TroubleRAG: Evaluating Retrieval Pipelines for Real-World Chemistry Troubleshooting

Anonymous Author(s)
Affiliation
Address
email

Abstract

Troubleshooting complex laboratory instruments, such as in chromatography and mass spectrometry, presents a significant information retrieval challenge due to highly specific, technical documentation. Existing chemistry RAG benchmarks primarily target short, general-purpose Q/A tasks, whereas chemists need tools that can address open-ended, laboratory troubleshooting questions in the context of scientific research. To bridge this gap, we introduce TroubleRAG, a comprehensive benchmark for evaluating Retrieval-Augmented Generation (RAG) pipelines in this domain. We first constructed a novel dataset of 113 high-quality troubleshooting scenarios, curated from synthetic data using LLM-based scoring and expert chemists validation. Using TroubleRAG, we conduct an empirical analysis of key retrieval design choices, including sparse, dense, and hybrid fusion; HyDE query expansion; and advanced chunking strategies. Our key finding is that widely recommended "best-practice" RAG configurations do not transfer: they underperform on specialized troubleshooting tasks. Guided by empirical analysis, we introduce a domain-tailored retrieval recipe that yields significant improvements, boosting Recall@5 by 8% and nDCG@5 by 8%. We also outline two extensions: (i) multimodal retrieval over tables and figures that routinely appear in instrument manuals, and (ii) multi-turn, interactive systems that request clarifying details to better reflect human-in-the-loop workflows. TroubleRAG is designed to advance robust, domain-aware RAG methodologies for practical laboratory support.

1 Introduction

Troubleshooting complex instruments like Liquid Chromatography-Mass Spectrometry (LC-MS) systems presents a critical bottleneck in scientific research, requiring experts to navigate through vast, unstructured technical manuals to diagnose and resolve instrument failures [15]. While Large Language Models (LLMs) offer potential assistance, their tendency to hallucinate makes them unreliable for high-stakes laboratory environments where incorrect guidance can lead to costly delays or equipment damage [6]. Retrieval-Augmented Generation (RAG) addresses this limitation by grounding LLMs in relevant documentation [10].

However, prevailing RAG best practices have been tuned for general-domain QA (e.g., Natural Questions [9]) or biomedical datasets with short, factual answers (e.g., PubMedQA [7]). These approaches prove inadequate for scientific troubleshooting scenarios, which demand the retrieval of lengthy procedural texts, synthesis of multi-step solutions, and interpretation of multimodal content including instrument diagrams, parameter tables, and error screenshots. It remains an open question:

Do general-purpose RAG "best practices" effectively transfer to specialized technical domains?

Submitted to the AI for Science workshop (NeurIPS 2025).

To bridge this gap, we introduce `TroubleRAG`, a comprehensive benchmark specifically designed for evaluating RAG pipelines on real-world chemistry troubleshooting tasks. Our work makes three key contributions:

• *Novel Dataset Construction*: Through discussions with experts at the NSF Center for Bioanalytic Metrology (CBM), we present a carefully curated dataset of 113 expert-validated troubleshooting scenarios derived from authentic LC-MS and HPLC technical documentation, designed to reflect the complexity and open-ended nature of real laboratory problems.

• *Systematic Empirical Analysis*: Through empirical evaluation across eight RAG configurations, we demonstrate that widely-recommended "best practice" approaches significantly underperform on specialized troubleshooting tasks, challenging the assumption that general-domain optimizations transfer to technical domains.

• *Domain-Tailored Methodology*: We develop a domain-specific retrieval recipe that combines SPLADE sparse retrieval, dense embeddings, and advanced reranking (FlagReranker), achieving substantial improvements of 8% in Recall@5 and 8% in nDCG@5s. Moreover, we outline critical extensions including multimodal retrieval capabilities for processing instrument figures and tables, and interactive multi-turn systems that mirror real-world human-in-the-loop troubleshooting workflows.

Our findings reveal that effective RAG for scientific troubleshooting requires domain-aware design choices that diverge significantly from general recommendations. For instance, we show that Hypothetical Document Embeddings (HyDE), often considered beneficial for query expansion, actually harm performance in technical domains where precision is paramount. Similarly, hierarchical chunking strategies underperform simple fixed-size approaches when complete procedural context is essential.

By providing `TroubleRAG` as a dedicated benchmark, we aim to develop robust, domain-aware retrieval systems that can provide reliable technical support in laboratory environments, ultimately accelerating scientific discovery through more effective human-AI collaboration in real-world chemistry troubleshooting.

