of mainstream LLMs on ChemDialog and characterize their capabilities and remaining challenges in multi-turn chemical dialogues. Our empirical analyses further examine how models use information from earlier turns across different chemical tasks, providing insights that may inform future research on LLMs for chemistry.

2 Related Work

2.1 Dialogue-based Benchmarks

Multi-turn evaluation tests whether a model can retain instructions, resolve references, revise earlier claims, and use information introduced across an interaction rather than answer isolated prompts. MT-Bench introduced multi-turn model comparison alongside LLM-based judging, while MT-Eval and MT-Bench-101 developed more explicit capability taxonomies for dialogue history and fine-grained turn-level behaviour (Zheng et al., 2023; Kwan et al., 2024; Bai et al., 2024a). Multi-IF evaluates multilingual instruction following across turns, and MultiChallenge targets instruction retention, inference-memory interactions, and knowledge retrieval in realistic conversations (He et al., 2024; Deshpande et al., 2025). Recent work extends interactive evaluation into specialised domains. MathChat studies iterative mathematical problem solving and instruction adherence; SciCode evaluates the integration of scientific knowledge into research code; and task-oriented medical dialogue evaluates domain reasoning through interactive information exchange (Liang et al., 2024; Tian et al., 2024; Liu et al., 2025a). ChemDialog brings this interaction-centred evaluation to chemical reactions. Each constructed dialogue is grounded in a reaction record, and later questions depend on structures, conditions, procedural roles, or corrections established earlier.

2.2 Chemical-related Benchmarks

Chemical LLM evaluation spans broad knowledge coverage, quantitative chemistry, molecular structure, and reaction reasoning. ChemLLMBench and ChemEval collect diverse chemistry tasks, while ChemBench examines chemical knowledge and reasoning across chemistry and materials questions (Guo et al., 2023; Huang et al., 2026; Mirza et al., 2025). QCBench focuses on quantitative chemistry; ChemIQ assesses chemical intelligence through structure-sensitive problems; and FGBench targets functional-group-level molecular reasoning (Xie et al., 2025; Runcie et al., 2025; Liu et al., 2025b). MaCBench introduces multimodal chemistry and materials questions, whereas ChemTables evaluates semantic interpretation of tables from chemical patents (Alampara et al., 2025; Zhai et al., 2021). Other benchmarks place greater emphasis on compositional or process-level reasoning. MolPuzzle evaluates multimodal molecular structure elucidation, modular chemical

operations probe reasoning beyond conventional chemical QA, and SUPERChem combines multimodal inputs with process-oriented chemistry questions (Guo et al., 2024; Li et al., 2025a; Zhao et al., 2025a). These resources provide broad coverage of chemical knowledge and reasoning, mostly through individual questions or independently scored instances. ChemDialog adds multi-turn state integration, measuring how chemical content is recovered and revised across three reaction topics and six interaction capabilities.

3 CHEMDIALOG BENCHMARK

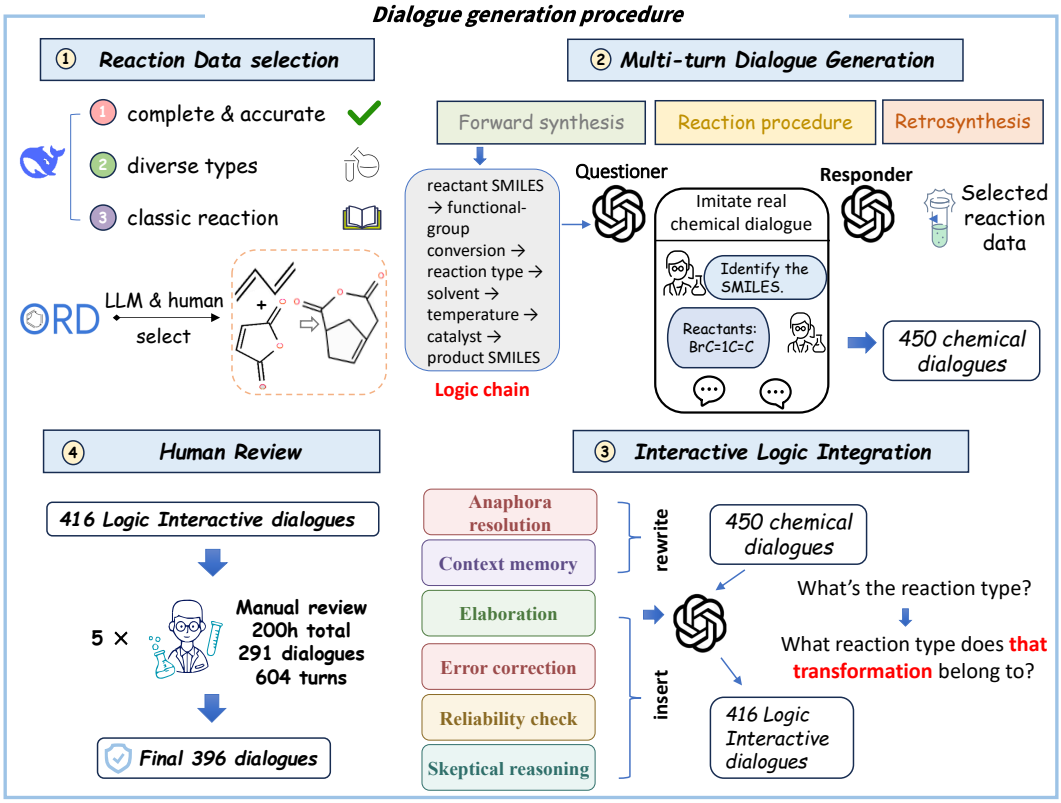


Figure 2: ChemDialog construction pipeline contains four stages. First, diverse and classic chemical reactions, accompanied by rich metadata, are selected as the basis for dialogue generation. Second, an LLM automatically generates dialogues from the selected reaction metadata according to predefined logical structures. Third, guided by the dialogue context and selected interaction logic, an LLM randomly generate interactive turns for each dialogue by rewriting or inserting operations. Finally, manual review is conducted to ensure the quality of the dataset.

3.1 DATASET CONSTRUCTION PIPELINE

To the best of our knowledge, no existing dataset focuses on multi-turn dialogue grounded in chemical reaction records. The absence of such a dataset makes it difficult to evaluate LLMs' ability to adapt during interactions and integrate chemical information across dialogue turns. Therefore, we present ChemDialog, a multi-turn dialogue dataset that integrates both chemical and interactive tasks, enabling a more comprehensive evaluation of LLMs in complex conversational settings. ChemDialog construction pipeline, as shown in Figure 2, consists of four stages: reaction data selection, multi-turn dialogue generation, interactive logic integration, and human review.

Reaction Data Selection. ChemDialog is constructed from source reactions collected from the Open Reaction Database (Kearnes et al., 2021). To ensure that the generated dialogues are grounded in reliable reaction data while covering a broad range of chemical scenarios, we screen the source

records based on three criteria: 1) complete & accurate: each selected record should contain sufficient and reliable information; 2) diverse types: the selected reactions should cover a broad range of reaction types and mechanisms, and 3) classic reaction: we prioritize representative reactions that are well established in the literature and commonly encountered in chemical research. An LLM is first employed for coarse screening, followed by manual review to further verify the quality of the selected reactions. This process results in 205 high-quality source reactions covering diverse reaction types and commonly used chemical transformations.

Multi-turn Dialogue Generation. To move beyond isolated question answering, we organize each selected reaction into a logically connected multi-turn dialogue. The dialogue design enables models to be evaluated across related chemical tasks rather than on each task in isolation. ChemDialog covers three chemical topics: ***forward synthesis (FS)***, ***reaction procedure (RP)***, and ***retrosynthesis (RS)***. For each topic, we define a logic chain of five to seven turns that follows its task structure. For example, the FS chain progresses from reactant SMILES through functional-group conversion, reaction type, solvent, temperature, and catalyst to product SMILES, as detailed in §A.3. By combining multiple chemical tasks into a coherent sequence, these chains evaluate both task-level chemical understanding and the ability to integrate information across dialogue turns.

Interactive Logic Integration. Chemical knowledge alone is insufficient for multi-turn dialogue: models are expected to use prior context well and respond appropriately to different communicative intents. To evaluate these abilities, we integrate six types of interaction logic through two complementary operations: rewriting an existing turn and inserting a new turn. The rewriting operation targets ***Anaphora resolution*** and ***Context memory***, for example, ***Anaphora resolution*** interprets expressions such as "that solvent" or "the previous transformation". The insertion operation covers ***Elaboration***, ***Error correction***, ***Reliability check***, and, ***Skeptical reasoning***, for example, ***Error correction*** introduces an incorrect premise to be identified and corrected (Details can be seen in §A.3). For each dialogue, an LLM selects interaction logic compatible with the dialogue content and performs one rewrite and one insertion. This context-aware process maintains dialogue coherence while enabling systematic evaluation of contextual understanding and interactive capabilities.

Human Review. Because automatic checks cannot fully assess chemical validity and dialogue quality, we combine automatic validation with human review. Automatic validation first examines each turn for evidence support, interaction-logic cues, and topic compatibility, discarding dialogues that fail any check. Human reviewers then inspect the remaining dialogues through a browser-based review platform, where they correct answers, annotate issues, and remove unusable samples. All review actions are recorded in a change log for traceability. In total, five chemistry reviewers spent 200 hours reviewing the dataset: they revised 604 turns across 291 dialogues and removed 20 of the 416 candidate dialogues (Appendix A.7).

3.2 Dataset Statistics

ChemDialog contains 396 dialogues from 205 source reactions. The dialogues have 2,725 turns in total. Of these, 2,329 non-seed turns are scored. The scored turns include 791 capability turns. The mean dialogue length is 6.88 turns, each turn averages 17.0 words (8.8 in questions and 8.2 in answers), and each dialogue averages 117.0 words. More details can be seen in Appendix A.5.

3.3 Evaluation Scheme

Evaluating LLMs in multi-turn chemical dialogues presents three main challenges: ❶ ***Heterogeneous answer formats***: chemical tasks produce outputs ranging from molecular structures and numerical conditions to open-ended explanations, making a single metric unsuitable for all tasks; ❷ ***Consistent comparison across diverse LLMs***: variations in response style, model family, scale, and generation configuration complicate direct comparison; and ❸ ***Chemical correctness versus interaction logic***: a response may be chemically plausible while failing to follow the dialogue context, or interactionally appropriate while containing incorrect chemistry.

To account for ❶, we match the metric to the answer type. Molecular structures are evaluated by exact matching of RDKit canonical and isomeric SMILES (Landrum, 2013); numerical values, *e.g.*, reaction times and yields, are assessed using numeric or range tolerances; and open-ended responses are evaluated using deterministic rules followed by a constrained LLM judge. The detailed rules are

provided in Appendix A.6. The reliability of the LLM judge is validated by its agreement with human evaluations, as detailed in Appendix A.15. To account for ❷, we prescribe a task-specific response format and extract only the designated answer fields for scoring. We also evaluate all models using the same dialogue turns, history, and decoding settings, reducing variation caused by response style or experimental configuration. To account for ❸, we compute two scores for each turn containing interaction logic: a chemical-content score $c_i \in [0, 1]$ and an interaction-logic score $\ell_i \in \{0, 0.3, 0.5, 0.8, 1.0\}$. We combine them by taking the lower score:

$$s_i = \min(c_i, \ell_i), \tag{1}$$

This design gives a high score only when the response is both chemically correct and interactionally appropriate. For turns without interaction logic, the final score is simply the chemical-content score, $s_i = c_i$. Appendix A.6 provides the complete scoring rules and interaction-type-specific rubrics.

4 Experimental Results and Discussion

4.1 Benchmark Performance Comparison

Table 1: Pass@1 performance (%) of 13 general-purpose LLMs on ChemDialog. FS, RS, and RP denote three chemical topics, *i.e.*, forward synthesis, retrosynthesis, and reaction procedure, respectively. AR, CM, EL, EC, RC, and SK denote six types of interactive logic, *i.e.*, anaphora resolution, context memory, elaboration, error correction, reliability check, and skeptical reasoning, respectively. **Red** and **blue** denote the highest and second-highest distinct scores in each metric column.

Model	Chemical Topics (%)				Interactive Capabilities(%)							Overall Avg (%)
	FS	RS	RP	Chem Avg	AR	CM	EL	EC	RC	SK	Logic Avg	
DeepSeek-v3.2	38.8	43.7	48.9	43.5	40.0	37.8	62.0	52.8	77.4	47.1	49.5	46.5
DeepSeek-v4-flash	42.2	57.6	51.7	50.3	**50.8**	**46.8**	63.8	**61.0**	73.7	**62.9**	**57.2**	53.7
DeepSeek-v4-pro	50.8	58.6	51.8	53.7	47.0	45.1	57.6	53.2	76.3	58.4	53.9	53.8
GLM-5.2	50.7	52.8	**53.3**	52.2	44.8	43.0	63.5	56.1	**81.0**	57.3	54.4	53.3
Kimi-K2.6	50.1	**60.7**	51.3	54.1	46.0	39.6	**72.6**	**58.6**	79.4	**58.8**	55.2	**54.6**
GPT-5.5	52.4	58.8	50.9	54.1	**51.1**	**47.8**	56.0	51.1	70.9	54.3	53.9	54.0
GPT-5.6-luna	46.4	52.7	51.0	49.9	49.0	45.0	53.5	45.2	75.7	55.9	52.6	51.3
GPT-5.6-terra	46.0	54.2	49.2	49.8	47.9	43.0	53.7	47.3	70.1	50.6	50.6	50.2
GPT-5.6-sol	**53.4**	**59.8**	**52.7**	**55.4**	50.6	46.5	53.8	47.3	72.0	51.4	52.5	53.9
Qwen3.7-flash	37.3	41.2	48.0	41.9	35.5	33.5	50.2	53.4	73.7	44.5	45.1	43.5
Qwen3.7-max	50.9	56.3	51.9	53.0	46.0	43.8	57.2	**58.6**	**79.8**	54.3	53.9	53.4
Qwen3.8-flash	41.3	46.8	49.7	45.7	41.8	40.8	56.3	53.3	71.3	54.4	50.2	48.0
Qwen3.8-max	**52.8**	59.3	**52.7**	**55.0**	48.0	46.2	**69.0**	56.8	79.6	51.5	**55.8**	**55.4**

We compare 13 general-purpose models from five families: DeepSeek, GLM, GPT, Kimi, and Qwen, in Table 1. All models are evaluated under the same protocol. At each scored turn, the model receives the current question, the unscored opening turn that provides the initial reaction information, and the complete gold-standard dialogue history preceding the question. Reference answers are used in place of the model's earlier outputs, so every model is conditioned on the same correct context and errors cannot propagate across turns. Explicit thinking is disabled, and each response has a maximum generation budget of 512 tokens. We sample one candidate per turn ($K = 1$) at temperature $T = 0.7$. The resulting answer is assessed with the proposed evaluation scheme and DeepSeek-v4-flash serves as the LLM judge. We conclude three main findings from Table 1.

Mainstream LLMs still have considerable room for improvement in both chemical-reaction understanding and chemistry-related interaction. As shown in Table 1, even the best-performing model achieves an Overall Avg of only 55.4%. The highest Chemical Avg is 55.4%, and the highest Logic Avg is 57.2%, indicating that substantial limitations remain in both dimensions.

Strong performance on chemical tasks does not necessarily translate into strong interactive performance. GPT-5.6-sol achieves the highest Chemical Avg (55.4%), yet its Logic Avg is only 52.5%, below the best Logic Avg of 57.2% achieved by DeepSeek-v4-flash. This discrepancy shows that a model's ability to answer chemical questions accurately does not guarantee that it can effectively interpret and respond to interaction-dependent questions in multi-turn chemical dialogues.

LLMs struggle more with implicit dependencies on dialogue history than with explicitly stated interaction intents. Context memory (33.5%–47.8%) and anaphora resolution (35.5%–51.1%) score consistently lower than reliability check (70.1%–81.0%) and elaboration (50.2%–72.6%). The former require models to recover a prior chemical fact or resolve a reference to an earlier turn, whereas the latter explicitly state the claim to verify or the topic to elaborate. This gap identifies implicit cross-turn information integration as a key bottleneck in chemical dialogue.

Table 2: Cross-task information integration under three dialogue-history conditions. ***FS product exact hit*** and ***RS reactant-set exact hit*** evaluate the final questions in forward synthesis and retrosynthesis dialogue, respectively. ***No history*** retains only the initial task-setting exchange and the final target question. ***Question only*** additionally includes the intermediate user questions but omits their answers. ***Gold history*** includes the complete reference dialogue and its correct intermediate answers, measuring the contribution of accurate cross-task information. Δ_{Q} =***Question only***−***No history*** and Δ_{Gold} =***Gold history***−***No history***.

