LLMs as Lab Engineers: A Benchmark for Troubleshooting in Chemistry

Anonymous ACL submission

Abstract

We introduce ChemBench, a comprehensive benchmark for evaluating LLMs' capabilities in chemistry troubleshooting scenarios. Unlike existing benchmarks focused on factual knowledge, ChemBench assesses model abilities to provide contextualized, practical guidance for complex analytical chemistry challenges, particularly for troubleshooting of liquid chromatography coupled mass spectrometry (LC-MS) instruments. We evaluate three enhancement approaches: chemistry-specialized models, human-guided Chain-of-Thought reasoning, and Retrieval-Augmented Generation (RAG). Our findings reveal that general-purpose commercial models often outperform domain-specialized ones, while RAG and reasoning significantly improve performance. The six-dimension evaluation framework (specificity, correctness, usefulness, feasibility, misinformation risk, and error handling) provides valuable insights into LLMs' real-world utility for chemistry researchers, establishing a foundation for developing more effective AI assistants for scientific research.[1]

1 Introduction

Artificial intelligence (AI) assistants, particularly advanced large language models (LLMs), are increasingly transforming scientific research. Evolving beyond simple chatbots, they now support diverse research tasks and show promise as "lab-side engineers" for real-time troubleshooting. However, rigorous evaluation is necessary before laboratories can confidently rely on these AI systems in high-stakes experimental environments.

Recent AI research offers several promising strategies for enhancing LLM performance in chemistry: domain-specialized models (e.g., ChemDFM (Zhao et al., 2025), ChemLLM (Zhang et al., 2024)) improve factual and terminological precision; expert-guided Chain-of-Thought (CoT) (White et al., 2023) structures reasoning processes to mirror chemists' problem-solving approaches; and retrieval-augmented generation (RAG) (Wang et al., 2024) enables models to access external knowledge bases, overcoming limitations in parametric knowledge while maintaining alignment with current instrument specifications.

Despite these advancements, current LLM benchmarks are generally inadequate for complex scientific tasks. They often focus on simple, factual questions, neglecting to test crucial abilities like seeking clarification or offering practical, vendor-specific guidance for lab personnel. This means an LLM can score highly on standard QA tasks yet still provide shallow or inaccurate guidance on actual laboratory challenges.

To address this critical gap, we introduce ChemBench, a benchmark specifically designed to evaluate LLMs' capabilities in chemistry troubleshooting scenarios. Troubleshooting complex analytical platforms, such as liquid chromatography coupled mass spectrometry (LC-MS), serves as an ideal testbed for a realistic benchmark, as it demands complex, often multi-turn dialogue from symptom to solution. By partnering with analytic chemistry specialists , we develop a rubric-based evaluation framework with six metrics: *specificity/-contextualization*, *correctness*, *usefulness*, *feasibility*, *misinformation risk*, and *error handling*, each scored on a 1–5 scale by expert chemists.

Through this framework, we explore three promising enhancement approaches: chemistry-specialized models, human-guided Chain-of-Thought reasoning, and RAG systems, by addressing three core research questions. First, we investigate whether domain-specialized LLMs outperform general models in chemistry troubleshooting tasks (**RQ1**). Second, we examine how human-

[1]The data is provided in the supplementary material. The code is available anonymously at https://anonymous.4open.science/r/ChemBench-1E32 for review purposes, and the public repository will be released after acceptance.

guided Chain-of-Thought reasoning enhances performance compared to single-step answers (**RQ2**). Finally, we evaluate the impact of RAG on troubleshooting effectiveness (**RQ3**). Through analyzing the research questions, our key findings are outlined below:

- General-purpose models like GPT and Claude consistently outperform specialized chemistry models on most criteria (with the exception of the tool-augmented ChemCrow).
- The models' Error Handling capability shows the most significant performance gap between model types, with only top-tier models demonstrating strong capabilities to acknowledge limitations and request clarifying information;
- Human-guided Chain-of-Thought reasoning significantly improves performance, particularly for mid-tier models like Gemini, suggesting that reasoning capabilities may be more valuable than domain-specific training.
- RAG integration substantially enhances answer quality across all metrics by incorporating specialized knowledge sources.

ChemBench provides valuable insights for deploying LLMs in real-world chemistry research environments and establishes a solid foundation for developing more effective AI assistants for complex scientific research.

2 Related Work

Large Language Models in Chemistry. Two main approaches have emerged to integrate Large Language Models (LLMs) with chemistry.

First, LLMs augmented chemistry tools, such as ChemCrow for synthesis planning and drug discovery (Bran et al., 2023), and an autonomous multi-LLM lab agent (Boiko et al., 2023), leverage reasoning while mitigating domain-specific limitations. Second, fine-tuned chemistry LLMs such as GPT-based property predictors and planners (Jablonka et al., 2024), ChemDFM (Zhao et al., 2025), and ChemLLM (Zhang et al., 2024) show strong performance on structured tasks.

However, current benchmarks for evaluating LLMs in chemistry are inadequate. Existing assessments like ChemLLMBench (Guo et al., 2023) and capability studies (White et al., 2023; Castro Nascimento and Pimentel, 2023) focus on basic chemical knowledge but fail to test interactive, step-by-step troubleshooting questions. No standard framework exists for evaluating open-ended troubleshooting scenarios that reflect real research work. Just as OpenAI released HealthBench (Arora et al., 2025) for evaluating LLMs in medical domains, chemistry needs a comprehensive benchmark that tests LLMs on realistic troubleshooting tasks. More related work and the details can be found in Appendix A.

3 Design of ChemBench

3.1 Dataset Construction

Human-generated Questions. Guided by chemists (backgrounds detailed in Section 6), we collected approximately 50 open-ended chemistry troubleshooting questions through discussions with them. Of these questions, half are are accompanied by expert-provided, gold-standard answers, while the rest are unanswered. These questions focus on analytical chemistry challenges, particularly chromatography and mass spectrometry issues that chemists routinely encounter in laboratory settings. For example, a typical question addresses challenges in high-performance liquid chromatography (HPLC): "*I have 4 components in my sample, but no matter what HPLC conditions I try, I only see a maximum of 3 peaks. I am using a C18 column with a water/Acetonitrile gradient from 70-90%. What should I do?*" More details of the collected questions are presented in Section B.1. After collecting these questions, we subjected the responses from various LLMs to qualitative evaluation by chemistry experts using the rubric-based criteria, introduced in Section 3.2.

AI-generated Questions. To expand our evaluation dataset and establish a comprehensive RAG training corpus, we augmented our human-generated questions with AI-generated question-answer pairs derived from authoritative chemistry troubleshooting documentation. Using a collection of troubleshooting manuals and technical guides (Table 2) from instrument manufacturers, we implemented a multi-stage generation and filtering pipeline detailed in Section B.2.

Source PDFs were first segmented into 2,000-token chunks with 200-token overlaps to preserve procedural coherence. For each chunk, we prompted `GPT-4o-mini` to generate a single question-answer pair using a carefully designed prompt[2] that elicited natural, first-person laboratory scenarios with comprehensive, step-by-step answers grounded strictly in the source content.

[2]The complete prompt is provided in Section B.2.

Then, from an initial pool of 200 generated pairs, we applied a two-stage filtering protocol:

Stage 1 (Automated Filtering): LLM-based scoring evaluated each pair on three 5-point criteria: groundedness (whether the answer is supported by source context), relevance (practical usefulness to users), and standalone clarity (whether the question is understandable without source context). Pairs scoring ≥ 4 on all criteria advanced to Stage 2, reducing the pool to 135 candidates.

Stage 2 (Expert Review): Five chemists independently evaluated the remaining pairs across four criteria: (i) question clarity and specificity, (ii) factual accuracy of the answer, (iii) directness in addressing the question, and (iv) sufficiency of the source chunk for generating a complete answer. Pairs scoring ≥ 3 from all reviewers were retained.