2 Related Work

Large Language Models in Chemistry. The integration of Large Language Models (LLMs) into chemistry has progressed from evaluating general knowledge recall to developing specialized agents. Initial efforts focused on comprehensive benchmarks that assessed LLMs across diverse chemical tasks [5, 11]. These evaluations demonstrated promising performance on standardized chemistry questions but were fundamentally limited by their focus on closed-ended problems with definitive ground-truth answers, easily measured through accuracy-based metrics.

Recent research has advanced towards tool-augmented chemical agents. Systems like ChemCrow [1] integrate LLMs with external tools (e.g., synthesis planners, databases) to autonomously plan and execute complex procedures.

Retrieval-Augmented Generation for Scientific Domains. Retrieval-Augmented Generation (RAG) has been identified as a key method to enhance the factual accuracy and relevance of LLM outputs by grounding them in retrieved external knowledge [16]. Recent work has begun to develop RAG benchmarks specifically tailored to chemistry and related scientific fields, showing significant improvements over standalone LLM performance on chemistry-focused question-answering tasks [19]. These studies have established the value of domain-specific retrieval for chemical knowledge tasks and have begun to explore optimal configurations for scientific research.

However, a critical limitation persists across existing chemistry-focused LLM and RAG evaluations: they are largely designed for closed-ended knowledge recall, easily measured by metrics like accuracy or F1 score [11]. This leaves a significant gap in understanding how these systems perform on the open-ended, complex, and procedural problems that define real-world scientific research.

Our work directly addresses this evaluation gap by shifting focus from the well-studied question *"What is the property of X?"* to the practically urgent scenario *"My instrument is broken; what should I do?"*. This transition requires fundamentally different retrieval and generation capabilities, demanding systems that can navigate lengthy, technical procedural documentation; synthesize multi-step solutions

from distributed information sources and handle the inherent ambiguity of real troubleshooting scenarios.

3 Benchmark Dataset

3.1 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 1. The corpus is representative of real-world technical documentation, containing a mix of free text, bulleted checklists, parameter tables, warning boxes, and instrument photographs.

Table 1: 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)

3.2 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.

3.3 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) **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 at the NSF Center for Bioanalytic38 Metrology (CBM). 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.

4 Retrieval Framework for Systematic Evaluation

We conduct a controlled evaluation of retrieval-augmented generation (RAG) pipelines using a modular framework designed for the chemistry troubleshooting domain. This design allows for the analysis of state-of-the-art components across the retrieval process, as illustrated in Figure 1. Our objective is to identify the most effective strategies for retrieving relevant technical documentation from a newly curated corpus of real-world chemistry troubleshooting queries and documents.

Retrieval Models. We evaluate a diverse set of retriever algorithms and strategies that form the foundation of the pipeline. This includes:

- **Lexical & Sparse Models:** The classic lexical baseline BM25 [13] and the learned sparse retriever SPLADE-v2 [3], which expands queries with related terms.
- **Dense Model:** The state-of-the-art dense retriever BGE-large-en-v1.5 [17], indexed using FAISS-HNSW [8] for efficient approximate nearest neighbor search.
- **Hybrid Model:** A configuration that combines sparse and dense scores via Reciprocal Rank Fusion (RRF) [2].
- **Hypothetical Document Embeddings (HyDE):** To bridge the lexical and semantic gap between concise queries and verbose technical passages, we implement HyDE [4]. This strategy uses an instruction-tuned language model to generate a hypothetical ideal response to the query. The embedding of this "pseudo-document" is then used for retrieval with the dense model, aiming to capture the query's intent more effectively than its raw form.

Document Chunking. We explore two fundamentally different approaches to text segmentation. The first is a standard fixed-size, sentence-aware window (512 tokens with a 20-token overlap). The second is Small2Big, a hierarchical strategy where smaller chunks are first retrieved to identify and return larger, more contextually rich parent documents.

Post-Retrieval Reranking. We rigorously evaluate the utility of a refinement step by ablating three distinct configurations: no reranker, the established monoT5 [12] cross-encoder, and the more recent FlagReranker. For configurations employing a reranker, the top-50 candidates from the first-stage retriever are re-scored. The final top-5 documents are then selected from this refined ranking. This comparison is designed to quantify the performance gain from reranking itself and to determine the most effective model for this specific domain.