Model	FS product exact hit (%)					RS reactant-set exact hit (%)				
	No history	Question only	Gold history	Δ_{Q}	Δ_{Gold}	No history	Question only	Gold history	Δ_{Q}	Δ_{Gold}
DeepSeek-v4-flash	12.2	**16.5**	18.0	+4.3	+5.8	7.3	8.8	34.0	+1.5	+26.7
DeepSeek-v4-pro	14.4	**16.5**	31.7	+2.1	+17.3	**11.7**	**13.1**	41.8	+1.4	+30.2
GLM-5.2	6.5	7.9	28.8	+1.4	+22.3	5.1	5.1	34.6	0.0	+29.5
GPT-5.6-sol	**15.8**	15.8	32.4	0.0	+16.5	9.5	8.8	**45.1**	-0.7	+35.6
Kimi-K2.6	**28.1**	**30.2**	**42.4**	+2.1	+14.4	**21.2**	**12.4**	**48.4**	-8.8	+27.2
Qwen3.8-max	8.6	14.4	**34.5**	+5.8	+25.9	8.0	**12.4**	41.8	+4.4	+33.8
Models Avg	14.3	16.9	31.3	+2.6	+17.0	10.5	10.1	41.0	-0.4	+30.5

4.2 Cross-Task Information Integration Analysis

Most chemistry benchmarks evaluate tasks independently and therefore cannot show whether a model can use information from one task to solve another. To test this cross-task integration, we vary only the dialogue history available when the model answers the final forward-synthesis or retrosynthesis question, while holding all other generation settings fixed. Table 2 compares three conditions, *i.e.*, ***No history***, ***Question only***, and ***Gold history***. Table 2 supports two main findings.

Correct dialogue history consistently improves final predictions, showing that LLMs can integrate information across chemical tasks. As shown in Table 2, ***Gold history*** increases the average exact-hit rate by 17.0% for forward synthesis and 30.5% for retrosynthesis relative to ***No history***. These consistent gains show that chemical facts established in earlier dialogue turns can help models recover the final product or reactants. We also find ***Gold history*** helps retrosynthesis better compared with forward synthesis. This may be because, in forward synthesis, the given reactants and conditions usually constrain the possible product, limiting the additional benefit of history. In contrast, for retrosynthesis, the same target can have multiple plausible precursor sets, so correct history provides greater benefit by helping the model identify the intended solution.

The question sequence can provide limited guidance, but most of the history benefit comes from the correct intermediate answers. Compared with ***No history***, ***Question only*** improves forward synthesis by only 2.6% on average and changes retrosynthesis by -0.4%. The intermediate questions indicate which aspects of the reaction should be considered, but without their answers they do not supply the chemical information needed for the final prediction. This explains why their overall benefit is much smaller than that of ***Gold history***. However, contradicted with ***Gold history***, ***Question only*** only helps forward synthesis. This may be because, for forward synthesis, the question sequence provides a useful reasoning path through the transformation, reaction type, and conditions before product prediction. For retrosynthesis, however, the questions do not reveal which of the multiple plausible disconnections and precursor sets is intended, so they provide no consistent improvement.

4.3 Inference-Time Sampling: Answer Quality and Consistency

LLM generation is stochastic, so a single response may not reliably represent a model's performance. We therefore investigate whether repeated sampling genuinely improves answer quality or merely increases the chance of obtaining one high-scoring candidate. For DeepSeek-v4-pro, GPT-5.6-sol,

and Qwen3.8-max, we sample $K \in \{1, 3, 5\}$ responses for each question across forward synthesis, retrosynthesis, and reaction procedure while keeping the evaluation setting fixed. We report Best@K to measure the highest score among the K candidates, Mean@K to measure their average quality, and Self-consistency@K (SC@K) to measure answer agreement after text and SMILES normalisation. The expermental results are reported in Table 3.

Increasing the sample count improves best-of-K performance but has limited impact on average answer quality. As shown in Table 3, Best@K improves as more samples are generated: Best@5 exceeds Best@1 by 6.3% to 13.8% across the nine model–task combinations, reflecting the increased chance of obtaining a high-scoring response. In contrast, Mean@5 differs from Mean@1 by only -2.4% to $+0.9\%$, indicating that the quality of a typical response remains largely stable as the sample count increases. Meanwhile, self-consistency decreases by 2.9% to 5.6% from $K = 3$ to $K = 5$, suggesting that additional samples expose greater variation among model responses. Together, these results indicate that single-sample evaluation provides a reasonable fixed-budget estimate of typical answer quality, while best-of-K evaluation better reflects the potential performance obtainable from multiple attempts (see Appendix A.9 for details).

Table 3: Inference-time sampling performance and answer consistency. Best@K and Mean@K denote the highest and average scores among K responses, respectively. SC@K denotes the modal-answer fraction after text and SMILES normalisation. Since SC@K is undefined for $K = 1$, we report it only for $K = 3$ and $K = 5$.

Model	Task	Best@K			Mean@K			SC@K		
		$K=1$	$K=3$	$K=5$	$K=1$	$K=3$	$K=5$	$K=1$	$K=3$	$K=5$
DeepSeek-v4-pro	FS	50.8	**60.6**	**64.6**	50.8	51.1	51.2	–	73.8	69.9
GPT-5.6-sol	FS	**53.4**	57.5	60.8	**53.4**	**52.9**	**52.8**	–	**81.3**	**78.4**
Qwen3.8-max	FS	**52.8**	**58.1**	**61.4**	**52.8**	**52.8**	**52.7**	–	**81.6**	**78.1**
DeepSeek-v4-pro	RS	58.6	**67.0**	**71.6**	58.6	56.7	**57.5**	–	70.4	65.7
GPT-5.6-sol	RS	**59.8**	**66.6**	**68.7**	**59.8**	**60.0**	**59.8**	–	**79.6**	**76.4**
Qwen3.8-max	RS	**59.3**	64.3	66.7	**59.3**	**57.4**	56.9	–	**77.5**	**74.1**
DeepSeek-v4-pro	RP	**51.8**	**59.7**	**62.1**	**51.8**	**52.8**	**52.7**	–	67.2	61.6
GPT-5.6-sol	RP	**52.7**	**57.1**	**59.3**	**52.7**	**52.5**	**52.7**	–	**78.0**	**74.9**
Qwen3.8-max	RP	**52.7**	56.7	59.0	**52.7**	52.0	52.5	–	**69.9**	**65.8**

4.4 Model Scale and Architecture Analysis

Model size is often used as a proxy for capability, but performance may also depend on architecture and parameter efficiency. To examine these factors within one model family, we evaluate five Qwen3 checkpoints under the same Pass@1 protocol, with the results shown in Table 4. As seen in the table, larger models generally perform better, while MoE models achieve strong performance with fewer activated parameters. To be specific, among dense models, Overall increases from 36.0% at 8B to 38.7% at 14B and 41.5% at 32B. Qwen3-30B-A3B reaches 38.6% with only 3B activated parameters, nearly matching the 14B dense model, while Qwen3-235B-A22B achieves the highest Overall score of 43.8% with 22B activated parameters, outperforming the 32B dense model. These results suggest that both model scale and architecture contribute to performance.

Table 4: Pass@1 performance of Qwen3 models across various parameter scales and architectures.

Model	Chemical Topics (%)				Interactive Capabilities (%)							Overall Avg (%)
	FS	RS	RP	Chemical Avg	AR	CM	EL	EC	RC	SK	Logic Avg	
Qwen3-8B (dense)	27.0	25.2	45.0	31.9	26.4	25.8	50.2	42.3	**83.1**	38.2	40.1	36.0
Qwen3-14B (dense)	28.8	30.4	45.6	34.5	32.7	28.3	51.4	51.8	70.6	**46.4**	42.9	38.7
Qwen3-30B-A3B (MoE)	26.1	32.3	45.7	34.2	**33.3**	29.6	47.3	**57.7**	72.8	39.8	43.0	38.6
Qwen3-32B (dense)	**31.7**	**35.0**	**46.8**	**37.4**	33.0	**31.1**	**53.9**	**60.0**	**73.3**	**47.7**	**45.5**	**41.5**
Qwen3-235B-A22B (MoE)	**36.2**	**39.3**	**48.6**	**41.0**	**37.6**	**35.5**	**59.4**	50.8	70.8	45.5	**46.7**	**43.8**

4.5 LLM-Judge Robustness and Human Validation

Because some open-ended answers are scored by an LLM judge, we test whether the results depend on the evaluator. We perform two checks: rescoring the same frozen answers with a second LLM judge and comparing the automatic scores with ratings from chemistry students. For the first check, DeepSeek-v4-flash and GPT-5.6-sol score the same Pass@1 answers without regeneration.

Experiments in Appendix A.13 show the aggregate results are stable across the two judges(Table 5).

Table 5: Frozen-answer judge rescoring. Both judges score the same sample-one answers with no regeneration, so every scored turn is matched. Delta is GPT judge minus primary judge and is in percentage points (pp); all columns except Pearson and Spearman are percentages; difference, correlation, and agreement columns are not ranked by colour.

Generator	Task	N	Primary	GPT (frozen)	Δ	Pearson	Spearman	Exact
DeepSeek-v4-flash	FS	829	42.2	42.0	-0.1	0.967	0.967	93.8
DeepSeek-v4-flash	RP	709	51.7	52.3	$+0.6$	0.961	0.955	75.5
DeepSeek-v4-flash	RS	791	57.6	54.5	-3.1	0.917	0.917	79.5
DeepSeek-v4-flash	All	2,329	50.3	49.4	-0.9	0.948	0.947	83.4
GPT-5.6-sol	FS	829	53.4	53.3	0.0	0.987	0.988	96.3
GPT-5.6-sol	RP	709	52.7	54.4	$+1.7$	0.972	0.971	80.3
GPT-5.6-sol	RS	791	59.8	59.3	-0.5	0.941	0.941	85.0
GPT-5.6-sol	All	2,329	55.4	55.7	$+0.3$	0.967	0.968	87.5

For human validation, five chemistry students rated 88 anonymised answers from 20 questions using the same five quality bands and ranked the answers within each question. As shown in Table 6, 65–70 of the 88 answers receive the same band from each student and the automatic judge. Besides, the correlation is also high, speicially Kendall τ ranging from 0.759 to 0.900 and Spearman ρ from 0.832 to 0.946. These results show that the LLM judge is broadly consistent with human evaluation. Further details are provided in Appendix A.15.

Table 6: Agreement between each human rater and the automatic judge. "Order = judge" counts questions with identical answer rankings, "Band map = judge" counts questions with identical band assignments for all answers, and "Same band" counts individual answers assigned to the same band. Mean absolute differences measure disagreement in score, band, and rank; score and band differences are in percentage points (pp), while rank differences are in positions. Kendall τ and Spearman ρ are correlation coefficients.

Rater	Order = judge	Band map = judge	Same band (88)	Mean $\lvert\Delta\rvert$ score (pp)	Mean $\lvert\Delta\rvert$ band (pp)	Mean $\lvert\Delta\rvert$ rank	Mean τ	Mean ρ
1	9/20	7/20	65/88	8.4	8.0	0.39	$+0.759$	$+0.832$
2	10/20	8/20	68/88	7.0	6.0	0.27	$+0.843$	$+0.911$
3	12/20	7/20	70/88	5.5	5.1	0.23	$+0.900$	$+0.946$
4	9/20	5/20	70/88	5.9	5.4	0.25	$+0.866$	$+0.926$
5	12/20	7/20	68/88	5.9	5.7	0.23	$+0.883$	$+0.936$

5 Conclusion

This work investigates the evaluation of understanding chemical reactions through multi-turn interaction in LLMs. Unlike existing benchmarks that evaluate chemical tasks independently through single-turn questions, this work connects related tasks within a dialogue and examines whether models can use information from earlier turns while responding to different interaction intents. To support this evaluation, we introduce ChemDialog, a human-verified benchmark containing 396 dialogues and 2,725 turns grounded in 205 source reactions. It covers forward synthesis, reaction procedure, and retrosynthesis, together with six types of interaction logic, and uses a format-aware evaluation scheme for both chemical correctness and context-appropriate responses. Our evaluation of 13 mainstream LLMs shows that current models still have considerable room for improvement and that strong chemical-task performance does not necessarily translate into strong interactive performance. Further analysis shows that correct information from earlier dialogue turns substantially improves subsequent predictions, although the gains differ across models and between forward synthesis and retrosynthesis. These findings highlight the need to evaluate cross-task information integration alongside task-level chemical knowledge. We hope ChemDialog provides a useful resource for studying and improving LLMs in multi-turn chemical interactions.

Limitations. ChemDialog has two main limitations. First, strict manual review ensures data quality but limits the current dataset size. Second, to make reliable ground-truth annotation and evaluation feasible, we construct dialogues using predefined task structures, which may not fully capture real-world chemical interactions.

AI USE STATEMENT

Generative AI tools, including OpenAI Codex with GPT-5.6, assisted dialogue generation, code development, and language polishing. The constructed dialogue collection was checked through the structured validation and review process described in Appendix A.7, and reviewer backgrounds and inter-rater agreement are reported in Appendix A.15. The authors remain responsible for the manuscript, claims, and released artifacts.

ETHICS STATEMENT

ChemDialog is intended for research evaluation rather than laboratory instruction. Model responses can appear plausible while containing unsafe or incorrect operational details. A public release should document provenance and licensing, remove identifying or restricted material, and state that benchmark outputs require expert verification before practical use.

REPRODUCIBILITY STATEMENT

Section 3.1 describes dialogue construction and Section 3.3 defines evaluation. Appendix A.3–A.6 provides additional detail. The anonymized details for this work is available at `https://anonymous.4open.science/r/iclr2027-anon-535C/`.

REFERENCES

Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, et al. Probing the limitations of multimodal language models for chemistry and materials research. *Nature Computational Science*, 5(10):952–961, 2025.

M. Andriushchenko, A. Souly, M. Dziemian, et al. Agentharm: A benchmark for measuring harmfulness of llm agents. In *International Conference on Learning Representations (ICLR)*, 2025.

Ge Bai, Jie Liu, Xingyuan Bu, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*, 2024a.

Yushi Bai, Xin Lv, Jiajie Zhang, et al. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024b.

Feiyang Cai, Katelin Zacour, Tianyu Zhu, et al. Chemfm as a scaling law guided foundation model pre-trained on informative chemicals. *Communications Chemistry*, 9(1):3, dec 2025.

M. Chen, J. Tworek, H. Jun, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

K. Cobbe, V. Kosaraju, M. Bavarian, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, et al. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, 2025.

Y. Du, S. Li, A. Torralba, et al. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning (ICML)*, 2023.

Kehan Guo, Bozhao Nan, Yujun Zhou, et al. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Meng-Hao Guo, Jiajun Xu, Yi Zhang, et al. R-bench: Graduate-level multi-disciplinary benchmarks for llm and mllm complex reasoning evaluation. *arXiv preprint arXiv:2505.02018*, 2025.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 59662–59688, 2023.

Yan Guo, Menglan Luo, Wenbo Zhang, et al. Few-shot molecular property optimization via a domain-specialized large language model. *Chemical Science*, 17(10):4928–4941, 2026.

Y. He, D. Jin, C. Wang, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instruction following. *arXiv preprint arXiv:2410.15553*, 2024.

Omer Hofman, Jonathan Brokman, Oren Rachmil, et al. Maps: A multilingual benchmark for agent performance and security. In *Findings of the Association for Computational Linguistics: EACL 2026*, Rabat, Morocco, 2026. Association for Computational Linguistics.

Y. Huang, Y. Bai, Z. Zhu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Y. Huang, L. Sun, H. Wang, et al. Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning (ICML)*, 2024.

Yuqing Huang, Rongyang Zhang, Xuesong He, et al. Chemeval: A multi-level and fine-grained chemical capability evaluation for large language models. In *International Conference on Learning Representations (ICLR)*, 2026.

C. E. Jimenez, J. Yang, A. Wettig, et al. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations (ICLR)*, 2024.

Q. Jin, B. Dhingra, Z. Liu, et al. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, 2019.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, et al. BIG-bench extra hard. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pp. 26473–26501, Vienna, Austria, July 2025. Association for Computational Linguistics.

Steven M. Kearnes, Michael R. Maser, Michael Wleklinski, et al. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021.

M. G. Kendall and B. Babington Smith. The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3):275–287, 1939.

Hanbum Ko, Chanhui Lee, Ye Rin Kim, et al. Reactionreasoner: Towards reasoning llm for chemical reaction prediction. In *NeurIPS 2025 Workshop on AI for Science*, 2025.

Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433, 2004.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, et al. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pp. 20153–20177, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Greg Landrum. Rdkit: Open-source cheminformatics, 2013. Software package, https://www.rdkit.org.

Hao Li, He Cao, Bin Feng, et al. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. *Bioinformatics*, 41, 2025a.

Junxian Li, Di Zhang, Xunzhi Wang, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2025)*, volume 39, pp. 415–423, 2025b.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, et al. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, 2023.

Zhenwen Liang, Dian Yu, Wenhao Yu, et al. Mathchat: Benchmarking mathematical reasoning and instruction following in multi-turn interactions. *arXiv preprint arXiv:2405.19444*, 2024.

Ruoyu Liu, Kui Xue, Xiaofan Zhang, et al. Interactive evaluation for medical LLMs via task-oriented dialogue system. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pp. 4871–4896, Abu Dhabi, UAE, January 2025a. Association for Computational Linguistics.

Xuan Liu, Siru Ouyang, Xianrui Zhong, et al. Fgbench: A dataset and benchmark for molecular property reasoning at functional group-level in large language models. *arXiv preprint arXiv:2508.01055*, 2025b.

G. Mialon, C. Fourrier, T. Wolf, et al. Gaia: A benchmark for general ai assistants. In *International Conference on Learning Representations (ICLR)*, 2024.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17(7):1027–1034, 2025.

D. M. Nguyen, T. C. Phan, N. Le Hai, et al. Codemmlu: A multi-task benchmark for assessing code understanding and reasoning capabilities of codellms. In *International Conference on Learning Representations (ICLR)*, 2025.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`, 2023.

E. M. Ponti, G. Glavaš, O. Majewska, et al. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pp. 2366–2376, 2020.

A. Romanou, N. Foroutan, A. Sotnikova, et al. Include: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Machine Learning (ICML)*, 2024.

Nicholas T. Runcie, Charlotte M. Deane, and Fergus Imrie. Assessing the chemical intelligence of large language models. *Journal of Chemical Information and Modeling*, 2025.

M. Sap, H. Rashkin, D. Chen, et al. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, pp. 4463–4473, 2019.

John Schulman, Barret Zoph, Christina Kim, et al. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*, 2022.

P. Shojaee, N.-H. Nguyen, K. Meidani, et al. Llm-srbench: A new benchmark for scientific equation discovery with large language models. *arXiv preprint arXiv:2504.10415*, 2025.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

Kunyang Sun, Dorian Bagni, Joseph M. Cavanagh, et al. Synllama: Generating synthesizable molecules and their analogs with large language models. *ACS Central Science*, 11(11): 2108–2120, 2025.

L. Sun, Y. Han, Z. Zhao, et al. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2024)*, 2024.

Minyang Tian, Luyu Gao, Shizhuo Zhang, et al. Scicode: A research coding benchmark curated by scientists. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 30624–30650, 2024.

Alex Wang, A. Singh, J. Michael, et al. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, 2018.

Alex Wang, Y. Pruksachatkun, N. Nangia, et al. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

B. Wang, C. Xu, S. Wang, et al. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *International Conference on Learning Representations (ICLR)*, 2024a.

Yuxia Wang, Haonan Li, Xudong Han, et al. Do-not-answer: Evaluating safeguards in LLMs. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 896–911, St. Julian's, Malta, March 2024b. Association for Computational Linguistics.

Zeming Wei, Tianlin Li, Xiaojun Jia, et al. Position: Agent-specific trustworthiness risk as a research priority. In *International Conference on Machine Learning (ICML)*, 2025.

Jiaqing Xie, Weida Wang, Ben Gao, et al. Qcbench: Evaluating large language models on domain-specific quantitative chemistry. *Journal of Chemical Information and Modeling*, 65(22): 12268–12278, 2025.

Yifei Yang, Runhan Shi, Zuchao Li, et al. Batgpt-chem: A foundation large model for chemical engineering. *Research*, 8:0827, sep 2025.

Botao Yu, Frazier N. Baker, Ziqi Chen, et al. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *Conference on Language Modeling (COLM 2024)*, 2024.

Zenan Zhai, Christian Druckenbrodt, Camilo Thorne, et al. Chemtables: A dataset for semantic classification on tables in chemical patents. *Journal of Cheminformatics*, 13(1):97, 2021.

Di Zhang, Wei Liu, Qian Tan, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024a.

Tianyi Zhang, Varsha Kishore, Felix Wu, et al. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*, 2020.

Yueqing Zhang, Wentao Liu, Yan Zhang, et al. Ecnu-chemgpt: A large language model for chemistry and retrosynthesis prediction. *CCS Chemistry*, 2026. Just Published.

Z. Zhang, L. Lei, L. Wu, et al. Safetybench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024b.

Zehua Zhao, Zhen Huang, Jiajun Li, et al. Superchem: A multimodal reasoning benchmark in chemistry. *arXiv preprint arXiv:2512.01274*, 2025a.

Zihan Zhao, Da Ma, Lu Chen, et al. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4):102523, 2025b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 46595–46623, 2023.

Kaijie Zhu, Jindong Wang, Qinlin Zhao, et al. Dyval 2: Dynamic evaluation of large language models by meta probing agents. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62599–62617. PMLR, 21–27 Jul 2024.

A Additional Evidence and Evaluation Details

The appendix follows the order of the main manuscript: related-work scope, dataset construction and statistics, evaluation metrics and worked scoring cases, dataset review and validation, experimental supplements, and finally the independent human LLM-as-a-judge protocol.

A.1 Review of LLM-oriented Benchmarks

The development of instruction-tuned dialogue systems has motivated evaluation suites that extend beyond a single task or dataset (Schulman et al., 2022). Unified platforms such as OpenCompass organise heterogeneous benchmarks under a common evaluation interface (OpenCompass Contributors, 2023). Earlier language-understanding and text-generation evaluations include GLUE, SuperGLUE, and BERTScore (Wang et al., 2018; 2019; Zhang et al., 2020); later suites such as C-Eval broadened coverage across disciplines and difficulty levels (Huang et al., 2023). Multilingual understanding is examined by INCLUDE and XCOPA (Romanou et al., 2024; Ponti et al., 2020), while complex reasoning is represented by R-Bench, verifier-based mathematical problems, and BIG-Bench Extra Hard (Guo et al., 2025; Cobbe et al., 2021; Kazemi et al., 2025). HaluEval and multi-agent debate address hallucination or factuality-related behaviour (Li et al., 2023; Du et al., 2023).

Evaluation has also diversified by risk and deployment setting. AgentHarm, SafetyBench, and Do-Not-Answer probe harmful or unsafe behaviour (Andriushchenko et al., 2025; Zhang et al., 2024b; Wang et al., 2024b); DyVal 2, Adversarial GLUE, and TrustLLM examine dynamic evaluation, robustness, or trustworthiness (Zhu et al., 2024; Wang et al., 2021; Huang et al., 2024; Wei et al., 2025). MAPS and GAIA target agent capabilities, while LongBench focuses on long-context understanding (Hofman et al., 2026; Mialon et al., 2024; Bai et al., 2024b). Domain-oriented benchmarks include SciEval, PubMedQA, and LLM-SRBench for scientific tasks (Sun et al., 2024; Jin et al., 2019; Shojaee et al., 2025); Social IQa for social interaction reasoning (Sap et al., 2019); and HumanEval, SWE-bench, and CodeMMLU for programming and software engineering (Chen et al., 2021; Jimenez et al., 2024; Nguyen et al., 2025). ChemDialog adds a reaction-dialogue benchmark to this set.

Open-ended benchmark scoring creates a separate evaluation problem. MT-Bench formalised LLM-as-a-judge comparison, and PandaLM studied automatic evaluation for instruction-tuned models (Zheng et al., 2023; Wang et al., 2024a). Such scores depend on the rubric, judge, presentation, and calibration. ChemDialog combines exact or structured chemical checks with constrained semantic scoring, tests an alternate judge, and defines an independent human-consistency study.

A.2 Chemical-domain LLMs

The main comparison uses general-purpose models. Chemistry-oriented systems are a separate comparison class, and most of them are single-task models. ChemFM is a foundation model pre-trained on chemical data and is evaluated mainly on retrosynthesis prediction (Cai et al., 2025). ECNU-ChemGPT combines general chemistry knowledge with retrosynthesis planning (Zhang et al., 2026). ReactionReasoner targets reaction prediction with reasoning-style training data and is currently a NeurIPS AI for Science workshop entry (Ko et al., 2025). DrugLLM optimises molecular properties in the few-shot setting (Guo et al., 2026), and SynLlama generates synthesisable molecules from purchasable building blocks (Sun et al., 2025). ChemDFM and BatGPT-Chem are broader foundation models for chemistry and chemical engineering (Zhao et al., 2025b; Yang et al., 2025); LlaSMol is instruction-tuned on a large collection of chemical tasks (Yu et al., 2024); and ChemVLM adds chemical images to text (Li et al., 2025b).

General chemical dialogue models remain rare. ChemDFM is the representative one: it is pre-trained on chemical literature and textbooks and then instruction-tuned, and it answers free-form chemistry questions in conversation (Zhao et al., 2025b). ChemLLM offers a conversational interface that mainly supports task execution such as name conversion and molecular description (Zhang et al., 2024a). Both are trained on chemical texts and task instructions rather than on multi-turn dialogue that follows a reaction record.

A.3 DATA CONSTRUCTION PROTOCOL

The source-to-dialogue workflow has two generation stages followed by structured validation and review. Stage 1 maps a reaction record to a task-specific evidence file and filters unsupported topics; the reaction records are stored in the Open Reaction Database schema (Kearnes et al., 2021). Question, answer, and validation roles then construct a five-to-seven-turn topic chain. Forward synthesis proceeds from reactants towards product and reaction identity; procedure understanding proceeds from reaction basis towards operational outcome; retrosynthesis proceeds from the target towards precursors.

Stage 2 assigns one context-dependent rewrite and one interaction insertion when supported. Rewrites implement ANAPHORA RESOLUTION or CONTEXT MEMORY. Insertions implement RELIABILITY CHECK, SKEPTICAL REASONING, ELABORATION, or ERROR CORRECTION. The validator checks capability cues, topic compatibility, evidence support, turn numbering, and anchor references. The final collection contains 791 capability turns because retrosynthesis dialogue 3 retains one rewritten capability turn but no inserted turn.

Each benchmark dialogue record stores stable dialogue and source-reaction identifiers, task, turn, question, answer, topic, capability, evidence, confidence, construction action, and anchor. Machine-specific source paths are rewritten to stable `source_records/<id>.txt` identifiers. Because ChemDialog is an evaluation benchmark, the appendix describes the generation logic without reproducing generator source code.

A.4 REPRESENTATIVE CONSTRUCTED DIALOGUES

Table 7: Representative constructed dialogue chains. Brackets mark the Stage-2 capability operation.

Task / source	Recorded topic chain
Forward, reaction100, D2	reactants → functional-group conversion → elaboration [new aryl–C#N linkage] → reaction type [anaphora rewrite] → solvent → temperature → product SMILES.
Procedure, reaction106, D2	reaction basis → medium → condition control → elaboration [Dean–Stark support] → time → work-up [anaphora rewrite] → product/yield.
Retrosynthesis, reaction104, D4	product SMILES → functional-group analysis [anaphora rewrite] → disconnection → reaction type → solvent → reliability check [solvent confirmation] → precursor SMILES.

The first chain records aryl iodide cyanation, formation of an aryl–C#N bond, DMF, 80 °C, and a final product structure. The second records ketalisation in toluene, Dean–Stark removal of water, 110 °C for 16 h, silica purification, and a reported 100% yield. The third records a Suzuki–Miyaura interpretation, acetonitrile/aqueous sodium carbonate, and precursor structures. Exact text and evidence remain in the benchmark dialogue records.

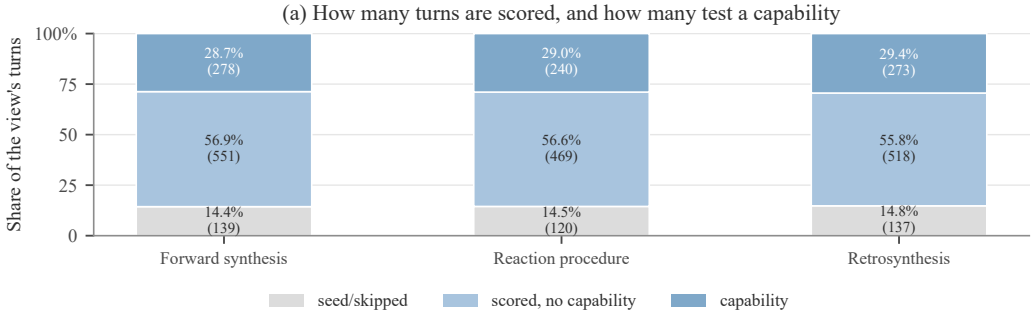
Table 8: A complete constructed dialogue: forward synthesis, dialogue 2, source reaction100. The first turn is the unscored seed, and capability turns are marked in the last column.

Turn	Question	Reference answer	Topic	Capability
T1 (seed)	Please first provide the substrate SMILES for this reaction.	`ClC=1C=C(C(=O)NC2=C(C=C(C=C2Cl)C(C(F)(F)F)(C(F)(F)F)F)Cl)C=CC1I;[C-]#N.[C-]#N.[Zn+2]`	reactant smiles	—
T2	What functional group conversion occurs in this reaction?	Aryl iodide to aryl nitrile (cyanation)	functional group conversion	—
T3	Which new linkage specifically reflects that conversion?	The key new linkage is the aryl-C#N bond formed at the carbon that originally bore iodine	functional group conversion	elaboration
T4	What reaction type does that transformation belong to?	Palladium-catalyzed aryl cyanation	reaction type	anaphora res.
T5	What solvent is used in this reaction?	N,N-dimethylformamide (DMF)	conditions solvent	—
T6	What is the reaction temperature?	80 degrees Celsius	conditions temperature	—
T7	What is the product SMILES?	`ClC=1C=C(C(=O)NC2=C(C=C(C=C2Cl)C(C(F)(F)F)(C(F)(F)F)F)Cl)C=CC1C#N`	product smiles	—

The dialogue above is the first chain listed in Table 7. It opens with the substrate SMILES, asks for the functional-group conversion, rewrites the follow-up so that T4 depends on the previous turn ('that transformation"), adds an elaboration turn on the new linkage, and closes with the solvent, the temperature, and the product SMILES. T3 is the high-scoring worked case of Table 12.

A.5 DATASET STATISTICS

Sixty-one source reactions occur in all three topics. Pairwise overlaps are 84 for forward–procedure, 102 for forward–retrosynthesis, and 66 for procedure–retrosynthesis. Capability counts are 200 anaphora-resolution, 196 context-memory, 93 elaboration, 93 error-correction, 106 reliability-check, and 103 skeptical-reasoning turns. Figure 3 shows, for each topic, the share of seed or skipped turns, scored non-capability turns, and capability turns on the left, and splits the 791 capability turns by task view and by capability on the right. Figure 4 gives the topic composition of each view in panel (a) and the length of the recorded questions and answers in panels (b) and (c).



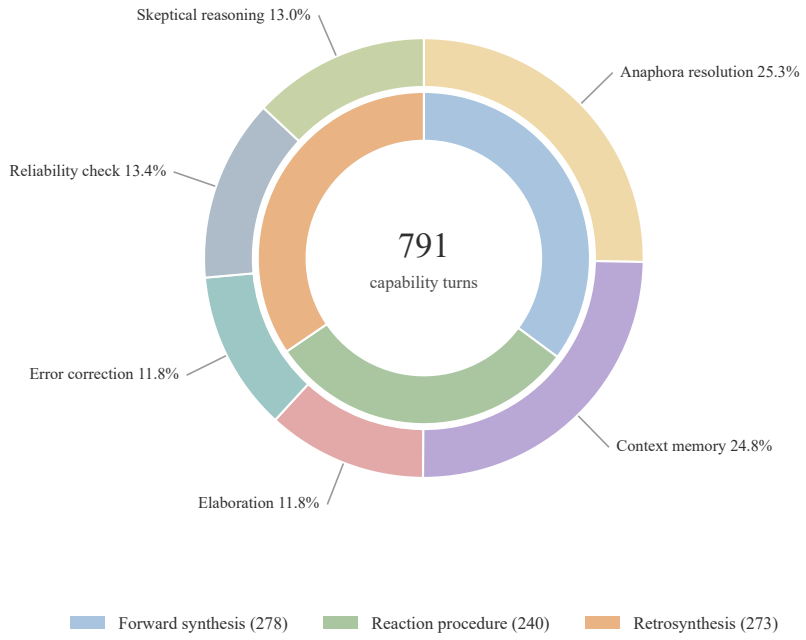


Figure 3: Turn composition of ChemDialog. (a) Turn types in each task topic. Grey segments are unscored seed turns. Light-blue segments are scored turns without a capability label. Dark-blue segments are capability turns. Each segment shows its share and turn count. (b) The 791 capability turns. The inner ring splits them by task topic. The outer ring splits them by capability. The two rings show separate marginal distributions.