This process yielded 113 high-quality QA pairs deemed most relevant and valuable by chemistry experts. This expanded dataset serves dual purposes: it provides additional test cases for evaluating LLM performance across a wider range of chemistry troubleshooting scenarios, and it establishes a foundation for training and refining RAG systems specialized for chemistry applications.

3.2 Rubric-based Evaluation Framework

To comprehensively assess LLM performance in chemistry troubleshooting scenarios, we have developed a multi-dimensional evaluation framework consisting of six rubric criteria. Each metric is scored on a 5-point scale (1-5, 1 = worst performance, 5 = best performance).

Specificity/Contextualization evaluates whether responses are tailored to the specific troubleshooting scenario rather than providing generic information, and whether answers address the particular experimental conditions in the query.

Correctness assesses factual accuracy and chemical soundness, including proper application of chemical principles, accurate mechanisms, and appropriate technical terminology.

Usefulness evaluates how helpful the response would be for a chemist facing the troubleshooting scenario, including clear step-by-step procedures, explanations of underlying principles, and appropriate diagnostic tests.

Feasibility assesses whether proposed solutions are practical, implementable, and reasonable within a typical chemistry laboratory setting, considering standard practices and resource requirements.

Misinformation Risk evaluates potential negative consequences if incorrect information in the response were followed, including safety risks (e.g., incompatible chemical combinations) and experimental integrity issues (e.g., contamination risks or invalid analytical approaches).

Error Handling evaluates the LLM's ability to appropriately acknowledge uncertainty when information is insufficient, request clarifying information when necessary, and guide users an iterative troubleshooting process.

Each response is independently assessed by ten chemists with advanced degrees and laboratory experience. Evaluators can then provide additional details to correct or refine LLM answers, which supports the chain-of-thought (CoT) reasoning enhancement described in Section 3.3.2. To standardize this process, we developed a detailed questionnaire that clearly informed the experts about the intended use of the evaluation results and provided clarifications for each scoring metric; a screenshot of this questionnaire is included in Appendix C.

3.3 Enhancement Approaches

In this section, we explore three primary approaches to enhance LLM performance in chemistry troubleshooting contexts.

3.3.1 Chemistry-specialized Models

To evaluate whether domain-specific post-training or fine-tuning enhances LLM performance in chemistry troubleshooting, we compare chemistry-focused models against general-purpose LLMs using our rubric-based evaluation framework. We assess three chemistry-specialized LLMs: **ChemCrow** (Bran et al., 2023), which integrates GPT-4 with specialized chemistry tool; and fine-tuned chemistry LLMs including **ChemDFM** (Zhao et al., 2025) and **ChemLLM** (Guo et al., 2023). These specialized models are benchmarked against general-purpose models (e.g., GPT, Claude) to determine whether domain-specific training or tool integration provides measurable advantages for chemistry troubleshooting tasks.

3.3.2 Human-guided Chain-of-Thought

Our human-guided chain-of-thought (CoT) approach leverages structured interactions between chemists and LLMs to guide reasoning processes and improve response quality. We developed an interactive prompting template (Figure 1) that incorporates explicit expert feedback to guide LLM reasoning in chemistry troubleshooting contexts.

Response Refinement for ChemBench

<*start*>system
Experts have evaluated your response and provided feedback about its specificity, accuracy, and usefulness.

Guidelines
Please analyze the expert feedback carefully and refine your original response by following these steps:
1. Identify Key Issues: Review the feedback and identify the specific aspects that require improvement. These might include: factual inaccuracies; lack of specificity; missing context or definitions; safety considerations
2. Explanation of Refinement: - What aspect of your original response needs refinement - Why this refinement is necessary (based on expert feedback) - How your approach will change in the revised answer
3. Revised Response: Provide your comprehensive revised answer that addresses all feedback points

<*end*>

<*start*>user
[INSERT EXPERT FEEDBACK HERE]
<*end*>

Figure 1: The prompt template for response refinement

This approach creates a dynamic evaluation environment where chemists provide targeted feedback after reviewing initial LLM responses. The interactive template analyzes expert feedback into three key steps: (1) identifying specific aspects requiring refinement based on the rubric criteria, (2) explaining why refinement is necessary and how the approach should change, and (3) generating a revised response that addresses all feedback points. This approach allows us to evaluate not only final answer quality but also LLM's capacity for iterative reasoning throughout multi-turn troubleshooting scenarios where initial information may be incomplete and additional clarification is necessary.

3.3.3 RAG with Chemistry Sources

Our RAG implementation integrates LLMs with specialized chemistry knowledge sources to enhance their troubleshooting capabilities. We developed a custom chemistry troubleshooting knowledge base leveraging corporate troubleshooting documentation and expert-curated solutions, using the same corpus from which AI-generated questions were derived (Table 2).

Subsequently, our RAG workflow (Figure 2) follows a retrieval-then-generation pipeline: for each troubleshooting query, the system retrieves the top-k most relevant document chunks, summarizes them, and generates an answer based on the original query integrated with the retrieved context.

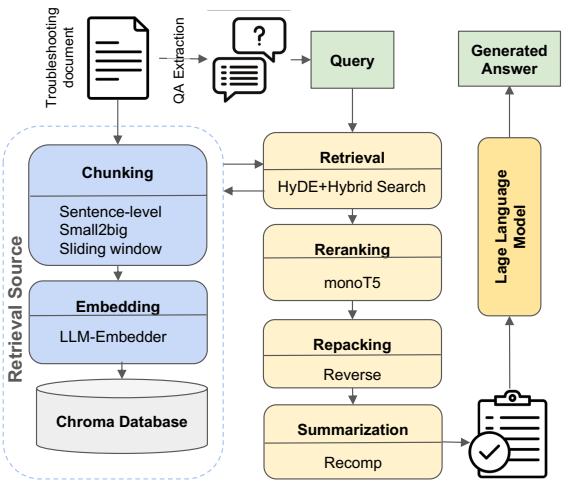


Figure 2: Retrieval-augmented generation workflow.

This implementation adheres to best practices outlined by Wang et al. (2024) while incorporating chemistry-specific prompt structures that maintain critical details about experimental conditions. Detailed implementation specifications are provided in Section 4.

3.4 Research Questions

To compare three enhancement approaches, we pursue three research questions in our benchmark evaluation.

RQ-1. *To what extent do domain-specialized LLMs outperform general-purpose LLMs in chemistry troubleshooting? How can existing LLMs be systematically ranked across key rubric criteria?*

This question examines whether chemistry-focused models demonstrate superior performance compared to general-purpose LLMs when addressing chemistry-specific troubleshooting scenarios. We utilize our rubric-based evaluation framework (Section 3.2) to assess various LLMs based on multiple dimensions relevant to chemistry troubleshooting tasks.

RQ-2. *How significantly does human-guided Chain-of-Thought reasoning enhance LLM performance compared to single-step answers?*

This question investigates whether structuring prompts to encourage step-by-step reasoning improves answers to troubleshooting questions. We evaluate how guided reasoning affects solution quality compared to direct responses, particularly for complex, multi-step chemical analysis.

RQ-3. *What measurable impact does RAG integration have on troubleshooting effectiveness?*

We examine how augmenting LLMs with exter-

Table 1: ChemBench evaluates general or professional LLMs on chemistry troubleshooting challenges.

Model	Type	#Size	Form	Ver.
GPT	General	N/A	api	5
Claude	General	N/A	api	Sonnet-4
DeepSeek	General	671B	api	R1
LLaMA	General	400B	api	4 Maverick
Gemini	General	N/A	api	2.5 Pro
Mistral	General	124B	api	Pixtral Large
ChemDFM	Professional	8B	open	v1.5
ChemLLM	Professional	7B	open	v1.5-DPO
ChemCrow	Professional	N/A	api	-

nal knowledge sources affects their performance in troubleshooting tasks, specifically for quantifying retrieval accuracy and generation performance, as well as the enhancement of overall solution quality when models can access relevant knowledgebase.

By addressing these research questions, our benchmark provides a structured evaluation of current capabilities of LLMs in chemistry troubleshooting, offering practical insights for both AI and chemistry researchers.