This systematic design provides a comprehensive grid of configurations for benchmarking and analysis, specifically tailored to the challenges of chemistry troubleshooting documentation retrieval.

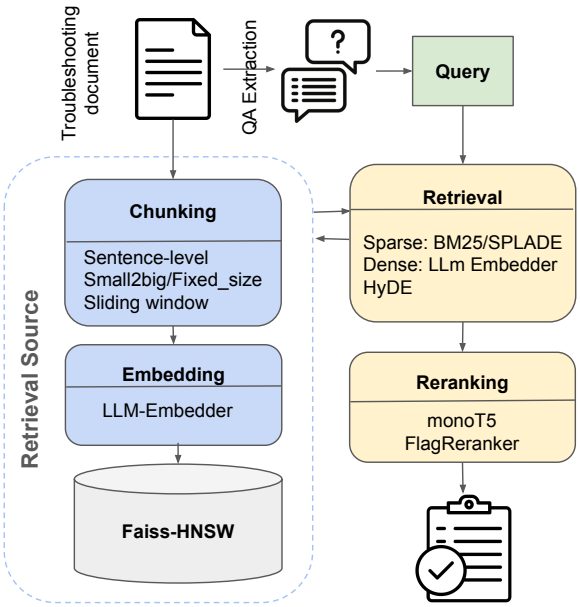


Figure 1: The modular retrieval pipeline used for evaluation. Source documents are processed (chunked) and indexed. User queries are processed by a first-stage retriever (sparse, dense, hybrid, or HyDE). The top-50 candidates are then optionally re-ranked by a cross-encoder (monoT5 or FlagReranker), and the final top-5 results are selected for output.

5 Evaluation

5.1 Experimental Setup

To measure the efficacy of different RAG components for chemistry troubleshooting, we conduct a series of controlled experiments on our newly curated dataset.

Our primary evaluation compares eight distinct pipeline configurations (ID1–ID8 in Table 2). These configurations are systematically derived from our modular framework (Section 4) to isolate the impact of individual components. We contrast these specialized configurations against a robust external baseline, termed Best Practice (BP). This baseline represents a generalized state-of-the-art setup derived from recent literature [16], which combines BM25 (sparse), LLM-Embedder (dense), HyDE query transformation, the Small2Big chunking strategy, and monoT5 reranking. This comparison allows us to rigorously test whether domain-specific tuning provides a significant advantage over a generic approach.

We quantify retrieval performance using four key metrics, all focused on the top-5 retrieved chunks. To measure ranking quality, we use Mean Average Precision (mAP) and Normalized Discounted Cumulative Gain (nDCG@5). To assess the ability to find the correct source, we use Recall@5. Finally, to evaluate semantic alignment beyond lexical overlap, we compute the maximum BERTScore [18] between the ground truth chunk and the retrieved candidates.

5.2 Results and Analysis

Our experimental evaluation, summarized in Table 2, reveals notable performance differences across the eight RAG configurations. The comprehensive comparison of these models across all four evaluation metrics is visualized in Figure 2, while Figure 3 provides a detailed distribution analysis and a focused comparison with the Best Practice (BP) baseline.

The results demonstrate that our domain-tailored configuration, ID7, achieves the best performance on our chemistry troubleshooting benchmark with a Recall@5 score of 0.9469. This configuration utilizing SPLADE for sparse retrieval, LLM Embedder for dense retrieval, fixed-size chunking, and

Table 2: TroubleRAG configurations with grouped components. Dense = LLM_Embedder; HyDE = hypothetical document expansion.

	Retrieval			Chunking	Reranking	Index	Metric
ID	**Sparse**	**Dense**	**HyDE**	**Method**	**Model**	**Vector DB**	**Recall@5**
1	BM25	✗	✗	Fixed 512	✗	—	0.7699
2	SPLADE	✗	✗	Fixed 512	✗	—	0.8849
3	✗	✓	✗	Fixed 512	✗	—	0.8673
4	SPLADE	✓	✗	Fixed 512	✗	FAISS-HNSW	0.8850
5	SPLADE	✓	✓	Fixed 512	✗	FAISS-HNSW	0.7168
6	SPLADE	✓	✗	Small2Big	✗	FAISS-HNSW	0.8673
7	SPLADE	✓	✗	Fixed 512	FlagReranker	FAISS-HNSW	**0.9469**
8	SPLADE	✓	✗	Fixed 512	monoT5	FAISS-HNSW	0.9115