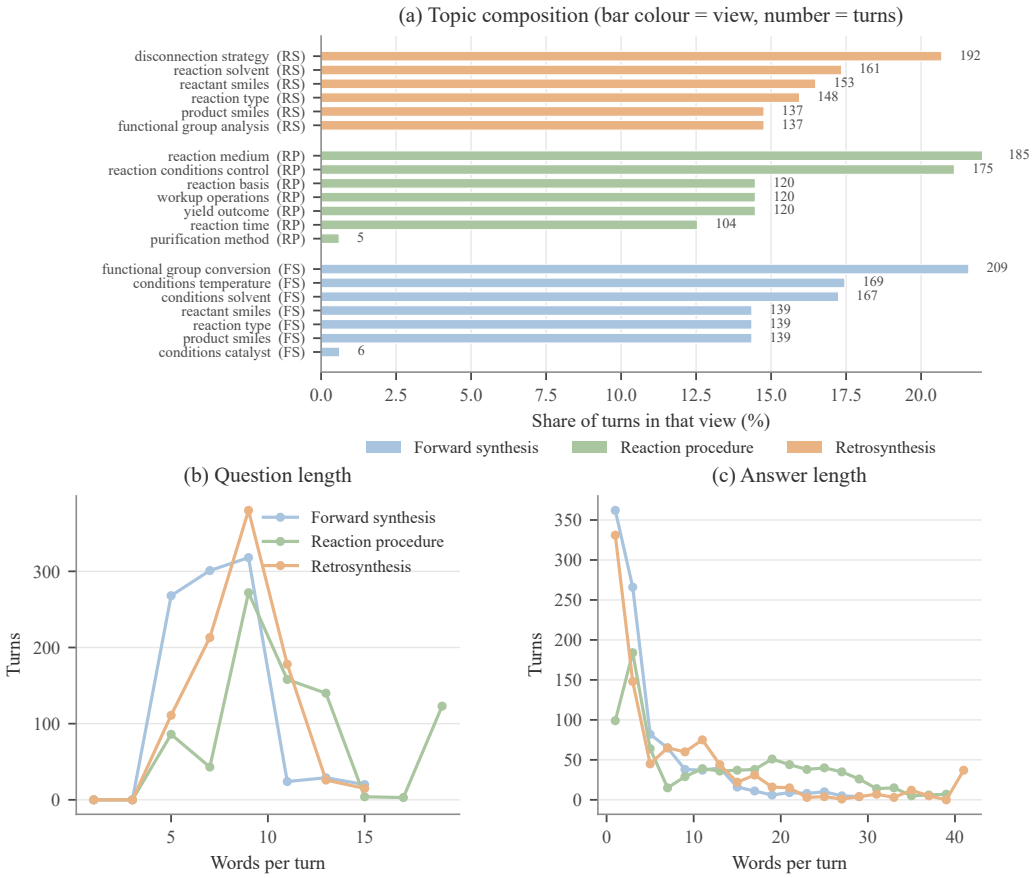


Figure 4: Topic composition and turn lengths. (a) Share of turns by topic within each view; the number of turns is printed next to every bar. (b, c) Distribution of question and answer length in words per turn, one line per view.

Table 9: Dataset and review-record statistics. Raw notes record the issues raised during review.

View	Dialogues	Turns	Mean turns	Non-seed	Capability	Unique reactions	Raw notes	Dialogues with records
Forward synthesis	139	968	6.96	829	278	139	73	76
Reaction procedure	120	829	6.91	709	240	120	7	10
Retrosynthesis	137	928	6.77	791	273	137	157	125
Total	396	2,725	6.88	2,329	791	205	237	211

A.6 DETAILED SCORING AND AGGREGATION

Evaluation protocol. The seed turn defines the task and is not scored. For every later turn the evaluator builds the requested history, obtains the model answer, scores chemical content, applies an interaction rubric when the turn tests a capability, and records error and format details. *Gold history* is the full reference dialogue history stored with the benchmark and therefore shows what a model can do when every earlier answer is correct; *model history* supplies the tested model's own earlier answers; *no history* keeps only the task seed and the final target question. The main sample-one protocol uses gold history, no thinking, $K = 1$, $T = 0.7$, and DeepSeek-v4-flash as the primary judge.

Structures and sets. RDKit canonical/isomeric SMILES retains stereochemistry before comparison (Landrum, 2013). Forward product accuracy uses exact canonical product matching. The history-control retrosynthesis question uses exact equality of the required reactant set. The broader precursor topic score is distinct: 1.0 for the complete required set, 0.5 for the correct main organic core with one missing/wrong support component, 0.2 for matching only a non-core support component, and 0 otherwise. History comparisons use exact reactant-set hits; the four-level component score is reported separately for retrosynthesis topic results.

Numbers, semantics, and procedure slots. Alias-normalised set overlap scores solvents and catalysts. Temperatures, times, and yields use parsed values, ranges, controlled labels, and topic-specific tolerances. Reaction type and functional-group change use standard labels before constrained semantic fallback. Procedure scoring checks required reactant, reagent, medium, apparatus, work-up, and purification slots and penalises stage conflicts. Product/yield outcome uses

$$c_{\text{yield}} = 0.7\,\text{EM}_{\text{product}} + 0.3\,s_{\text{yield value}}. \tag{2}$$

Capability scoring. The logic rubric uses $\{0, 0.3, 0.5, 0.8, 1.0\}$. A bare yes/no receives no logic credit. Wrong atom pairs, bond orders, reaction centres, or substitution relationships trigger a critical-fact check. Capability final score is $s_i = \min(c_i, \ell_i)$. Thus a fluent correction with the wrong chemical fact cannot score above its content score.

Table 10 lists the rule behind every topic and Table 11 the additional requirement of every interaction capability.

Table 10: Scoring rule of every topic, grouped by task topic. Seventeen rows cover the sixteen distinct topics: reaction type appears in both the forward-synthesis and the retrosynthesis chain. Temperatures and times use the numeric tolerances below, and the yield outcome follows the weighted formula in the text.

View	Topic	Rule
FS	Catalyst	Alias-normalised set overlap against the recorded catalyst set.
FS	Solvent	Alias-normalised set overlap against the recorded solvent set.
FS	Temperature	Exact value 1.0; within 5 °C 0.6; within 10 °C 0.3; otherwise 0; room temperature and ambient are treated as equivalent.
FS	Functional-group conversion	Normalised label match of the group change.
FS	Product SMILES	Exact RDKit canonical/isomeric SMILES hit.
FS	Reaction type	Normalised reaction-class label.
RP	Purification	Match of the recorded purification method.
RP	Condition control	Required-slot coverage of the condition slots, with a penalty for stage conflicts.
RP	Medium	Alias-normalised set overlap against the recorded medium.
RP	Time	Converted to hours: exact 1.0, within 20% 0.5; textual labels use windows (overnight 12–18 h 1.0, 8–12 or 18–24 h 0.5; several hours 3–8 h; a few hours 2–5 h; 'briefly" 0).
RP	Work-up	Recall of the required work-up operations.
RP	Yield outcome	$c_{\text{yield}} = 0.7\,\text{EM}_{\text{product}} + 0.3\,s_{\text{yield value}}$.
RS	Disconnection	Match of the key disconnection.
RS	Functional-group analysis	Level-set match of the analysed groups.
RS	Precursors	1.0 for the complete required set; 0.5 for the correct main core with one missing or wrong support component; 0.2 for a support component only; otherwise 0.
RS	Solvent	Alias-normalised set overlap against the recorded solvent set.
RS	Reaction type	Normalised reaction-class label.

Table 11: Interaction capabilities and their extra requirement. Every capability turn also receives a logic score $\ell \in \{0, 0.3, 0.5, 0.8, 1.0\}$, and the final score is $\min(c, \ell)$; a bare yes/no answer receives no logic credit, and wrong atom pairs, bond orders, reaction centres, or substitution relationships trigger a critical-fact check.

Capability	Turn type	Extra requirement on top of the chemical content
Anaphora resolution	Rewritten turn	The referring expression must be resolved to the entity the record supports.
Context memory	Rewritten turn	The fact carried over from the earlier turn must match the record.
Elaboration	Inserted turn	The added detail must be supported by the evidence file.
Error correction	Inserted turn	The model must reject the incorrect premise, correct it, and state the corrected conclusion; all three parts are required(three, two, or one part present scores 1.0, 0.8, or 0.5).
Reliability check	Inserted turn	The model must accept or reject the premise correctly and give a supporting fact from the record(both or one part present scores 1.0 or 0.5).
Skeptical reasoning	Inserted turn	The model must identify the weak or unsupported claim in the premise.

Failures and valid samples. The effective mean includes records with statuses `evaluated`, `evaluated_with_judge_fallback`, `empty_model_answer`, or `generation_truncated` when a final score is present. Empty output receives zero. API/provider failure or a pure-judge item without a judge result remains unavailable, not an automatic zero. Each JSON reports attempted and effective counts, API error, empty-answer, truncation, format-validity, conflict, override, and fallback records.

Sampling and weighted averages. Best@K samples K answers for an item and counts the item as solved when at least one sample is correct. For item i and candidates $k = 1, \ldots, K$,

$$\textsc{Best@K} = \frac{1}{N}\sum_{i=1}^{N} \max_{k} s_{ik}, \qquad \text{Mean@}K = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K} s_{ik}. \tag{3}$$

Answers are normalised as canonical SMILES sets when parseable and otherwise as normalised text. If the modal normalised answer occurs m_i times, $\text{SC@}K = N^{-1}\sum_i m_i/K$.

For task counts n_t and capability counts n_c, the reported aggregates are

$$\begin{aligned} \text{Chemical Avg} &= \frac{\sum_{t\in\{FS,RS,RP\}} n_t \,\text{Score}_t}{\sum_t n_t}, \\ \text{Logic Avg} &= \frac{\sum_c n_c \,\text{Score}_c}{\sum_c n_c}, \qquad \text{Overall Avg} = \frac{\text{Chemical Avg} + \text{Logic Avg}}{2}. \end{aligned} \tag{4}$$

The weights are the effective numbers of scored turns in the corresponding JSON records.

Table 12: Two high- and two low-scoring capability cases copied from sample-one detail records.

Run / location	Question	Reference	Model answer	(c, ℓ, s)
DS-v4-flash, FS D2 T3	Which new linkage reflects that conversion?	Aryl–C#N at the carbon that bore iodine.	The aryl iodide carbon bonds to nitrile carbon; the C–C#N linkage replaces C–I at the same position.	$(1, 1, 1)$
GPT-5.6-sol, RS D4 T6	Is the solvent acetonitrile and aqueous sodium carbonate?	Yes, both.	"Yes, acetonitrile and aqueous sodium carbonate."	$(1, 1, 1)$
DS-v4-flash, RS D101 T6	Is `C1CCOC1` not a reactant and just a solvent?	Yes. `C1CCOC1` is only a solvent; the reactant SMILES are `CC(=O)Nc1nc2ccc(-c3cnc(C)c(N)c3)cc2s1;O=S(=O)(Cl)c1ccccc1`.	Confirms the solvent role of THF but gives no reactant SMILES.	$(0, 1, 0)$
GPT-5.6-sol, RS D40 T4	Which bond is the key retrosynthetic cut?	The specified C–C coupling junction.	Proposes a different central biaryl C–C cut and Suzuki coupling.	$(0, 1, 0)$

Four worked evaluation cases. The low cases keep the interactive form but miss the chemical target: one proposes a different disconnection, and one confirms the yes/no question without giving the required reactant structures. The final score therefore takes the lower of content and logic. Figure 5 shows the same four turns as score cards.

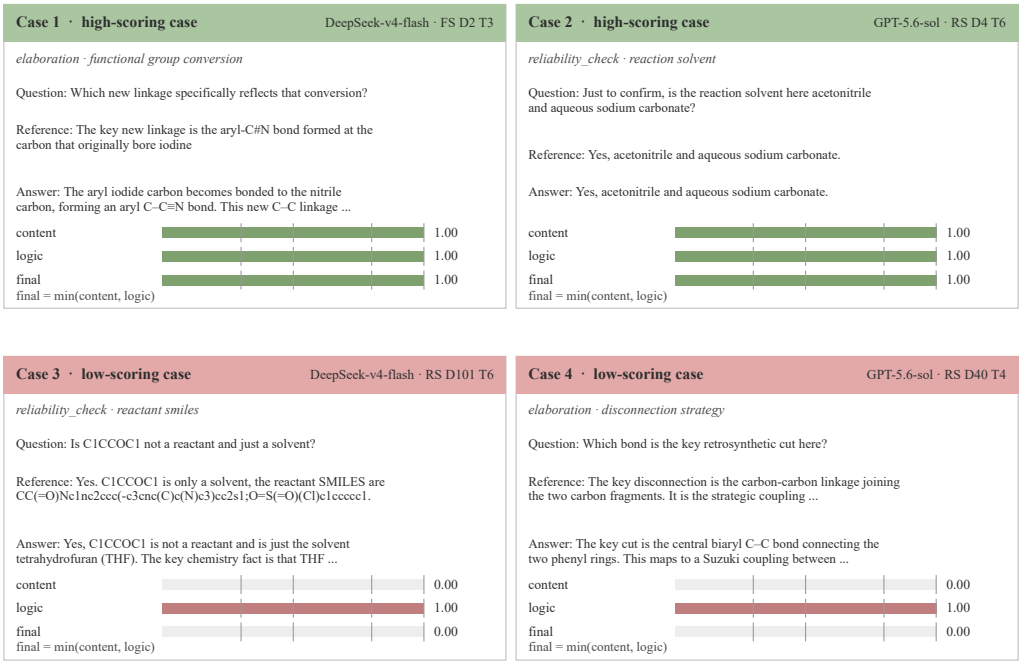


Figure 5: Score cards for the four capability turns in Table 12. Green cards are high-scoring cases, and red cards are low-scoring cases. Each card shows the question, the reference answer, and an abbreviated model answer. Bars give the content score c, the logic score ℓ, and the final score $s = \min(c, \ell)$. Ticks mark the logic levels $\{0, 0.3, 0.5, 0.8, 1.0\}$. In both low cases, $\ell = 1.00$ but $c = 0.00$, so $s = 0.00$. Case 3 omits the required reactant SMILES. Case 4 proposes a different disconnection.

Evaluation runs and control conditions. The experiment set contains 39 sample-one runs, nine $K = 3$ runs, nine $K = 5$ runs, 15 Qwen3 scale runs, 12 no-history/no-thinking runs, eight question-only runs on four endpoints, two no-history/thinking runs, two gold-history/thinking runs, two model-history runs, six GPT-judge runs, and one frozen-answer rescoring run covering two generators. Thinking configurations also change the generation budget from 512 to 65,536 tokens.