4 Experimental Setup

Model Selection. We selected diverse LLMs, spanning both general-purpose and domain-specialized categories. General-purpose models include GPT (OpenAI, 2025), Claude (Anthropic, 2024), Llama (Dubey et al., 2024), DeepSeek (DeepSeek-AI, 2025), Gemini (Comanici et al., 2025), Mistral (Jiang et al., 2023), while chemistry-specialized models include ChemDFM (Zhao et al., 2025), ChemLLM (Zhang et al., 2024), and ChemCrow (Bran et al., 2023). Detailed specifications for each model are provided in Table 1.

Implementation Details. For our RAG implementations, we deployed open-source LLMs on a local server integrated with LangChain. This approach was necessary because some RAG knowledge sources contained confidential troubleshooting documentation from instrument manufacturers that could not be uploaded to cloud-based API services. Following Wang et al. (2024), our RAG implementation used sentence-level chunking with a small-to-big strategy; 175-token chunks with a 20-token overlap for query matching, and 512-token chunks that include broader context. Embeddings were generated using the LLM-Embedder (Zhang et al., 2023) and stored in a Chroma vector database. Retrieval combined dense search with HyDE-generated pseudo-documents (Gao et al., 2023) and sparse BM25 scores, using a hybrid weight of $\alpha = 0.3$. Results were reranked with monoT5 (Nogueira et al., 2020), repacked in reverse order, and summarization was performed using the Recomp (Xu et al., 2023) extractive summarizer. For each query, the system retrieved the top-5 most relevant document chunks used for answer generation and retrieval performance evaluation.

Evaluation Metrics. For open-ended troubleshooting questions, we employed a rubric-based evaluation by chemists using six metrics (each rated on a 1-5 scale): **specificity/contextualization**, **correctness**, **usefulness**, **feasibility**, **misinformation risk**, and **error handling**, as detailed in Section 3.2.

For RAG evaluation, we separately evaluated retrieval and generation performance. Retrieval performance was assessed using two primary metrics: (1) **Recall@5** (Manning et al., 2008) measures whether the gold source chunk corresponding to the correct answer was retrieved within the top five results, and (2) **nDCG@5** (Järvelin and Kekäläinen, 2002) evaluates the quality of the retrieved ranking, evaluating the placement of relevant chunks relative to the gold source. Generation performance utilizes standard metrics (**ROUGE** (Lin, 2004), **BLEU** (Papineni et al., 2002)) and semantic similarity (**SBERT** (Reimers and Gurevych, 2019)) between model-generated answers and gold standard answers.

In addition to these metric-based evaluations, we implemented an external LLM-based evaluator using `GPT-4o-mini`[3] and `Qwen3 Max`[4], respectively. We developed a prompt template to guide LLMs in evaluating response quality rated on a 1-to-5 scale.

> You are an expert chemist. Your task is to evaluate the quality of an answer given a question and a reference (gold) answer.
> Question: `{question}`
> Reference Answer: `{gold-standard answer}`
> Generated Answer: `{answer}`
> Please rate the generated answer on a scale from 1 to 5, where: 1 = Very poor, 2 = Poor, 3 = Fair, 4 = Good, 5 = Excellent.

5 Results and Analysis

In this section, we present a comprehensive analysis of our benchmark results, examining how different LLMs perform across chemistry troubleshoot-

[3]https://openrouter.ai/openai/gpt-4o-mini
[4]https://openrouter.ai/qwen/qwen3-max

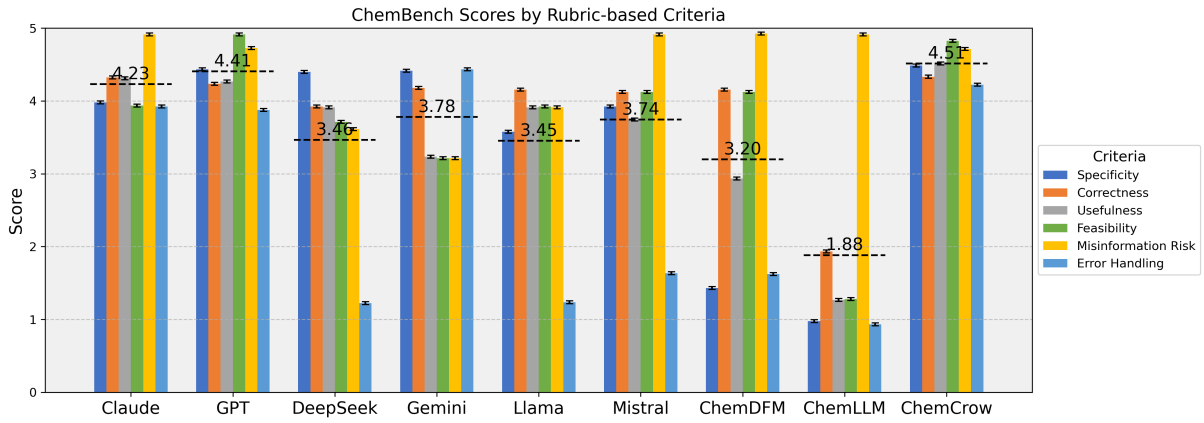


Figure 3: Performance of different models on ChemBench stratified by six criteria.

ing scenarios and comparing various enhancement approaches, according to our research questions.

5.1 Expert Evaluation Across Models

We analyze the results of expert evaluations across our criteria (Figure 3) to answer RQ-1 regarding domain-specialized models.

General vs. Chemistry-specialized. Counter-intuitively, chemistry-specialized models generally underperform compared to general commercial models. Figure 3 clearly demonstrates that GPT and Claude consistently outperform specialized models like ChemDFM and ChemLLM across most evaluation criteria. ChemLLM performed particularly poorly (average score 1.88), often generating incoherent responses. ChemDFM showed moderate performance but notable weaknesses in specificity (1.43) and usefulness (2.93).

ChemCrow represents the exception, achieving superior performance (average 4.50) comparable to top commercial models. However, ChemCrow differs fundamentally in its approach. It augments GPT-4 with chemistry tools rather than using fine-tuning alone. This suggests that tool augmentation may be more effective than domain-specific fine-tuning, particularly when fine-tuning is limited by smaller model sizes and parameter counts.

Rubric-based Analysis. Among all evaluation dimensions, error handling capability shows the most significant performance gap between model types. This metric demonstrates the largest contrast between high and low-performing models. Only three models (i.e., GPT (4.41), Claude (3.92), and ChemCrow (4.21)) demonstrate strong capability to acknowledge response limitations, request clarifying information, and suggest verification steps. The remaining models, including both general and specialized options, score below 1.7 in this dimension, indicating a fundamental inability to handle uncertainty appropriately.

In terms of correctness and feasibility, commercial general models (GPT and Claude) maintain a clear advantage, suggesting their superior reasoning capabilities extend effectively to chemistry domains despite lacking specialized training.

Overall Ranking. Analyzing overall performance across metrics reveals three distinct clusters:

- **Tier 1** (Average >4.4): GPT (4.41) and ChemCrow (4.50) demonstrate exceptional performance across all evaluation dimensions, offering both domain expertise and reasoning capabilities.
- **Tier 2** (Average 3.5-4.3): Claude (4.23), Gemini (3.78), and Mistral (3.74) provide strong general performance with moderate chemistry-specific capabilities, making them viable options for chemistry despite lacking specialized training.
- **Tier 3** (Average <3.5): DeepSeek (3.46), Llama (3.45), ChemDFM (3.20), and ChemLLM (1.88) show significant limitations in addressing chemistry troubleshooting tasks, with ChemLLM performing substantially worse than all other models.

These findings challenge conventional intuition about domain specialization, suggesting that the general capabilities of state-of-the-art commercial models may be more valuable than chemistry-specific training. Furthermore, tool augmentation approaches like ChemCrow appear more promising than fine-tuning for enhancing LLM performance in specialized scientific domains.