A.7 HUMAN REVIEW PLATFORM AND OBSERVED WORKLOAD

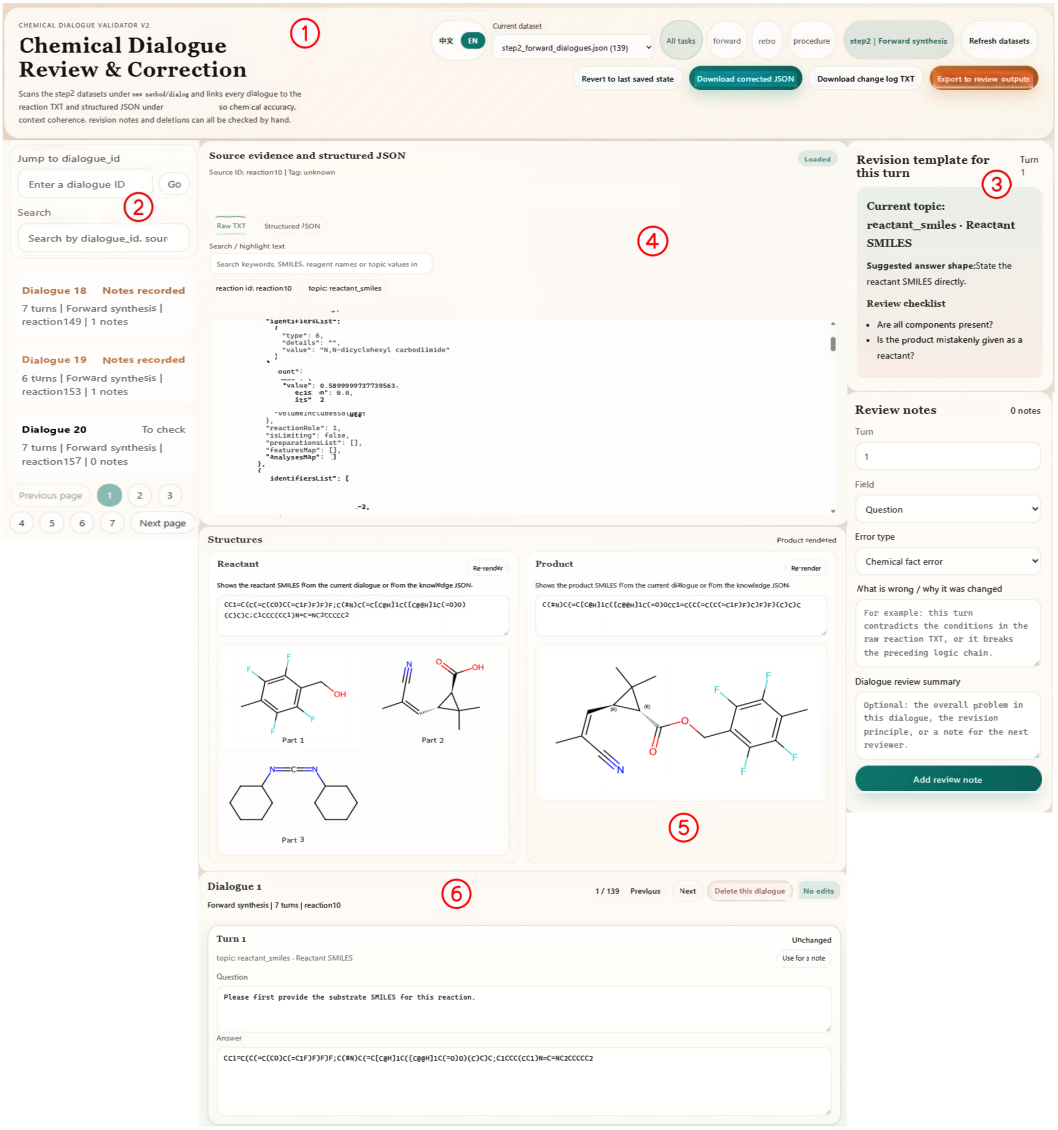


Figure 6: The validation platform used for review. ① dataset and task filter with the dialogue count, the interface language, and four actions (revert to the last saved state, download corrected JSON, download the change log, export review outputs); ② searchable dialogue list with direct `dialogue_id` jump, per-dialogue status, and pagination; ③ revision template for the current turn, with its topic, suggested answer shape, and checklist, next to the review-note panel (turn, field, error type, reason, dialogue summary); ④ source evidence beside the structured knowledge record, with a raw-text/JSON switch and keyword highlighting; ⑤ RDKit-rendered reactant and product cards with the SMILES they render; ⑥ dialogue and turn editor with navigation, per-turn status, question and answer fields, and use-for-a-note. Local paths in the screenshot are masked.

The local validation platform combines an HTTP backend with a browser interface, and Figure 6 shows the screen the reviewers worked in. Six visible regions keep evidence adjacent to the editable answer: ① the top bar, with the interface language, the dataset selector and its dialogue count, the task filter, the active view, and four actions (revert to the last saved state, download corrected JSON, download the change log, export review outputs); ② a searchable dialogue list with direct `dialogue_id` navigation, a status label and metadata for every dialogue, and pagination; ③ a revision template for the current turn, giving its topic, the suggested answer shape, and a short

checklist; ④ the source evidence, with the source identifier, a raw-text/JSON switch, and keyword highlighting over the structured record; ⑤ RDKit-rendered reactant and product cards that print the SMILES they render; and ⑥ the dialogue and turn editor, with dialogue navigation, per-turn edit status, question and answer fields, and a use-for-a-note action. Notes are written in the panel beside ③, which records the turn, the field, one of four error types, the reason, and a dialogue summary. Corrected dialogue records are separated from a `review_meta.json` review sidecar and a readable change log, and deleted records retain identifiers, source IDs, reasons, and timestamps in the review trail.

Reviewing one dialogue follows the same sequence in every topic. The reviewer opens the dialogue, reads the raw reaction record next to the structured knowledge, and checks every turn against them; each problem becomes a note that records the field, the topic, and one of four error types (chemical fact, overly general answer, terminology or phrasing, task-logic shift). Where the record supports a correction, the reviewer edits the question or the answer; otherwise the note remains an issue record. The reviewer then keeps or deletes the dialogue. Every action is appended to the change log with the original text, the revised text, the reason, and a timestamp, so the review can be audited turn by turn.

Table 13: Recorded review workload. Pre-delete is reconstructed as retained dialogues plus deletion-log entries. Modified dialogues and refined turns are the platform's record of every dialogue and turn changed during review.

View	Pre-delete	Retained	Deleted	Modified dlg.	Refined turns
Forward	150	139	11	115	201
Procedure	124	120	4	56	157
Retrosynthesis	142	137	5	120	246
Total	416	396	20	291	604

The recorded workload is substantial. Review covered 416 candidate dialogues and retained 396(20 deleted). The review platform records 291 of the 396 retained dialogues (73.5%) as modified, across 604 turns, and reviewers wrote 537 notes, of which 314 (58.5%) record a chemical-fact issue, followed by overly general answers (39), terminology or phrasing (28), and task-logic shifts (46) and 110 other issues. Effort concentrates in retrosynthesis, where 120 of 137 retained dialogues (88%) were modified, and in reactant-SMILES topics, which account for 114 notes (21.2%). Five reviewers majoring in chemistry, with relevant backgrounds in basic and organic chemistry, spent a workweek, totaling 200 person-hours.

Table 14: Traceable before–after review corrections.

Location	Before	After and rationale
Forward D42 T2, transformation	"Acylation of an amine to form an amide."	"Aryl chloride to aryl amine (Buchwald–Hartwig amination)"; the original label contradicted the reaction record.
Forward D43 T1, reactants	An incompatible set included a non-core component.	The answer was replaced with the core substrate components; K_2CO_3 was separated as a base/work-up component rather than a core reactant.
Retro D21 T4, disconnection	The explanation began from an incorrectly selected reactant.	The revision derives a C–N cut from the target product, preserving the intended product-oriented dependency.

Figure 7 shows the two ways review handled a problem, with the state before review on the left and the state after review on the right. In forward-synthesis dialogue 76, turn 4, a context-memory turn, the recorded reaction type contradicted the source record and the answer was replaced by the copper-mediated O-arylation that the record describes. In retrosynthesis dialogue 22 the source reaction is a two-step sequence, so the dialogue was removed instead of repaired; its identifier, source reaction, reason, and removal time stay in the collection, together with the other 19 removed dialogues.

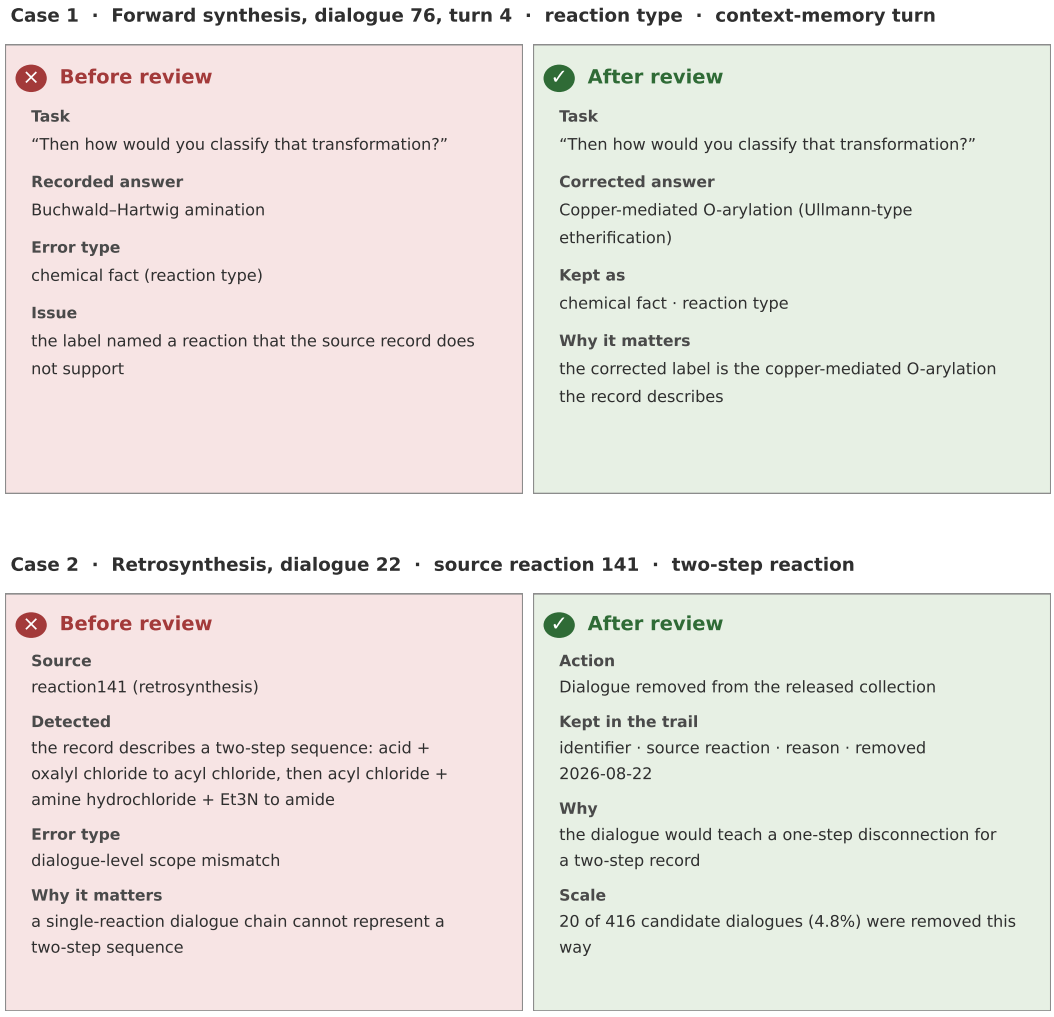


Figure 7: Two examples of review actions. Each case shows the state before review (left, red) and after review (right, green). Case 1 (top): in forward-synthesis dialogue 76, turn 4, the recorded reaction type did not match the source record. The reviewer replaced it with the copper-mediated O-arylation described in the record. Case 2 (bottom): the source reaction of retrosynthesis dialogue 22 is a two-step sequence. A single-reaction dialogue cannot represent it, so the dialogue was removed. Its identifier, source reaction, reason, and removal time remain in the review trail.

A.8 SAMPLE-ONE TOPIC AND CAPABILITY RESULTS

All Sample-1 tables use gold history, no thinking, $K = 1$, $T = 0.7$, and the DeepSeek-V4-Flash judge. Seed topics are omitted because the first turn is not scored. Throughout the performance tables, **red** and **blue** denote the highest and second-highest distinct value in each comparable metric column; tied values receive the same rank color, and difference, count, workload, and association columns carry no rank color. Figure 8 renders the same runs as four heatmaps: one block per task view with its topics and one block for the six capabilities, with the 13 models in the order of Table 1 and one shared 0–100 color scale. Figure 9 gives their weighted chemical, logic, and overall averages as bars.

Table 15: Forward-synthesis sample-one results (%).

Model	Cat.	Solv.	Temp.	FG conv.	Product	Type	Overall
DeepSeek-v4-flash	84.2	36.2	50.9	50.3	18.0	48.7	42.2
DeepSeek-v4-pro	**99.0**	48.1	**64.1**	55.4	31.7	47.8	50.8
DeepSeek-v3.2	62.2	42.1	53.7	35.1	12.9	47.1	38.8
GLM-5.2	**89.2**	51.5	**64.0**	54.6	28.8	47.8	50.7
GPT-5.5	58.3	**53.6**	57.1	**67.5**	25.9	49.0	52.4
GPT-5.6-luna	72.5	40.8	60.7	51.1	24.5	**49.6**	46.4
GPT-5.6-sol	50.0	47.0	61.4	**69.1**	32.4	48.6	**53.4**
GPT-5.6-terra	65.4	46.4	58.5	51.4	18.1	**49.1**	46.0
Kimi-K2.6	76.0	36.3	55.6	63.0	**42.4**	47.3	50.1
Qwen3.7-flash	50.0	45.9	56.7	28.3	10.8	42.7	37.3
Qwen3.7-max	83.3	**54.9**	60.2	54.4	30.2	48.8	50.9
Qwen3.8-flash	75.8	45.8	58.0	39.0	12.9	46.0	41.3
Qwen3.8-max	83.3	49.7	62.2	61.5	**34.5**	48.8	**52.8**

Table 16: Reaction-procedure sample-one results (%).

Model	Purif.	Control	Medium	Time	Work-up	Yield	Overall
DeepSeek-v4-flash	**62.0**	34.9	84.4	87.7	**30.1**	15.9	51.7
DeepSeek-v4-pro	**62.0**	30.8	**85.9**	87.4	29.6	20.5	51.8
DeepSeek-v3.2	56.0	32.6	83.8	87.1	25.8	8.4	48.9
GLM-5.2	**62.0**	**38.0**	85.4	**90.2**	**30.9**	16.3	**53.3**
GPT-5.5	56.0	30.6	84.3	87.1	27.5	**21.0**	50.9
GPT-5.6-luna	**58.0**	33.3	84.5	88.1	28.4	15.3	51.0
GPT-5.6-sol	50.0	35.4	83.2	87.6	29.2	**24.4**	**52.7**
GPT-5.6-terra	50.0	31.8	82.0	86.7	24.8	16.0	49.2
Kimi-K2.6	**62.0**	33.8	84.7	**88.8**	22.8	20.7	51.3
Qwen3.7-flash	**62.0**	33.1	84.5	87.3	17.5	9.5	48.0
Qwen3.7-max	56.0	36.7	85.3	87.3	28.6	15.1	51.9
Qwen3.8-flash	56.0	34.9	83.5	87.6	24.3	11.3	49.7
Qwen3.8-max	56.0	**37.1**	**85.7**	88.3	28.2	18.5	**52.7**

Table 17: Retrosynthesis sample-one results (%). Precursors use the four-level component score; exact reactant-set hits are reported separately in Table 2.

Model	Disconnect	FG analysis	Precursors	Solvent	Type	Overall
DeepSeek-v4-flash	48.6	65.1	45.1	**59.7**	72.8	57.6
DeepSeek-v4-pro	51.6	59.0	50.0	57.0	77.9	58.6
DeepSeek-v3.2	37.7	42.3	29.8	54.3	55.5	43.7
GLM-5.2	43.1	52.0	45.1	57.9	68.3	52.8
GPT-5.5	**51.9**	**70.1**	44.8	50.6	80.6	58.8
GPT-5.6-luna	42.6	61.4	39.6	45.2	79.4	52.7
GPT-5.6-sol	46.3	**66.6**	**55.9**	51.0	**84.8**	**59.8**
GPT-5.6-terra	45.9	60.7	42.3	46.8	79.3	54.2
Kimi-K2.6	**52.4**	56.6	**57.8**	57.2	82.3	**60.7**
Qwen3.7-flash	26.5	40.7	21.0	59.1	62.0	41.2
Qwen3.7-max	38.0	55.4	46.9	**63.6**	**82.6**	56.3
Qwen3.8-flash	33.9	51.2	28.0	54.6	70.5	46.8
Qwen3.8-max	48.4	62.1	51.3	56.6	82.1	59.3

Table 18: Sample-one capability final scores pooled across topics (%).

Model	AR	CM	EL	EC	RC	SK	Logic Avg
DeepSeek-v4-flash	**50.8**	**46.8**	63.8	**61.0**	73.7	**62.9**	**57.2**
DeepSeek-v4-pro	47.0	45.1	57.6	53.2	76.3	58.4	53.9
DeepSeek-v3.2	40.0	37.8	62.0	52.8	77.4	47.1	49.5
GLM-5.2	44.8	43.0	63.5	56.1	**81.0**	57.3	54.4
GPT-5.5	**51.1**	**47.8**	56.0	51.1	70.9	54.3	53.9
GPT-5.6-luna	49.0	45.0	53.5	45.2	75.7	55.9	52.6
GPT-5.6-sol	50.6	46.5	53.8	47.3	72.0	51.4	52.5
GPT-5.6-terra	47.9	43.0	53.7	47.3	70.1	50.6	50.6
Kimi-K2.6	46.0	39.6	**72.6**	**58.6**	79.4	**58.8**	55.2
Qwen3.7-flash	35.5	33.5	50.2	53.4	73.7	44.5	45.1
Qwen3.7-max	46.0	43.8	57.2	**58.6**	**79.8**	54.3	53.9
Qwen3.8-flash	41.8	40.8	56.3	53.3	71.3	54.4	50.2
Qwen3.8-max	48.0	46.2	**69.0**	56.8	79.6	51.5	**55.8**

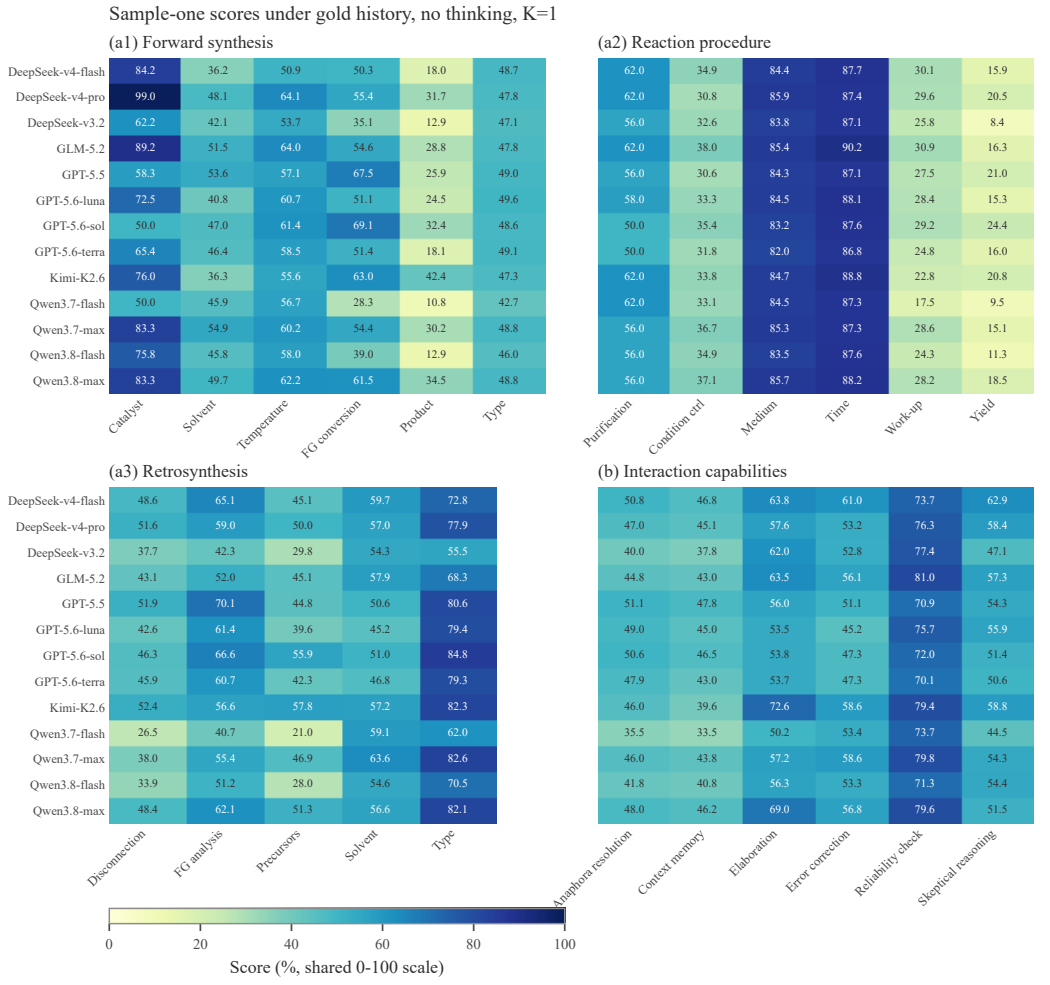


Figure 8: Sample-one topic and capability profiles under gold history, no thinking, and $K = 1$. Each block holds the 13 models in the order of Table 1: (a1) forward-synthesis topics, (a2) reaction-procedure topics, (a3) retrosynthesis topics, and (b) the six interaction capabilities. All four blocks share one 0–100 colour scale; the model labels appear on the left-hand blocks.

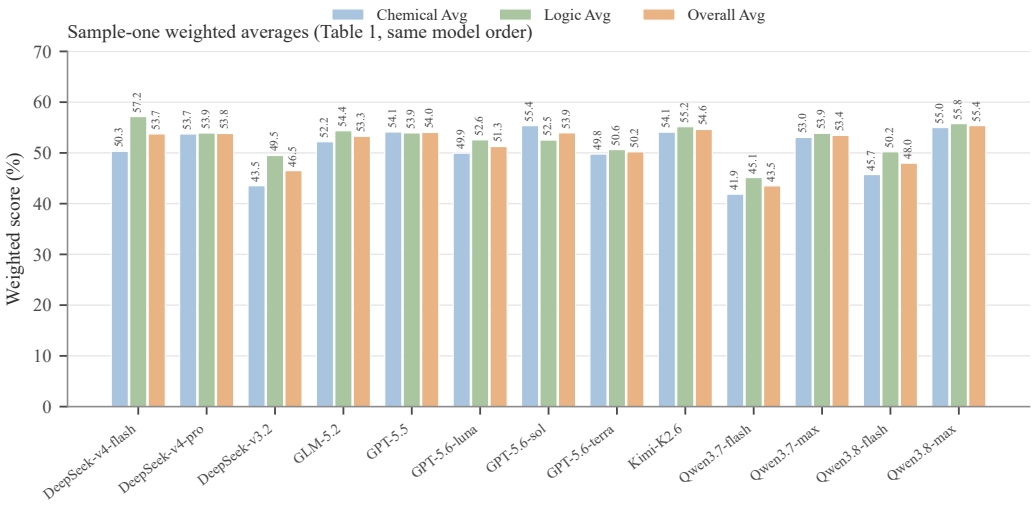


Figure 9: Sample-one weighted averages (in %) in the same model order as Table 1: the chemical average (weighted by the scored turns of the three task topics), the logic average (weighted by the capability turns of the six capabilities), and their mean.

A.9 SAMPLE-THREE AND SAMPLE-FIVE AGGREGATE, TOPIC, AND STABILITY RESULTS

Following the Sample-1 results, Tables 19 and 20 reproduce the main-table structure for best-of-three and best-of-five scores, with all values reported as percentages. The six subsequent task-specific tables retain the detailed FS, RP, and RS topic breakdowns. Table 3 then places BEST@1/BEST@3/BEST@5, Mean@K, and Self-consistency@K in one comparison.

Mean@K stays close to the best-of-one score in every cell (Best@1 to Mean@5 changes between -2.4 and $+0.9$ percentage points), while self-consistency falls from $K = 3$ to $K = 5$ by 2.9–5.6 percentage points. Repeated sampling therefore enlarges the candidate pool rather than improving ordinary candidate quality. Figure 10 plots one panel per topic with the Best@1, Best@3, and Best@5 scores of DeepSeek-V4-Pro, GPT-5.6-Sol, and Qwen3.8-Max, color marking the model; Table 3 reports the Mean@K and self-consistency columns behind those curves.

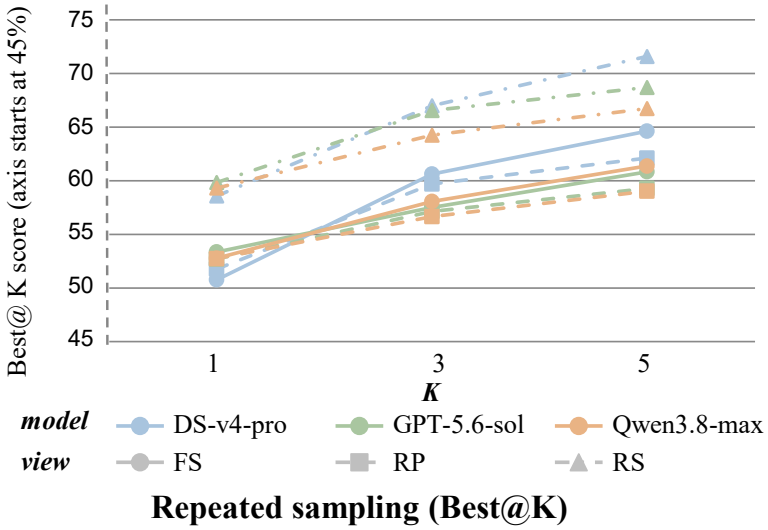


Figure 10: Best@K task-view scores of three models at $K \in \{1, 3, 5\}$. Colour marks the model, and line style and marker mark the task topic. The vertical axis starts at 45%.

Table 19: Best-of-three performance in the same aggregate format as the main sample-one table. Chemical Avg and Logic Avg use effective-turn weights; Overall is their simple mean(%).

Model	Chemical Topics				Interactive Capabilities							Overall Avg
	FS	RS	RP	Chemical Avg	AR	CM	EL	EC	RC	SK	Logic Avg	
DeepSeek-v4-pro	**60.6**	**67.0**	**59.7**	**62.5**	**55.4**	**51.4**	**71.0**	**59.6**	**82.2**	**69.4**	**62.2**	**62.3**
GPT-5.6-sol	57.5	**66.6**	**57.1**	**60.5**	**55.8**	**53.2**	58.9	48.0	76.3	55.9	57.4	58.9
Qwen3.8-max	**58.1**	64.3	56.7	59.7	52.0	50.1	**75.7**	**68.3**	**83.8**	**60.5**	**61.6**	**60.7**

Table 20: Best-of-five performance in the same aggregate format as the main sample-one table. Chemical Avg and Logic Avg use effective-turn weights; Overall is their simple mean(%).

Model	Chemical Topics				Interactive Capabilities							Overall Avg
	FS	RS	RP	Chemical Avg	AR	CM	EL	EC	RC	SK	Logic Avg	
DeepSeek-v4-pro	**64.6**	**71.6**	**62.1**	**66.2**	**57.2**	**55.0**	**77.6**	**59.1**	**83.8**	**72.7**	**64.9**	**65.5**
GPT-5.6-sol	60.8	**68.7**	**59.3**	**63.0**	**57.3**	**54.7**	61.5	47.7	78.4	57.8	58.9	61.0
Qwen3.8-max	**61.4**	66.7	59.0	62.5	53.6	50.6	**78.9**	**73.8**	**85.2**	**63.5**	**63.7**	**63.1**

Table 21: Forward synthesis, best-of-three score(%).

Model	Cat.	Solv.	Temp.	FG conv.	Product	Type	Overall
DeepSeek-v4-pro	**91.7**	**60.8**	**70.0**	**69.3**	**43.9**	**51.4**	**60.6**
GPT-5.6-sol	**75.0**	52.5	**65.7**	**75.6**	33.8	49.2	57.5
Qwen3.8-max	66.7	**57.4**	64.9	**71.3**	**38.1**	**50.2**	**58.1**

Table 22: Reaction procedure, best-of-three score(%).

Model	Purif.	Control	Medium	Time	Work-up	Yield	Overall
DeepSeek-v4-pro	**50.0**	**45.8**	**87.7**	**90.9**	**38.7**	**31.1**	**59.7**
GPT-5.6-sol	**56.0**	42.1	84.2	**89.2**	**38.8**	**28.0**	**57.1**
Qwen3.8-max	**56.0**	**42.2**	**86.9**	87.8	36.8	24.1	56.7

Table 23: Retrosynthesis, best-of-three score(%).

Model	Disconnect	FG analysis	Precursors	Solvent	Type	Overall
DeepSeek-v4-pro	**58.8**	**69.4**	**58.6**	**67.3**	83.8	**67.0**
GPT-5.6-sol	**60.4**	**74.5**	**60.5**	55.9	**85.1**	**66.6**
Qwen3.8-max	57.2	64.7	52.4	**63.7**	**85.8**	64.3

Table 24: Forward synthesis, best-of-five score(%).

Model	Cat.	Solv.	Temp.	FG conv.	Product	Type	Overall
DeepSeek-v4-pro	**94.7**	**67.4**	**71.3**	**76.6**	**46.8**	**51.8**	**64.6**
GPT-5.6-sol	75.0	57.8	**68.1**	**78.8**	39.6	49.2	60.8
Qwen3.8-max	**83.3**	**66.1**	68.0	72.3	**41.7**	**49.9**	**61.4**

Table 25: Reaction procedure, best-of-five score(%).

Model	Purif.	Control	Medium	Time	Work-up	Yield	Overall
DeepSeek-v4-pro	**62.0**	**51.0**	**88.6**	**91.0**	**42.5**	**32.0**	**62.1**
GPT-5.6-sol	**56.0**	**46.7**	85.1	**88.4**	**41.6**	**30.5**	**59.3**
Qwen3.8-max	**56.0**	45.4	**88.3**	88.0	38.6	29.2	59.0

Table 26: Retrosynthesis, best-of-five score(%).

Model	Disconnect	FG analysis	Precursors	Solvent	Type	Overall
DeepSeek-v4-pro	**63.1**	**73.1**	**66.9**	**69.1**	**88.8**	**71.6**
GPT-5.6-sol	62.0	**76.7**	**62.3**	59.8	**86.3**	**68.7**
Qwen3.8-max	**62.8**	66.4	56.0	**63.9**	**86.3**	66.7

A.10 QWEN3 PARAMETER SCALE

To examine how model scale affects chemical dialogue performance, we sweep the Qwen3 family across five checkpoints under the no-thinking, gold-history protocol: Qwen3-8B (dense), Qwen3-14B (dense), Qwen3-30B-A3B (mixture-of-experts, 3B activated), Qwen3-32B (dense), and Qwen3-235B-A22B (mixture-of-experts, 22B activated). Chemical Avg weights task means by effective scored turns, Logic Avg weights capability means by effective capability turns, and Overall is their simple mean. All values are reported as percentages.

As shown in Table 4 and Figure 11, every measure improves from 8B to 235B-A22B: FS from 27.0% to 36.2%, RP from 45.0% to 48.6%, RS from 25.2% to 39.3%, Chemical Avg from 31.9% to 41.0%, Logic Avg from 40.1% to 46.7%, and Overall from 36.0% to 43.8%. Retrosynthesis gains the most across the sweep (+14.1 percentage points), while reaction procedure gains the least (+3.6 percentage points). The trend is not monotone. The 30B-A3B MoE checkpoint (3B activated) reaches 38.6% Overall and 26.1% on FS, slightly below 14B dense (38.7% and 28.8%), although it already surpasses 14B dense on RS (32.3% vs. 30.4%) and RP (45.7% vs. 45.6%). At the top, 235B-A22B (22B activated) exceeds 32B dense on every measure, including Overall (43.8% vs. 41.5%), FS (36.2% vs. 31.7%), RS (39.3% vs. 35.0%), and RP (48.6% vs. 46.8%). Thus, scaling improves overall performance, but architecture and training recipe also matter: MoE can match or surpass larger dense models with fewer activated parameters.

Table 27: Forward-synthesis scores of the Qwen3 scale checkpoints (%).

Model	Cat.	Solv.	Temp.	FG conv.	Product	Type	Overall
Qwen3-8B (dense)	20.7	14.5	**53.1**	22.4	0.7	44.0	27.0
Qwen3-14B (dense)	**55.8**	19.6	46.4	27.3	**3.6**	44.9	28.8
Qwen3-30B-A3B (MoE)	39.2	18.2	37.2	24.6	2.2	47.8	26.1
Qwen3-32B (dense)	**40.2**	**24.5**	50.8	**29.8**	**3.6**	**47.9**	**31.7**
Qwen3-235B-A22B (MoE)	35.8	**21.5**	**57.5**	**36.9**	**14.4**	**48.8**	**36.2**

Table 28: Reaction-procedure scores of the Qwen3 scale checkpoints (%).

Model	Purif.	Control	Medium	Time	Work-up	Yield	Overall
Qwen3-8B (dense)	**62.0**	27.3	80.9	86.5	17.1	6.3	45.0
Qwen3-14B (dense)	**62.0**	29.5	79.4	**86.9**	18.5	**7.6**	45.6
Qwen3-30B-A3B (MoE)	50.0	29.0	**83.6**	**87.7**	15.3	5.5	45.7
Qwen3-32B (dense)	**56.0**	**32.0**	82.0	85.9	**18.9**	**7.6**	**46.8**
Qwen3-235B-A22B (MoE)	**56.0**	**33.3**	**82.4**	86.7	**23.9**	**9.8**	**48.6**

Table 29: Retrosynthesis scores of the Qwen3 scale checkpoints (%).