5.2 Chain-of-Thought Reasoning Evaluation

To address RQ-2, we evaluate CoT reasoning by implementing two additional interaction cycles with selective expert feedback incorporated into the

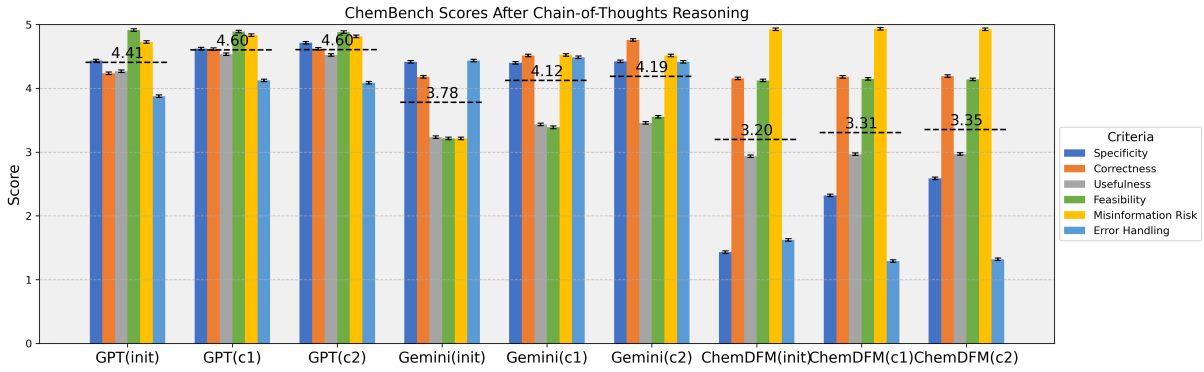


Figure 4: Performance of three representative models on ChemBench after CoT reasoning. C1 and C2 represent the first and second interaction cycle, respectively.

prompts. Experts rate responses using the same criteria after each cycle. We select representative models from each tier (GPT, Gemini, and ChemDFM) to track performance changes across multiple reasoning steps (Figure 4).

Performance Across Interaction Cycles. Our analysis shows CoT reasoning improves performance across most models with varying magnitudes. GPT shows consistent improvement, with its average score increasing from 4.41 (initial) to 4.60 (cycle 1) and maintaining at 4.60 (cycle 2). Gemini demonstrates even more substantial gains, with average scores rising from 3.78 to 4.12 and ultimately 4.19. In contrast, ChemDFM shows only marginal improvements, with average scores of 3.20, 3.31, and 3.35 across the three cycles.

Among evaluation dimensions, correctness and misinformation risk show the largest improvements across cycles, particularly for Gemini. Specificity remains relatively stable for GPT and Gemini, but improves sharply for ChemDFM, because expert feedback consistently highlighted it (e.g., "*...too general without any specific adjustments*" and "*...much more specificity is needed*"). Error handling improves significantly for general models but remains unchanged for ChemDFM, suggesting fundamental limitations in its adaptability.

Model-Specific Sensitivity to Reasoning. The three models show different sensitivities to CoT reasoning: GPT, already starting from a high baseline (4.41), shows modest improvement, primarily in correctness and feasibility. Gemini demonstrates the highest sensitivity to CoT (+0.41), with substantial gains across almost all dimensions, suggesting its architecture effectively incorporates feedback despite a lower starting point. ChemDFM shows minimal responsiveness to CoT (+0.15), with improvements concentrated in specificity while other dimensions remain largely unchanged.

These patterns suggest that CoT reasoning effectiveness relies on a model's fundamental reasoning architecture. The particularly strong improvement demonstrated by Gemini through CoT reasoning is noteworthy, as it suggests that well-structured interactive prompting can substantially narrow the performance gap between first and second-tier models in chemistry troubleshooting contexts.

5.3 RAG Performance Evaluation

To address RQ-3, we first examined the retrieval performance independently, then compared the generation performance between answers produced with and without RAG integration.

Retrieval Performance. We assessed retrieval quality using two standard ranking metrics, Recall@5 and nDCG@5, which respectively measure coverage of relevant information and ranking quality among the top-5 retrieved chunks. As shown in Figure 6, the best-practice RAG pipeline achieved a mean Recall@5 of 0.867 and a mean nDCG@5 of 0.789, indicating that most gold-standard sources were successfully retrieved within the top five results and that relevant chunks were generally ranked highly.

RAG vs. No-RAG. We compared answers generated with and without RAG across multiple metrics, as shown in Figure 5. The results clearly demonstrate the effectiveness of RAG, showing a consistent performance boost across all measures.

• **BERTScore-F1** (left) shows that RAG-generated answers exhibit a right-shifted distribution, peaking around 0.88, whereas answers w/o RAG cluster near 0.86. This indicates greater semantic similarity to the gold answers.

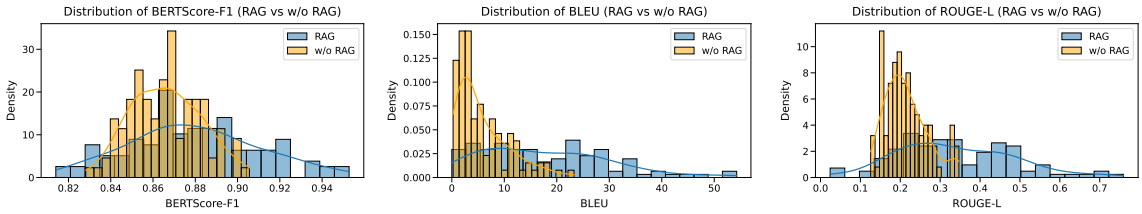


Figure 5: Comparison of LLM-generated answers with and without retrieval augmentation across three evaluation metrics. RAG consistently yields higher BERTScore-F1 (left), BLEU (center), and ROUGE-L (right) scores, reflecting improved semantic similarity, lexical precision, and contextual coherence relative to the reference answers.

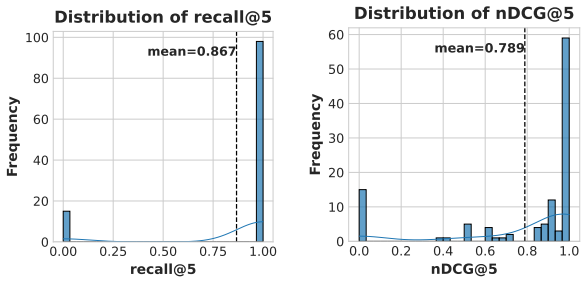


Figure 6: Retrieval metrics for the best-practice RAG pipeline (recall@5 and nDCG@5).

• **BLEU** (center) highlights improvements in lexical overlap. RAG-generated answers extend to higher values (often exceeding 20), while those w/o RAG remain concentrated below 10.

• **ROUGE-L** (right) follows the same trend, measuring the overlap of longer sequences. RAG scores are more broadly distributed and shift toward higher values (0.3–0.6), while No-RAG scores remain largely below 0.25. This improved overlap suggests that RAG helps preserve contextual and structural consistency.

These results demonstrate that RAG integration significantly enhances answer quality for chemistry troubleshooting queries, and improves the faithfulness and informativeness by incorporating specialized knowledge accurately.

Discussion on LLM-as-a-Judge. Since expert human evaluations are resource-intensive and time-consuming to scale, we additionally explored using LLMs themselves as evaluators of response quality. We prompted an external LLM (GPT, followed by Qwen) to rate each answer on a 1–5 scale based on factual accuracy, completeness, and alignment with the gold answer.

As shown in Figure 7, the evaluation results from both judges align well with our previous findings:

• **GPT's evaluation** (left) shows that RAG-generated answers consistently receive higher ratings, with the majority of scores clustered around 4–5, while No-RAG (LLM-only) answers are concentrated around 3–4.

• **Qwen's evaluation** (right) demonstrates an even more pronounced effect. RAG-generated answers are heavily concentrated in the 4–5 range, while No-RAG answers predominantly receive scores of 2 or 3. This consistent pattern across both evaluators indicates that even automated LLM judges recognize the added factual grounding and contextual completeness introduced by retrieval augmentation. The results further confirm the effectiveness of RAG in elevating answer quality. These findings suggest that LLM-based evaluation can serve as a scalable proxy for expert judgment when assessing domain-specific reasoning quality, though further work is needed to ensure robustness and mitigate evaluator bias.