Model	Disconnect	FG analysis	Precursors	Solvent	Type	Overall
Qwen3-8B (dense)	22.3	15.9	10.1	**40.5**	36.5	25.2
Qwen3-14B (dense)	**31.3**	26.2	14.4	26.0	54.6	30.4
Qwen3-30B-A3B (MoE)	27.7	**28.4**	12.7	35.6	58.9	32.3
Qwen3-32B (dense)	26.6	27.3	**21.6**	36.9	**65.1**	**35.0**
Qwen3-235B-A22B (MoE)	**35.0**	**35.4**	**22.7**	**40.7**	**64.4**	**39.3**

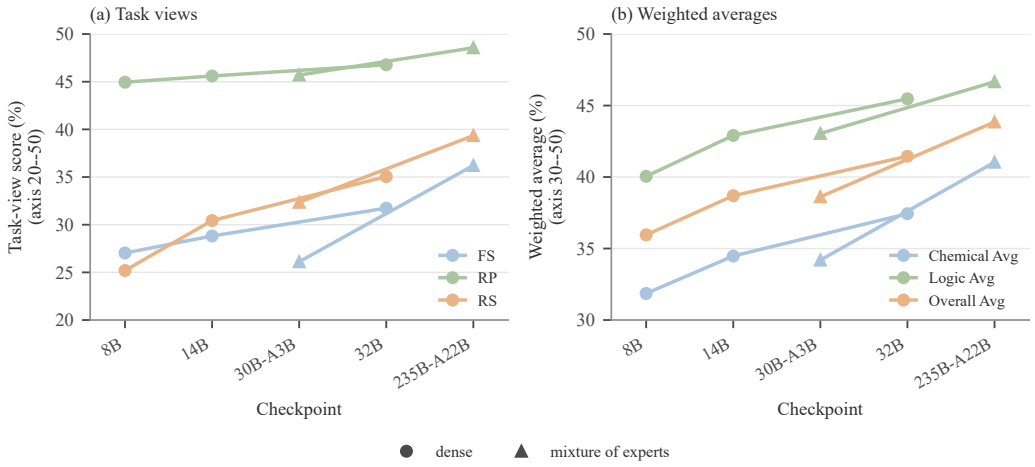


Figure 11: Scores of five Qwen3 checkpoints under gold history and no thinking (Table 4). (a) Task-view scores. (b) Weighted averages. Circles mark dense checkpoints and triangles mixture-of-experts checkpoints. The x-axis follows total parameter count but is not to scale. The vertical axes start at 20% in (a) and 30% in (b). Dense and MoE checkpoints also differ in routing, data mixture, and training recipe.

A.11 THINKING BY HISTORY CONFIGURATION

Table 30: DeepSeek-v4-pro thinking–history matrix. Exact final score is the product hit for FS and the exact reactant-set hit for RS. No-history chains contain only the final target after the seed. All values are percentages.

Task	History	Thinking	Exact final (%)	Valid N
FS	No history	No	14.4	139
FS	Gold history	No	31.7	829
FS	No history	Yes	**71.7**	138
FS	Gold history	Yes	**78.4**	829
RS	No history	No	11.7	137
RS	Gold history	No	**41.8**	791
RS	No history	Yes	34.3	137
RS	Gold history	Yes	**74.5**	791

Table 31: Matched differences in exact final score from Table 30. History effect is Gold minus No history at fixed thinking configuration. Thinking-configuration effect is thinking minus no thinking at fixed history. Differences are in percentage points(%).

Task	Hist. effect, no think	Hist. effect, think	Think effect, no hist.	Think effect, gold
FS	+17.3	+6.7	+57.4	+46.8
RS	+30.2	+40.2	+22.6	+32.7

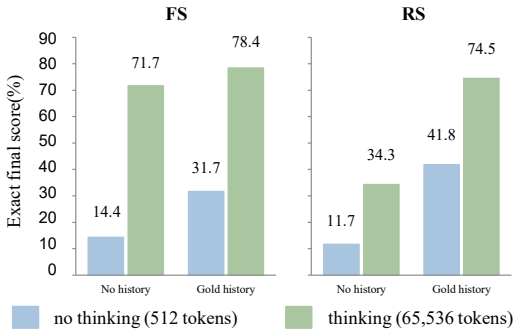


Figure 12: Exact final score of DeepSeek-v4-pro under four history–thinking configurations. FS uses the exact product-SMILES hit and RS uses the exact reactant-set hit. Blue bars use no thinking with a 512-token budget; green bars use thinking with a 65,536-token budget.

Under no history, each dialogue keeps the first seed turn of the topic and the question of the final target turn (the forward-synthesis product-SMILES turn or the retrosynthesis reactant-SMILES turn), but removes all intermediate turns; only the final target turn is scored, so N equals the number of dialogues (139 for FS and 137 for RS). Under gold history, every scored turn in the dialogue is evaluated, giving 829 and 791 turns respectively. The thinking runs set the maximum generation length to 65,536 tokens to avoid truncation, whereas no-thinking uses 512; the actual number of tokens consumed may be lower. The thinking effect therefore combines reasoning mode with a much larger maximum generation budget, and the two cannot be fully separated in this comparison. History effects within each row keep the thinking configuration fixed. For forward synthesis, thinking without history already reaches 71.7%, whereas gold history without thinking reaches only 31.7%, so the thinking effect ($+57.4$ percentage points) is much larger than the history effect ($+17.3$ percentage points) on this task. On retrosynthesis, history and reasoning are complementary rather than substitutable, whereas on forward synthesis reasoning can largely replace the intermediate information supplied by history. The thinking effect is much larger than the history effect on FS, while the history effect is larger on RS and further increases once thinking is enabled. These patterns are consistent with the chemical interpretation in the main text. Figure 12 plots the exact final score of the four configurations in the two panels, grouped by history condition with one bar per thinking setting; Table 31 lists the matched differences.

A.12 HISTORY QUALITY AND QUESTION-ONLY CONTROL

Table 32: Model-generated history compared with the unanswered question chain and the gold history under no thinking. Model history replaces every intermediate assistant answer with the model's own answer, while question-only keeps the questions but removes all intermediate answers. Colour marks the highest and second-highest value in each row. All values are percentages(%).

Model	Task	No history	Model history	Question-only	Gold history
DeepSeek-v4-pro	FS	14.4	12.9	**16.5**	**31.7**
GPT-5.6-sol	FS	15.8	**20.1**	15.8	**32.4**
DeepSeek-v4-pro	RS	11.7	6.5	**13.1**	**41.8**
GPT-5.6-sol	RS	9.5	**15.7**	8.8	**45.1**

The question-only control keeps the gold seed and every intermediate user question but removes all intermediate assistant answers, so the final target question is answered from the question sequence alone. It scores the same final target as the no-history runs (139 forward and 137 retrosynthesis dialogues), which makes the two directly comparable dialogue by dialogue. Table 2 reports the six models on both topics. Question-only matches or exceeds no history in 10 of the 12 rows: forward-synthesis differences range from 0.0 to $+5.8$ percentage points and retrosynthesis differences from -8.8 to $+4.4$ percentage points. Gold history is the highest cell in all 12 rows, at 18.0–34.5% on forward synthesis and 34.0–48.4% on retrosynthesis, which corresponds to gains over no history of $+5.8$ to $+25.9$ percentage points (forward) and $+26.7$ to $+35.6$ percentage points (retrosynthesis), and to gains over question-only of $+1.4$ to $+20.9$ percentage points (forward) and $+25.2$ to $+36.3$ percentage points (retrosynthesis). Table 32 replaces the unanswered question chain with the model's own answers for DeepSeek-v4-pro and GPT-5.6-sol and adds the gold-history column of the same sample-one runs: model history is below no history on both DeepSeek-v4-pro rows (FS 14.4% to 12.9%, RS 11.7% to 6.5%) and above it on both GPT-5.6-sol rows (FS 15.8% to 20.1%, RS 9.5% to 15.7%), so the four differences span -5.1 to $+6.2$ percentage points, and relative to question-only they span -6.6 to $+6.9$ percentage points; gold history is again the highest cell in all four rows (31.7% and 41.8% for DeepSeek-v4-pro, 32.4% and 45.1% for GPT-5.6-sol). All control runs use a single sample and no thinking, so differences of a few percentage points are not separated from sampling variation.

Model history is built by replacing every intermediate assistant answer with the answer the same model produced in the corresponding sample-one run, so the model sees its own outputs as history. This effect is model-dependent: GPT-5.6-sol improves over no history when using its own answers (FS 15.8% to 20.1%, RS 9.5% to 15.7%), whereas DeepSeek-v4-pro degrades (FS 14.4% to 12.9%, RS 11.7% to 6.5%). Incorrect intermediate content propagates into later answers and can actively mislead the model, so a model's own history is useful only when its intermediate answers are correct.

Table 33: Forward-synthesis results under model history and gold history. $\Delta(\%)$ = Gold − Model history.

Model	History	Cat.	Solv.	Temp.	FG conv.	Product	Type	Overall
DeepSeek-v4-pro	Model history	16.7	36.9	47.4	45.5	12.9	29.9	35.9
DeepSeek-v4-pro	Gold history	99.0	48.1	64.1	55.4	31.7	47.8	50.8
DeepSeek-v4-pro	Δ	+82.3	+11.2	+16.7	+10.0	+18.7	+17.9	+14.9
GPT-5.6-sol	Model history	16.7	35.0	47.1	63.9	20.1	39.2	42.8
GPT-5.6-sol	Gold history	50.0	47.0	61.4	69.1	32.4	48.6	53.4
GPT-5.6-sol	Δ	+33.3	+12.0	+14.3	+5.2	+12.2	+9.4	+10.5

Table 34: Retrosynthesis results under model history and gold history. $\Delta(\%)$ = Gold − Model history.

Model	History	Disconnect	FG analysis	Precursors	Solvent	Type	Overall
DeepSeek-v4-pro	Model history	36.9	59.0	14.4	27.8	43.3	35.7
DeepSeek-v4-pro	Gold history	51.6	59.0	50.0	57.0	77.9	58.6
DeepSeek-v4-pro	Δ	+14.7	0.0	+35.6	+29.2	+34.6	+22.9
GPT-5.6-sol	Model history	42.7	66.9	24.4	28.1	47.0	41.2
GPT-5.6-sol	Gold history	46.3	66.6	55.9	51.0	84.8	59.8
GPT-5.6-sol	Δ	+3.6	−0.4	+31.4	+23.0	+37.9	+18.7

Table 35: Capability final scores under model history and gold history, pooled over forward synthesis and retrosynthesis. $\Delta(\%)$ = Gold − Model history.

Model	History	AR	CM	EL	EC	RC	SK	Logic Avg
DeepSeek-v4-pro	Model history	43.3	45.4	38.6	21.6	31.8	43.7	39.3
DeepSeek-v4-pro	Gold history	53.2	53.6	57.5	55.0	72.4	66.6	58.3
DeepSeek-v4-pro	Δ	+9.9	+8.2	+18.9	+33.4	+40.7	+22.9	+19.1
GPT-5.6-sol	Model history	55.3	50.7	47.2	16.7	32.8	40.8	43.8
GPT-5.6-sol	Gold history	60.4	54.6	57.5	46.9	72.7	56.0	58.1
GPT-5.6-sol	Δ	+5.2	+3.9	+10.3	+30.2	+39.9	+15.2	+14.4

Table 36: Model history against gold history on the two matched models. Chemical Avg is weighted by the scored turns of the two topics available in both conditions (829 forward, 791 retrosynthesis); Logic Avg is weighted by the capability turns of those topics; Overall Avg is their simple mean(%).

Model	History	Chemical Avg	Logic Avg	Overall Avg
DeepSeek-v4-pro	Model history	35.8	39.3	37.5
DeepSeek-v4-pro	Gold history	54.6	58.3	56.5
DeepSeek-v4-pro	Δ	+18.8	+19.1	+18.9
GPT-5.6-sol	Model history	42.0	43.8	42.9
GPT-5.6-sol	Gold history	56.5	58.1	57.3
GPT-5.6-sol	Δ	+14.5	+14.4	+14.4

A.13 VALIDATING THE AUTOMATIC JUDGE

This section reports two complementary judge experiments. The first experiment regenerates answers under each judge condition, so its differences combine generation variation with evaluator choice. The second experiment removes generation entirely by freezing the same sample-one answers and letting both judges rescore them without regeneration, which isolates the judge effect and is the setting referenced in the main text.

Table 37: Complete alternate-judge replications (regenerated answers). Each judge condition regenerated answers; differences therefore combine generation variation and evaluator choice. All values are percentages.

Generator	Judge	FS	RP	RS	Chemical Avg	Logic Avg	Overall
DeepSeek-v4-flash	DeepSeek-v4-flash	42.2	51.7	57.6	50.3	**57.2**	53.7
DeepSeek-v4-flash	GPT-5.6-sol	48.3	**53.3**	57.4	52.9	**55.6**	**54.2**
GPT-5.6-sol	DeepSeek-v4-flash	**53.4**	52.7	**59.8**	**55.4**	52.5	53.9
GPT-5.6-sol	GPT-5.6-sol	**52.9**	**54.1**	**58.6**	**55.2**	54.3	**54.8**

Table 38: Subset with exactly identical selected-answer strings across the two regenerated runs, aligned by task, dialogue ID, and turn. Delta is GPT judge minus primary judge and is in percentage points. The subset covers 355/2,329 and 1,064/2,329 turns; its primary-judge means are 82.5% and 67.3%, against 50.3% and 55.4% over all scored turns. Pearson and Spearman are correlation coefficients.

Generator	Matched N	Primary mean	GPT mean	Δ	Pearson	Spearman	Exact agreement
DeepSeek-v4-flash	355	**82.5**	**82.7**	+0.1	**0.992**	**0.999**	**97.5**
GPT-5.6-sol	1,064	**67.3**	**67.9**	+0.6	**0.977**	**0.979**	**92.1**

Among the two generators, 355/2,329 and 1,064/2,329 outputs match exactly across the paired runs. Those turns are the easier part of the sample: their primary-judge means are 82.5% and 67.3%, against 50.3% and 55.4% over all scored turns, and the two judges agree on them almost perfectly (Spearman 0.999 and 0.979; exact agreement 97.5% and 92.1%). This analysis reduces generation variation but cannot remove it entirely, because the identical strings still come from independent regenerated runs.

The frozen-answer rescoring removes generation entirely and compares the two judges on exactly the same text. Over the full sample, the score band is unchanged for 83.4% and 87.5% of turns, so 12–17% of turns move to a different band; Spearman correlations are 0.947 and 0.968, and per-task differences range from -3.1 to $+1.7$ percentage points. At the aggregate level, the judge change lowers Overall Avg by 0.9 percentage points for DeepSeek-v4-flash generation and raises it by 0.3 percentage points for GPT-5.6-sol generation. This is the strict comparison reported in the main text and shows that the reported levels and model comparisons are not artifacts of the primary judge.

Figure 13 shows the three comparisons in order. Panel (a) holds the regenerated runs of Table 37. Panel (b) sets the identical-output subset next to the frozen rescoring over all scored turns. Panel (c) holds the frozen-answer rescoring of Table 5 over every scored turn.

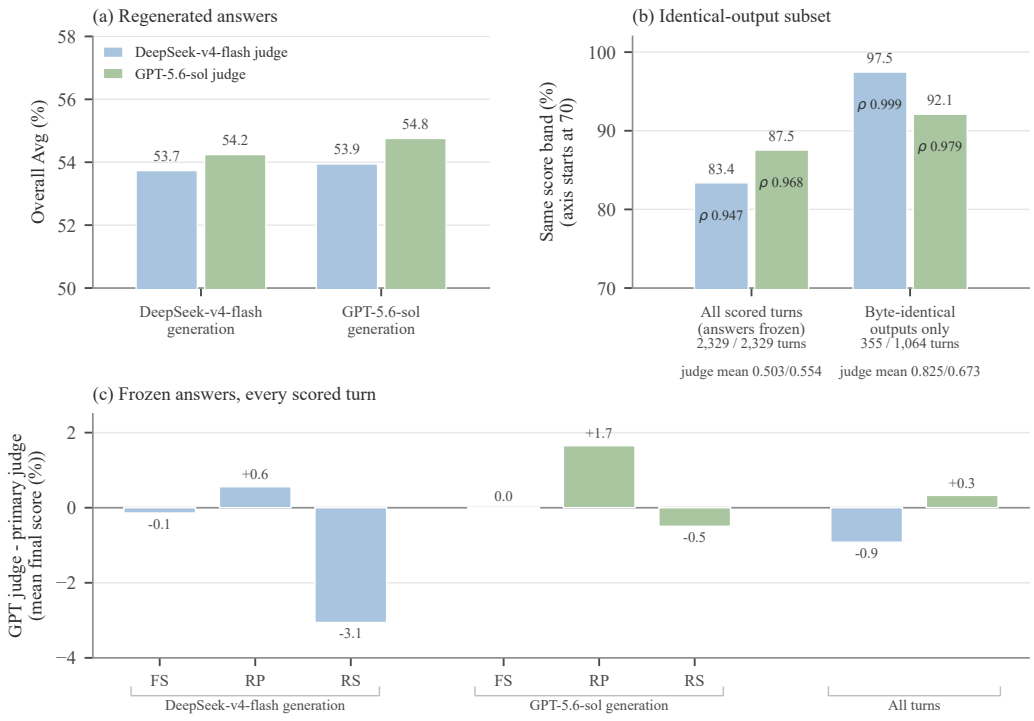


Figure 13: Judge comparisons in the order of the evidence. (a) Regenerated answers: each judge condition re-generated answers, so the Overall Avg difference of Table 37 mixes generation with scoring. (b) Identical-output subset: turns whose two regenerated runs produced the same answer text (355 and 1,064 of the 2,329 scored turns per generator), shown next to the frozen rescoring over all scored turns. Keeping the same score band covers 97.5% and 92.1% of the identical-output turns against 83.4% and 87.5% over all scored turns, and the identical-output turns are the easier part of the sample (primary-judge means 82.5% and 67.3% against 50.3% and 55.4%), with ρ 0.999 and 0.979 on them. (c) Frozen-answer rescoring: both judges score the same sample-one answers, covering every scored turn, and the turn-level differences of Table 5 stay within -3.1 and $+1.7$ percentage points.