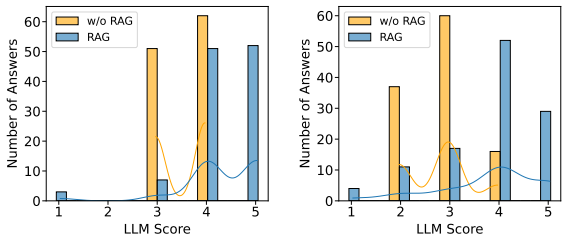


Figure 7: Performance comparison of AI-generated answers with and without RAG, scored by external LLM evaluators GPT (left) and Qwen (right).

6 Conclusion

ChemBench offers valuable insights into LLM performance for chemistry troubleshooting tasks. Surprisingly, general models outperformed domain-specialized models (except ChemCrow), suggesting that robust reasoning capabilities may be more valuable than domain-specific training alone. CoT reasoning demonstrated significant performance improvements, particularly for mid-tier models. RAG consistently enhanced answer quality by incorporating specialized knowledge.

Limitations

Despite our efforts to create a comprehensive benchmark for evaluating LLMs in chemistry troubleshooting, several limitations warrant consideration. On the one hand, our benchmark focuses primarily on analytical chemistry troubleshooting scenarios (particularly HPLC and LC-MS), which may not fully represent the breadth of chemistry subdomains such as organic synthesis, inorganic chemistry, or biochemistry. On the other hand, while our expert evaluation framework provides valuable insights, we only partnered with 30 specialists, considerably fewer than OpenAI's HealthBench which collected assessments from more than 200 physicians globally. Gathering sufficient expert evaluations is both time-consuming and resource-intensive; we plan to continue expanding our pool of evaluators to make our statistical results more convincing and comprehensive.

Ethics Statement

Our study maintains strict ethical standards concerning data collection and expert involvement.

Expert Annotation and Consent. Our annotators are comprised of experienced chemists recruited from diverse professional backgrounds, including both leading scientific research consortia and major instrument manufacturers. All participating experts provided informed consent for their annotations to be used in this research and for their contributions to be released alongside the final dataset.

Data Privacy. The content annotated consists entirely of simulated scientific dialogues generated from publicly available troubleshooting documentation (Table 2). No personal, sensitive, or confidential information is involved in the annotation process.

Open Science. All research procedures follow institutional guidelines, and no additional ethics board review is required under the ACL Ethics Policy. The dataset and annotation guidelines will be made publicly available upon publication to support reproducibility and future research in proactive scientific assistance systems.

References

Anthropic. 2024. The Claude 3 model family: Opus, sonnet, haiku. The most recent major release from Anthropic, representing the current state-of-the-art model family from the company.

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quinonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

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

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. Chemcrow: Augmenting large-language models with chemistry tools.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, and 4 others. 2024. SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis. *arXiv preprint*. ArXiv:2403.01976 [cs].

Cayque Monteiro Castro Nascimento and André Silva Pimentel. 2023. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655.

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

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. 2024. SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models. *arXiv preprint*. ArXiv:2406.09098 [cs].

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and 1 others. 2023. What can large language models do in chemistry? a comprehensive benchmark on

eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *arXiv preprint*. ArXiv:2009.03300 [cs].

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

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–441.

Albert Q. Jiang, Alexandre Han, Romain Mohta, Pascal Bëlisle, Eric Étienne, Flora Reizenstein, Shikhar Sharma, Wenhan Feng, Guillaume Lample de Lépinau, Stéphane Mazaré, Alexis Choplin, Maël de Brébisson, Teo Fan, and Hugo Lefevre. 2023. Mistral 7B a sub-20b parameter language model. *arXiv preprint arXiv:2310.06825*.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024. SceMQA: A Scientific College Entrance Level Multimodal Question Answering Benchmark. *arXiv preprint*. ArXiv:2402.05138 [cs].

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to information retrieval. Standard definition of Recall@k in Information Retrieval.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*.

OpenAI. 2025. GPT-5 system card. Official technical and safety documentation.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint*. ArXiv:2206.04615 [cs].

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. SciEval: A Multi-Level Large Language Model Evaluation Benchmark for Scientific Research. *arXiv preprint*. ArXiv:2308.13149 [cs].

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, and 1 others. 2024. Searching for best practices in retrieval-augmented generation. *arXiv preprint arXiv:2407.01219*.

Andrew D White, Glen M Hocky, Heta A Gandhi, Mehrad Ansari, Sam Cox, Geemi P Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, and 1 others. 2023. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2(2):368–376.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, Shufei Zhang, Mao Su, Hansen Zhong, Yuqiang Li, and Wanli Ouyang. 2024. Chemllm: A chemical large language model. *Preprint*, arXiv:2402.06852.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.

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

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. *arXiv preprint*. ArXiv:2304.06364 [cs].

A Related Work

A.1 Large Language Models in Chemistry

The integration of Large Language Models (LLMs) with chemistry has emerged as a promising interdisciplinary research direction, with efforts broadly falling into two approaches.

The first approach integrates LLMs with specialized chemistry tools. Bran et al. (2023) developed ChemCrow, an LLM-based system augmented with chemistry-specific tools for synthesis planning and drug discovery tasks. Similarly, Boiko et al. (2023) created an autonomous chemical agent that utilizes multiple LLMs to conduct laboratory experiments. These agent-based systems capitalize on LLMs' reasoning abilities while addressing domain-specific limitations.

The second approach involves domain-specific fine-tuning of LLMs for chemistry applications. Jablonka et al. (2024) demonstrated promising performance of GPT models in property prediction and synthesis planning with minimal examples. Specialized models have emerged, including ChemDFM (Zhao et al., 2025), trained on chemistry literature to understand chemical concepts, and ChemLLM (Zhang et al., 2024), fine-tuned for understanding chemical structures and reactions.

Despite growing interest in chemistry-focused LLMs, systematic evaluation frameworks remain limited. While initial assessments of LLM capabilities in chemistry contexts exist (White et al., 2023; Castro Nascimento and Pimentel, 2023), and ChemLLMBench (Guo et al., 2023) evaluates performance across eight well-defined tasks with clear ground truth answers, the field lacks benchmarks for open-ended chemistry troubleshooting scenarios. These complex scenarios better represent real-world research challenges but require expert evaluation due to their ill-defined nature and multiple valid solution paths. This significant gap constrains our understanding of how effectively LLMs can assist with complex chemistry problems.

A.2 Evaluation of Large Language Models

In recent years, the evaluation of LLMs like GPT has become a significant field of inquiry. In the early days of transformer evaluation, n-gram–based metrics such as ROUGE, BLEU, and BERTScore were used to assess grammatical and semantic fidelity. However, as field-specific LLMs emerged, researchers recognized the need for benchmarks that probe deeper reasoning and domain knowledge. Generalist suites like MMLU (Hendrycks et al., 2021), BIG-Bench (Srivastava et al., 2023), and AGIEval (Zhong et al., 2023) include science questions but rely on short, single-turn QA and fail to assess the multi-step problem solving required in real laboratory settings.

Recent multi-level benchmarks such as SciEval (Sun et al., 2024) and SciKnowEval (Feng et al., 2024) assess LLMs across a spectrum from factual recall to experimental design and safety reasoning in scientific domains. SciEval uses Bloom's taxonomy to probe basic knowledge and research-level tasks in chemistry, physics, and biology, revealing that SOTA models (e.g. GPT-4) struggle on dynamic problems and experimental analysis. SciKnowEval extends this by adding ethical and practicality judgements, testing whether models can avoid unsafe recommendations or propose feasible laboratory procedures. Although quite comprehensive, these models do not test interactive troubleshooting.

Benchmarks focused on literature comprehension and multimodal reasoning evaluate LLMs ability to interpret scientific research papers, both text and figures. For example, SciAssess (Cai et al., 2024) presents questions targeted towards comprehension and analysis of real scientific article excerpts, while SceMQA (Liang et al., 2024) gives answers to questions pertaining to college entrance exam problems within the core STEM fields.

Ultimately, several benchmarks exist to test basic knowledge and operation of chemical experimentation, but fail to fully capture step-wise guidance necessary for real-world experimentation. Domain-specific benchmarks, like ChemLLMBench (Guo et al., 2023) directly benchmarks core chemistry tasks related to chemical experimental design and core chemistry knowledge, including nomenclature, yield prediction, and reagent selection. ChemLLMBench uses objective metrics alingside human chemist evaluations, performing competitively with GPT-4. Once again, however, this benchmark reveals that automated metrics do not capture the nuances of feasibility curated under expert review.

B Dataset Construction

B.1 Human-generated Questions

We collected approximately 50 open-ended chemistry troubleshooting questions through discussions with expert chemists. Among, half of these questions include expert-provided gold-standard an-

swers (Table 3), while the rest are unanswered (Table 4). The whole collection of our question set will be submitted as the supplementary materials.

These questions focus on analytical chemistry challenges, particularly chromatography and mass spectrometry issues that chemists routinely encounter in laboratory settings. The collection includes questions about peak resolution problems, quantitation difficulties, method development challenges, system troubleshooting, and interpretation of unexpected results. For example, a typical question addresses separation challenges in high-performance liquid chromatography (HPLC): "*I have 4 components in my sample, but no matter what HPLC conditions I try, I only see a maximum of 3 peaks. I am using a C18 column with a water/Acetonitrile gradient from 70-90%. What should I do?*" Other questions address issues such as ion suppression in LC-MS, reproducibility problems, abnormal peak shapes (tailing, fronting, splitting), baseline disturbances, sensitivity challenges, and unexpected mass spectra interpretations. These scenarios require consideration of multiple possible causes, systematic diagnostic approaches, and development of practical solutions that balance effectiveness and feasibility.

B.2 AI-generated Questions

Additionally, we extracted 113 QA pairs from the troubleshooting documents as well as the source chunks for each answer. Then, we use them to evaluate RAG retrieval and generation performance. Three pairs of them are listed in Table 5. The whole collection of our question set will be submitted as the supplementary materials.

Source Corpus and Scope. To build a domain-specific corpus, we collected technical PDF documents spanning chromatography and mass spectrometry instrumentation, including troubleshooting manuals, user guides, training workbooks, and site-preparation specifications.

These documents, sourced from instrument vendors and application notes, cover common laboratory issues such as mobile phase preparation, column care, vacuum leaks, and detector noise. The detailed information of these documents are listed in Table 2. The corpus is representative of real-world technical documentation, containing a mix of free text, bulleted checklists, parameter tables, warning boxes, and instrument photographs.

LLM-Assisted Question-Answer Generation. We employed a multi-stage process to generate high-quality question-answer (QA) pairs. First, the source PDFs were segmented using a `RecursiveCharacterTextSplitter` with a chunk size of 2,000 tokens and an overlap of 200 tokens. We chose a large window (2,000 tokens with 200 overlap) to preserve long procedural sections and avoid fragmenting multi-step instructions during generation. For each resulting text chunk, we prompted an Azure OpenAI chat completion model (`gpt-4o-mini`) to produce a single QA pair. The prompt was specifically designed to elicit natural, first-person laboratory scenarios that are open-ended in nature. we used the following prompt:

> You are a helpful and knowledgeable lab assistant trained in HPLC and LC/MS troubleshooting. Based on the following technical content, generate ONE natural-sounding lab question in the style of a scientist seeking help, using a realistic first-person scenario. avoid being too general. (e.g., "I'm setting up my autoflex series instrument for a new experiment, but I'm concerned about the environmental conditions in my lab. The temperature fluctuates quite a bit, and I'm worried it might affect my results. What should I do?" or " I'm setting up my autoflex series instrument for a new experiment and I'm seeing some fluctuation in the data I'm collecting. What should I do?" Then provide a detailed, step-by-step, and comprehensive answer using ONLY the provided content. Be thorough in your explanation, include any relevant background, rationale, troubleshooting options, and possible causes or implications.

This approach encourages the generation of comprehensive, procedural answers rather than simple fact retrieval. To build the dataset, we randomly sampled 200 chunks from the segmented manuals and prompted the model once per chunk, resulting in 200 synthetic QA pairs that mimic the queries of a laboratory technician.

Multi-Stage Filtering Protocol. To ensure the dataset's quality and practical relevance, we implemented a rigorous two-stage filtering protocol involving both automated scoring and expert human review.

Stage 1: Automated LLM-Based Filtering. Each of the 200 generated QA pairs was automatically scored by an LLM assessor along three distinct criteria, each on a five-point scale: (1)

Table 2: Source troubleshooting documents

Document Title	Document Type	Page #	Source Company	Target Software
The Chromatography Detective: Troubleshooting Tips & Tools for LCMS	Troubleshooting Manual	67	Agilent	General LC/LCMS systems
Agilent Triple Quadrupole LC/MS System User Guide	Official User Manual	145	Agilent	MassHunter 12.1 or higher
Agilent Triple Quadrupole LC/MS System Introduction Workbook	Training Workbook	124	Agilent	MassHunter 12.1 or higher
compact	Site Preparation Specification	9	Bruker	N/A (Hardware focus)
solariX series	Site Preparation Specification	27	Bruker	N/A (Hardware focus)
timsTOF	Site Preparation Specification	12	Bruker	N/A (Hardware focus)
scimaX series	Site Preparation Specification	27	Bruker	N/A (Hardware focus)
autoflex series	Site Preparation Specification	10	Bruker	N/A (Hardware focus)
impact series	Site Preparation Specification	11	Bruker	N/A (Hardware focus)
ultrafleXtreme	Site Preparation Specification	9	Bruker	N/A (Hardware focus)
maXis series	Site Preparation Specification	10	Bruker	N/A (Hardware focus)
neofleX series	Site Preparation Specification	15	Bruker	N/A (Hardware focus)

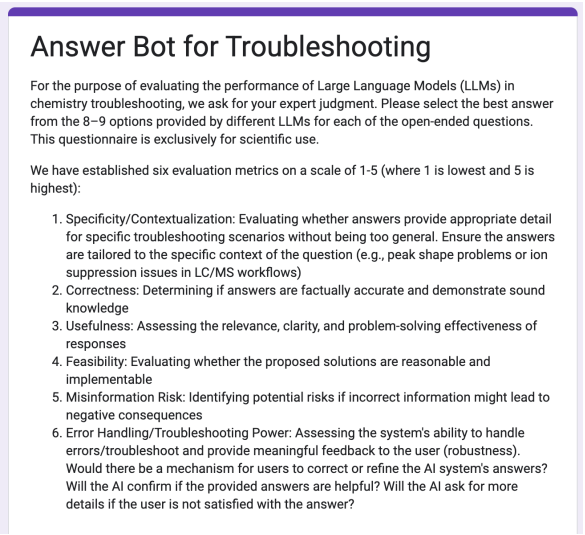
Groundedness: Whether the answer is fully supported by the provided source context. (2) **Relevance:** The practical usefulness of the question to a chromatography user. (3) **Standalone Clarity:** Whether the question is understandable without needing the source context. We automatically filtered out any pair that did not achieve a score of 4 or higher on all three criteria, which reduced the candidate pool to 135 QA pairs.

Stage 2: Expert Chemists Review. The remaining 135 candidates were independently reviewed by 5 chemists. Each expert rated the pairs on a four-point scale across four criteria: (i) question clarity and specificity, (ii) factual accuracy of the answer, (iii) whether the answer directly addresses the question, and (iv) the sufficiency of the source chunk for generating a complete answer. Any item receiving a score below 3 on any criterion from either reviewer was discarded. This final, stringent review process yielded the final dataset of **113** high-quality QA pairs. For each pair, the source document filename and page number are stored to ensure traceability.