A.14 TURN-POSITION ANALYSIS

Table 39: Mean final score by turn index across the 13 sample-one models. Turn indices follow the topic-specific logic chains: FS (functional-group conversion, reaction type, solvent, temperature, catalyst, product), RP (reaction basis, medium, condition control, time, work-up, product/yield), RS (product SMILES, functional-group analysis, disconnection, reaction type, solvent, reactant set). Some turns mix two topics. All values are percentages(%).

Topic	Turn 2	Turn 3	Turn 4	Turn 5	Turn 6	Turn 7
FS	45.1	**57.1**	45.7	50.3	**60.7**	23.1
RP	**97.2**	45.8	36.8	**81.0**	26.7	15.3
RS	57.2	37.8	**68.2**	**64.3**	58.2	34.3

Each topic follows a fixed topic chain, so a turn index carries both the amount of dialogue state already seen and the topic being asked about. The numbers behind Table 39 come from the 13 sample-one runs under gold history: for every run we take that run's own turn-wise mean final score and then average the 13 model values. Turn 1 is the seed and is not scored.

The chains differ by topic. Forward synthesis moves from functional-group conversion to reaction type, then solvent and temperature, and ends on the product. Reaction procedure moves through the reaction medium and condition control to reaction time, work-up, and the yield outcome. Retrosynthesis moves from functional-group analysis and disconnection strategy through reaction type and solvent to the reactant set. Several turns mix two topics, and the labels under Figure 14 name the topics that account for at least 40% of the turns at that position, in order of frequency.

The curves are non-monotonic and follow topic difficulty more than context length. All three topics bottom out on the final structural target: the product hit reaches 23.1% in FS, the yield outcome 15.3% in RP, and the reactant-set hit 34.3% in RS. The peaks are the property and condition turns: reaction medium 97.2% and reaction time 81.0% in RP, reaction type or disconnection 68.2% in RS, and temperature 60.7% in FS. FS climbs into turn 6 before dropping, RP rebounds to 81.0% at turn 5 and then falls to 26.7% at work-up, and RS peaks in the middle of the chain. The last turn also covers fewer dialogues than the earlier ones (FS 134 of 139, RP 109 of 120, RS 107 of 137), so the final point combines a harder topic with a smaller subset of dialogues. This smaller subset does not change the overall pattern, because the final structural targets are consistently the weakest across all three topics.

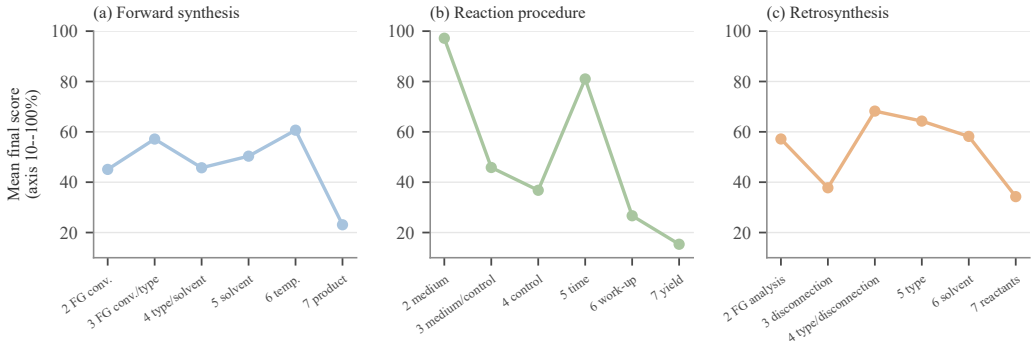


Figure 14: Mean final score by turn index for the three topics across the 13 sample-one models. Each panel is one topic; the x-axis labels give the turn index together with the topics that make up at least 40% of the turns at that position, the vertical axis starts at 10%, and the last turn is covered by fewer dialogues than the preceding turns (134 of 139 in FS, 109 of 120 in RP, 107 of 137 in RS).

A.15 INDEPENDENT HUMAN LLM-AS-A-JUDGE CONSISTENCY STUDY

The human-consistency study contains 20 questions: four forward, seven procedure, and nine retrosynthesis items, with four questions from each of five chemistry categories (reaction type, functional-group conversion, disconnection strategy, condition control, and other descriptive topics). Each item displays a prompt, the reference answer, and four or five anonymised answers selected across the automatic-score range. Candidates are taken from sample-one records that at least ten models answered with at least five distinct automatic scores; list-like and near-duplicate answers are removed, and the selected answers span the quality bands. Model identity, automatic score, and scoring mode remain in a separate key, and display order is permuted deterministically from the study, rater, and question identifiers. Raters therefore saw neither the model identity nor the automatic score when assigning bands and rankings.

A rater assigns each answer one of five bands $\{0, 0.2, 0.5, 0.8, 1.0\}$ (irrelevant, minimally correct, partially correct, mostly correct, complete) and orders the answers within the question, which gives 88 answer judgements and 20 rankings per rater; 12 items carry four answers and eight carry five. The browser interface stores progress per rater, requires a complete study before export, and writes JSON or CSV records of question, answer key, rank, band, and an optional note. Validation rejects unknown keys, missing answers, duplicate ranks, invalid bands, duplicate rater identifiers, or a study-identifier mismatch (Figure 15 shows the interface). Analysis reports per-question Spearman and Kendall τ, score-based correlation, score-difference-weighted pairwise concordance, inversion rate, automatic top-1 in human top-1/top-2, Kendall W, ICC(2,1), per-category medians, and a 2,000-resample bootstrap interval for median Spearman. These statistics follow the standard rank-agreement and reliability literature (Cohen, 1960; Krippendorff, 2004; Kendall & Babington Smith, 1939; Shrout & Fleiss, 1979). A pre-specified rule labels the study as supporting agreement when the median Spearman correlation is at least 0.6 with at least 75% positive question-level correlations, as borderline when the median is at least 0.3 with at least 50% positive correlations, and as not supporting agreement otherwise.

Five raters completed the study. All were chemistry master's or bachelor students who had taken basic chemistry and organic chemistry courses, and each spent 25–40 minutes on the questionnaire. All five exports passed validation, with no rejected file and no duplicate rater identifier, which gives 5 raters, 440 answer judgements, and 100 rankings in total.

The raters agreed with the automatic scores and with each other. The median question-level Spearman correlation between the automatic ranking and the rater panel is 1.00 (95% bootstrap interval 0.92–1.00) and all 20 items are positive, so the pre-specified rule labels the study as supporting agreement. The mean Kendall τ is 0.91, the correlation between automatic scores and mean human scores is 0.99, the mean score-difference-weighted pairwise concordance is 98.0%, and the mean inversion rate is 4.5%. Agreement is high inside every chemistry category, with median Spearman correlations of 0.87 to 1.00 (Table 41). Table 6 breaks the same comparison down by rater: between 9 and 12 of the 20 questions reproduce the judge order exactly, 65 to 70 of the 88 answers carry the judge band, the mean rank difference stays between 0.23 and 0.39, and the per-rater correlations range from 0.76 to 0.90 (Kendall τ) and from 0.83 to 0.95 (Spearman ρ). The automatic top-1 answer falls inside the human top-1 for 19 of the 20 items and inside the human top-2 for all 20. Figure 16 shows the per-item correlations and the automatic-versus-human score agreement. The panel is itself highly consistent (median Spearman 1.00, mean Kendall τ 0.91, pairwise concordance 98.0%, Kendall W 0.89, ICC(2,1) 0.93). This internal consistency indicates that the human reference is stable rather than driven by individual noise.

Table 40: Human consistency results over the five raters. Judgements and rankings are totals; each rater produced 88 answer judgements and 20 rankings. Kendall W and ICC(2,1) are averaged over the 20 questions. Pairwise concordance is reported as a percentage; Kendall W, ICC(2,1), and Spearman correlations are reported as coefficients.

Questions	Judgements	Rankings	Raters	Median Spearman	Mean Kendall τ	Pairwise conc. (%)	Kendall W/ICC
20	440	100	5	1.00	0.91	98.0	0.89/0.93

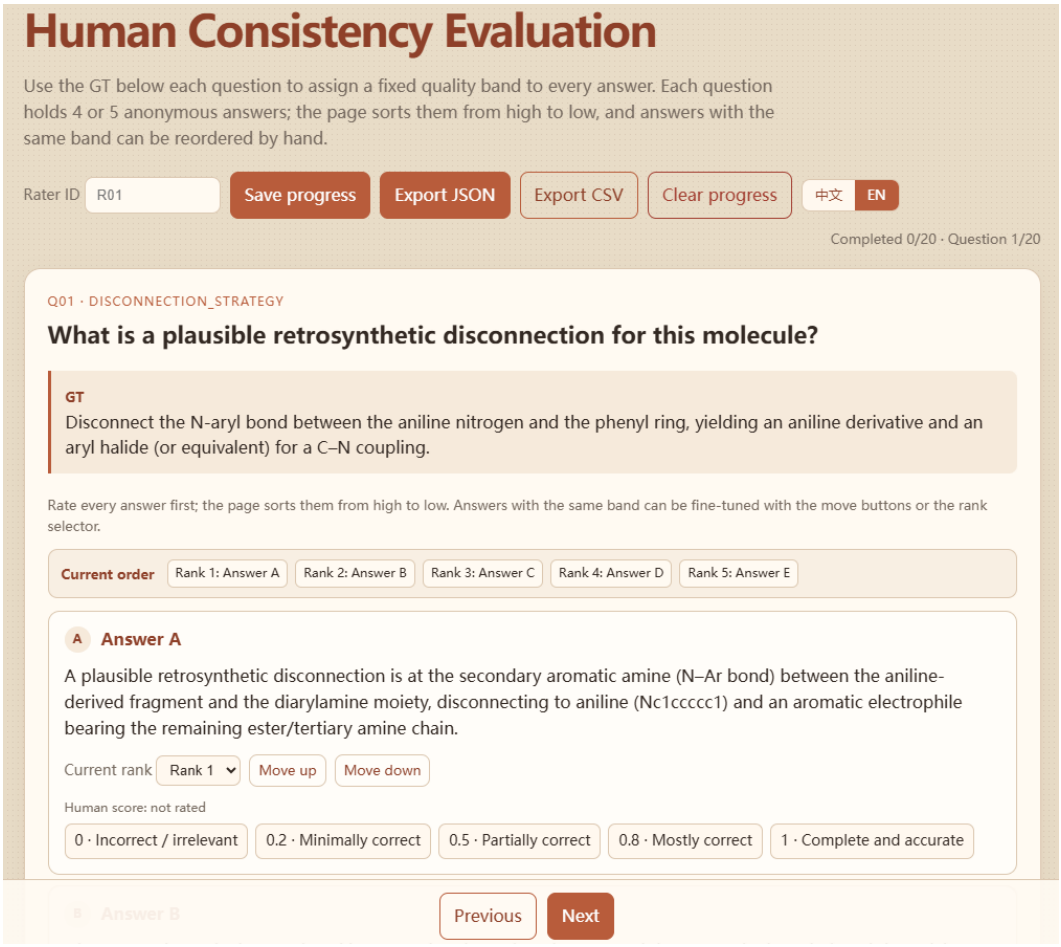


Figure 15: The questionnaire interface used for the human study. Each question shows the ground-truth reference (GT) followed by four or five anonymised answers; the rater assigns one quality band to every answer and then orders the answers, and the page keeps the current rank of each answer next to move-up and move-down controls. The header records the rater identifier, saves progress, and exports JSON or CSV, and the counter reports the completed questions.

This study is separate from the GPT-judge comparison: it uses human raters only, and the agreement statistics above are computed from their exports alone.

Table 41: Human consistency by chemistry category, four questions each. Median Spearman is a coefficient; pairwise concordance is reported as a percentage and averaged over the four questions of the category; the last column counts items whose automatic top-1 answer falls inside the human top-1.

Category	Questions	Median Spearman	Pairwise conc. (%)	Top-1 items
Reaction type	4	0.99	98.0	4/4
Functional-group conversion	4	0.87	97.0	3/4
Disconnection strategy	4	0.95	97.0	4/4
Condition control	4	1.00	99.0	4/4
Other descriptive	4	1.00	99.0	4/4

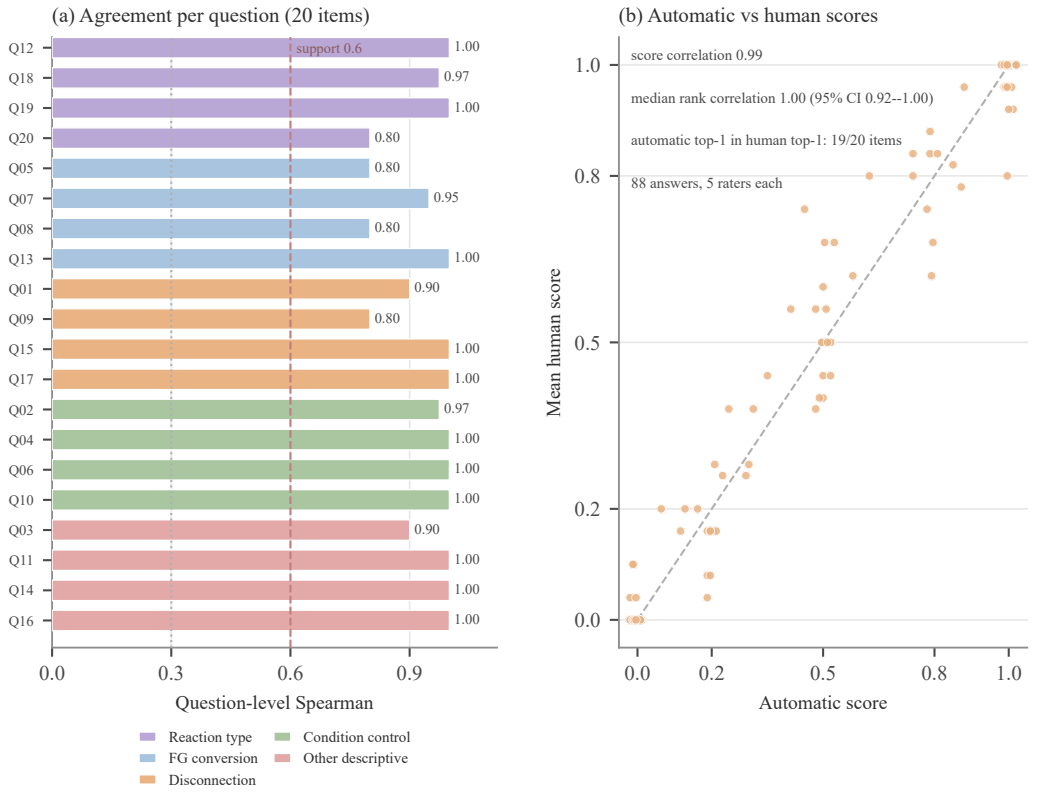


Figure 16: Agreement between the automatic judge and five human raters. (a) Spearman correlation per question between the automatic ranking and the panel ranking. The panel ranking orders answers by their mean rank over the five raters. Colour marks the chemistry category. The dashed line marks the 0.6 support threshold, and the dotted line marks the 0.3 borderline threshold. (b) Automatic score against the mean human score for all 88 answers. The dashed line is the identity line. The automatic top-1 answer falls inside the human top-1 for 19 of 20 questions and inside the human top-2 for all 20.