C Questionnaire for Chemistry Experts

To standardize the evaluation and ensure expert consistency, we provided chemists with a detailed questionnaire that clearly outlined the study's scientific purpose. This instrument required experts to use a 1-to-5 scale to assess LLM-generated answers against six criteria. The complete structure of the questionnaire, including the six metrics and the instruction for selecting the best answer, is shown in Figure 8.

To ensure the dataset's quality and practical relevance, a separate questionnaire was created for the filtering of the AI-generated QA pairs. This



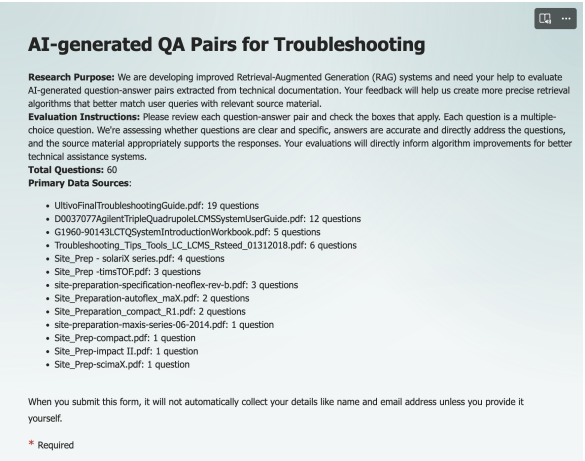
Answer Bot for Troubleshooting

For the purpose of evaluating the performance of Large Language Models (LLMs) in chemistry troubleshooting, we ask for your expert judgment. Please select the best answer from the 8–9 options provided by different LLMs for each of the open-ended questions. This questionnaire is exclusively for scientific use.

We have established six evaluation metrics on a scale of 1-5 (where 1 is lowest and 5 is highest):

1. Specificity/Contextualization: Evaluating whether answers provide appropriate detail for specific troubleshooting scenarios without being too general. Ensure the answers are tailored to the specific context of the question (e.g., peak shape problems or ion suppression issues in LC/MS workflows)
2. Correctness: Determining if answers are factually accurate and demonstrate sound knowledge
3. Usefulness: Assessing the relevance, clarity, and problem-solving effectiveness of responses
4. Feasibility: Evaluating whether the proposed solutions are reasonable and implementable
5. Misinformation Risk: Identifying potential risks if incorrect information might lead to negative consequences
6. Error Handling/Troubleshooting Power: Assessing the system's ability to handle errors/troubleshoot and provide meaningful feedback to the user (robustness). Would there be a mechanism for users to correct or refine the AI system's answers? Will the AI confirm if the provided answers are helpful? Will the AI ask for more details if the user is not satisfied with the answer?

Figure 8: The description of questionnaire about rubric-based evaluation.



AI-generated QA Pairs for Troubleshooting

Research Purpose: We are developing improved Retrieval-Augmented Generation (RAG) systems and need your help to evaluate AI-generated question-answer pairs extracted from technical documentation. Your feedback will help us create more precise retrieval algorithms that better match user queries with relevant source material.
Evaluation Instructions: Please review each question-answer pair and check the boxes that apply. Each question is a multiple-choice question. We're assessing whether questions are clear and specific, answers are accurate and directly address the questions, and the source material appropriately supports the responses. Your evaluations will directly inform algorithm improvements for better technical assistance systems.
Total Questions: 60
Primary Data Sources:

- UltivoFinalTroubleshootingGuide.pdf: 19 questions
- D0037077AgilentTripleQuadrupoleLCMSSystemUserGuide.pdf: 12 questions
- G1960-90143LCTQSystemIntroductionWorkbook.pdf: 5 questions
- Troubleshooting_Tips_Tools_LC_LCMS_Rsteed_01312018.pdf: 6 questions
- Site_Prep - solariX series.pdf: 4 questions
- Site_Prep -timsTOF.pdf: 3 questions
- site-preparation-specification-neoflex-rev-b.pdf: 3 questions
- Site_Preparation-autoflex_maX.pdf: 2 questions
- Site_Preparation_compact_R1.pdf: 2 questions
- site-preparation-maxis-series-06-2014.pdf: 1 question
- Site_Prep-compact.pdf: 1 question
- Site_Prep-impact II.pdf: 1 question
- Site_Prep-scimaX.pdf: 1 question

When you submit this form, it will not automatically collect your details like name and email address unless you provide it yourself.

* Required

Figure 9: The description of questionnaire about AI-generated QA filtering.

instrument, shown in Figure 9, asks experts to validate if the questions are clear, the answers are accurate, and the source material appropriately supports the responses. Because reviewing the entire pool of 135 generated questions was deemed too time-consuming for the chemists, we segmented the dataset, presenting 60 questions per expert review batch. This approach ensured a focused and efficient final screening process.

Table 3: Selected chemistry troubleshooting questions with expert answers from the ChemBench dataset.

Question	Answer
The signal for my analyte is low in the mass spectrometer.	Try positive and negative mode. Try different mobile phases. Try microflow or nanoflow HPLC. Derivatize your compound to make it ionize better.
The signal for my analyte is low in the UV-VIS detector.	Try a detector cell with a longer pathlength. Try mobile phases or columns that sharpen your peaks. Derivatize your compound to give it a chromophore that absorbs in the UV-Vis spectra.
What could cause loss of sensitivity in LC/MS?	Possible causes include: blocked or damaged nebulizer, contaminated transfer capillary, contaminated source, contaminated solvents, or detector gain loss.
What can cause changing retention times in LC/MS?	Causes include: insufficient buffer capacity, contamination buildup, insufficient equilibration time, active sites on the column, inconsistent on-line mobile-phase mixing, selective evaporation of mobile-phase components, varying column temperature, and check-valve malfunction.
What can cause high back pressure in LC/MS?	Causes include column blockage, small particle size, microbial growth, high mobile-phase viscosity, plugged frits, and salt precipitation.
What should I do if there is high column backpressure?	Check flow path (needle seat, capillaries, filter, and frits). Use lower viscosity solvents or higher temperature.
How can you address decreasing pressure?	Vent mobile-phase reservoir, replace inlet-line frit, check for pinched tubing, replace or clean check valves, replace pump seals, and degas solvent.
What can cause broad peaks in LC/MS?	Causes include sample overload, large injection volume, large extra column volume, high mobile-phase viscosity, peak dispersion in injector valves and poor column efficiency.
How can you address ghost peaks?	Flush the column to remove contaminants, clean the MS interface, use fresh solvents, and use sample cleanup methods.
What type of bottles should be used to avoid algae growth in solvents?	Use brown borosilicate bottles.
How often should aqueous and organic mobile phases be replaced?	Every second day.
What should be done weekly to maintain the LC system?	Change seal wash solvent, inspect solvent filters, and check system backpressure.
Why do peaks become distorted and plate counts drop after a few hundred injections? What are possible causes of short column lifetime?	The column could be contaminated. To address this, use a guard column, clean the column with a suitable solvent, or replace it. Other causes of column short lifetime could be pH outside the recommended range, use of strongly acidic or alkaline solutions, exposure to high temperatures or pressures beyond specifications
How can contamination in assays be prevented?	Use high-purity solvents, clean glassware thoroughly, filter samples, and use guard columns to prevent contamination.
What causes baseline noise in chromatograms?	Possible causes include air bubbles, pump issues, detector problems, or impurities in the mobile phase.
What is the best way to store columns when not in use?	Store columns in a cool place with end caps on and filled with an appropriate solvent to maintain their condition.
How can column life-time be extended?	Regularly clean the column, use guard columns, filter samples and solvents.
What is a common problem with shifting retention times?	Shifting retention times can occur due to changes in mobile phase composition or column temperature fluctuations.

Table 4: Selected unanswered open-ended chemistry troubleshooting questions from the ChemBench dataset.

Question	Answer
I have 4 components in my sample, but no matter what HPLC conditions I try, I only see a maximum of 3 peaks. I am using a C18 column with a water/Acetonitrile gradient from 70-90% What should I do?	N/A
I am having trouble getting accurate quantitation of the first peak from my LC-MS method, which is eluting very early in the assay and showing significant ion suppression. What should I do? I am developing an HPLC method for measuring anthracene, but having difficulties with reproducibility and seeing broad rolling baseline problems. I am using a C18 column with a water/Acetonitrile gradient from 70-90%. What should I do?	N/A
I am developing an HPLC-MS assay for several amide derivatives of the amino acid proline. I am seeing some strange behavior including split peaks and other abnormal peak shapes. What should I do?	N/A
I am developing an HPLC-MS assay for a drug product. I am getting good signal and good resolution from impurities, but the major peak has a tailing shoulder that makes integration and quantitation difficult. What should I do?	N/A
I am developing an HPLC-MS assay for a natural product. I am getting good signal and good resolution from impurities, but the major peak has a fronting that makes integration and quantitation difficult. What should I do?	N/A
I am developing an HPLC-MS assay and seeing some strange behavior. I am observing the doubling of some peaks and some strange peak shapes. In addition, I am sometimes observing an increased backpressure upon sample injection. What should I do?	N/A
I am having difficulties with an HPLC-MS assay for a natural product. My assay was working fine yesterday, but today the intensity of the main peak for the compound of interest has decreased and I am seeing a new peak growing in. What should I do?	N/A
My HPLC-MS system is working fine, but whenever I inject a sample of our new investigational compound I get weird results. I am seeing very strong tailing and extra peaks for this sample, although other samples and standards behave normally. What should I do?	N/A
I am repeating some HPLC-MS assays from last year, using the same column, which has been in storage for several months. My results are similar to what I previously observed, but the peak shapes are terrible, with many peaks broadened and some doubled. What should I do?	N/A
I am studying some newly synthesized BOC-amino acids by LC-MS. The HPLC traces look encouraging, but I am having difficulty seeing the M+1 peak for these new compounds. What should I do?	N/A
My LC-MS assay is showing a peak in the MS at m/z = 441 for almost every peak I examine. I am not sure what is going on. This peak wasn't seen in these samples last week. What should I do?	N/A
I am about to start the development of a new LC/MS method. After validation, the method will be transferred to the QC Lab, where it will be used for many years. What can I do to ensure good batch to batch reproducibility?	N/A
I am encountering some problems with my sample preparation. I am using SPE with a reversed-phase sorbent. Usually, the recovery is around 80% but sometimes it drops down to 50% or less. What could be the problem?	N/A

Table 5: Generated chemistry troubleshooting QA pairs using AI, with RAG responses derived from domain-specific documentation

Question	Standard Answer	RAG Answer
I am trying to generate calibration curves in MassHunter Quantitative Analysis but I'm unsure how to properly associate my internal standard with each target compound. What should I do?	To associate your internal standard (ISTD) with each target compound in MassHunter Quantitative Analysis, click the down arrow in the ISTD Compound Name cell for each target compound row and select the appropriate ISTD name. After that, enter the ISTD concentration for each ISTD compound in the ISTD Conc. cell, using the correct concentration value (e.g., 50.0000). This will ensure that your calibration curves are accurately generated based on the selected ISTD.	To properly associate your internal standard with each target compound in MassHunter Quantitative Analysis, follow these steps: 1. **Open Method Setup Tasks**: In MassHunter Quantitative Analysis interface, locate the sidebar on the left side of the method table window. 2. **Access ISTD Setup**: Click on 'Method Setup Tasks', then select 'ISTD Setup'. 3. **Assign Internal Standards**: - For each target compound row, click the down arrow in the 'ISTD Compound Name' cell. - Select the corresponding deuterated compound that you want to assign as the internal standard for that target compound. 4. **Enter ISTD Concentration**: In the 'ISTD Concentration' cell for each internal standard, type the concentration value (e.g., '50.0000').
I recently replaced the AC board in my Ultivo system, but the quad heater isn't heating up properly. What should I do?	Start by power cycling the Ultivo to see if a Qcard reboot allows the quad to start heating. Next, check the light status of the heaters at the back of the instrument near the Foreline Pump power cable; the lights should be on when heating and cycle on and off while holding temperature. If the quad is cold and no light is on, this indicates that the AC board is not sending power to the Quad heater. Additionally, verify all 24V readings using the PVH; if any 24V is missing, remove the Ultimod PS and test the voltages from an available power source, focusing on pins 3 & 4. If the voltage is still absent, you may need to replace the Ultimod PS. Lastly, consider swapping the heater/sensor cable connected to the cyclone and vortex boards; if the heater begins to heat after the swap, the issue may lie with the AC board or the heater cable itself.	To troubleshoot the quad heater not heating up properly after replacing the AC board in your Ultivo system, follow these steps: 1. **Power Cycle the Instrument**: Turn off the Ultivo system and then turn it back on. This may reset the system and allow the quad to start heating. 2. **Check Heater Light Status**: Look at the back of the instrument near the foreline pump power cable. Verify the light status of the heaters. If the quad is cold and no light is on, the AC board may not be sending power to the quad heater. 3. **Verify 24V Readings**: Access the pvh and check that all 24V readings are within the correct range. If 24V is missing: - Remove the ultimod power supply (PS) and plug it into any available power source. - Using the upper left pin as ground, measure all other voltages, focusing on pins 3 & 4. If the voltage is still missing, replace the ultimod PS. 4. **Inspect Heater/Sensor Cable**: Swap heater/sensor cable connected to the cyclone and vortex boards. If the heater begins to heat after the swap, the issue may be with the AC board or the heater cable. 5. **Check Connections and Cables**: Inspect the wires from the bulkhead connector to AC board for any damage or poor connections. If no issues are found, ensure that ESW version matches and reinstall it if necessary. By following these steps, you should be able to identify and resolve the issue with the quad heater not heating properly.

(continued on next page)

Question	Standard Answer	RAG Answer
I recently switched to using trifluoroacetic acid (TFA) in my ESI positive mode HPLC method, but now I'm seeing unexpected acetate anion contamination in my ESI negative mode results. What should I do?	To address the acetate anion contamination you're observing in ESI negative mode after using TFA, you should take the following steps: First, change the solvent frits and the analytical column, as these components may retain the TFA and lead to carryover. Next, rinse the LC system for an extended period to ensure that any residual TFA is removed. Additionally, consider using acetic acid or formic acid as alternatives to TFA in your mobile phase to prevent this issue in the future. This approach will help minimize contamination and improve the reliability of your results.	To address the unexpected acetate anion contamination in your ESI negative mode results after switching to trifluoroacetic acid (TFA) in your HPLC method, follow these steps: 1. **Check for TFA Adducts**: Verify if the expected TFA adducts (e.g., m/z 1034, 1334, etc.) are present in your negative ion mode results. This can help confirm if the contamination is related to TFA. 2. **Inspect the Analytical Column**: If you do not see contamination in the MS2 scan mode, the contamination might be located in the analytical column. Consider rinsing or replacing the column if necessary. 3. **Rinse the System**: Rinse the LC system thoroughly. Use acetic or formic acid as an alternative to TFA for rinsing. Ensure that you flush the system with a suitable solvent for an extended period to remove any residual contaminants. 4. **Clean the CDS Bottle**: Empty the CDS bottle, acid etch it, and rinse it with an organic solvent. Then fill it with acetonitrile and flush the CDS for at least an hour. Some contamination may require longer flushing. 5. **Replace Filters and Tubing**: Check and replace any filters in the system. Use PTFE tubing for the Milli-Q water system to minimize contamination. 6. **Use New Calibrant Tune Mix**: If contamination persists, consider using a new calibrant tune mix solution from a different lot to avoid carryover from previous analyses. 7. **Monitor for Abnormal Background Ions**: Be aware of any abnormal background ions that may appear in negative ion mode, particularly if using a nitrogen generator. If you observe such ions, further investigate the nitrogen generator's condition. By following these steps, you should be able to mitigate the unexpected acetate anion contamination in your ESI negative mode results.