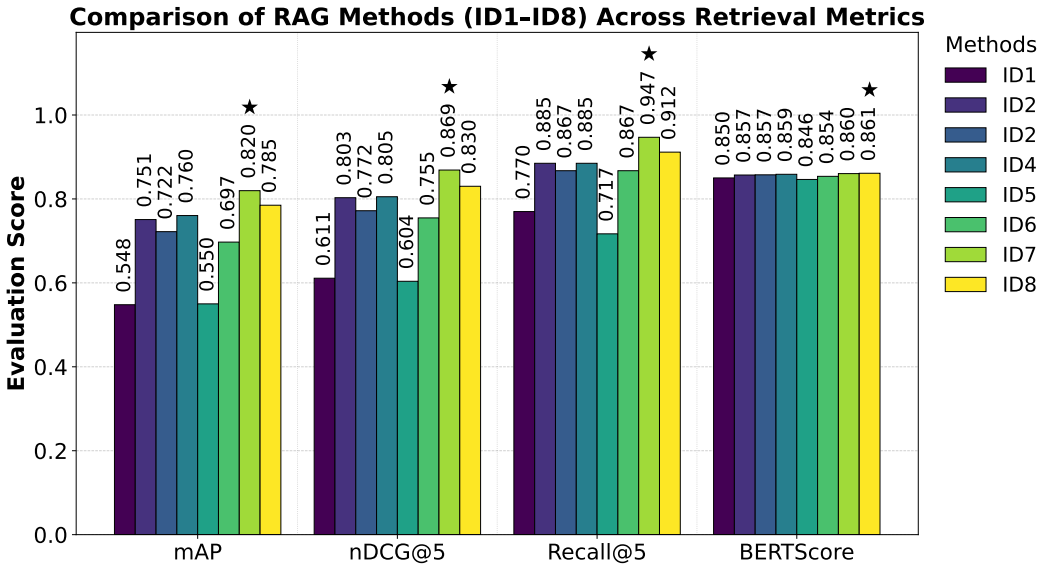


Figure 2: Retrieval performance of RAG methods (ID1–ID8) across mAP, nDCG@5, Recall@5, and BERTScore; stars mark the best method per metric.

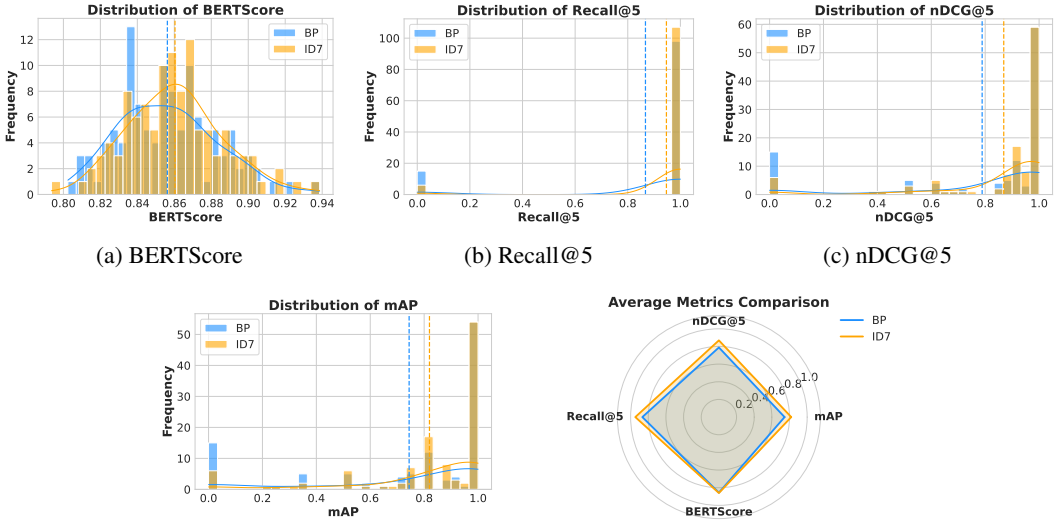
the FlagReranker, substantially outperforms the generic BP baseline across all metrics, as clearly visualized in the radar chart (Figure 3e). This performance advantage is critical for real-world applications, where successfully retrieving the correct troubleshooting procedure for a malfunctioning chromatograph or mass spectrometer can save hours of laboratory downtime.

Key Findings:

• **The FlagReranker is optimal for technical precision.** The significant gain of ID7 over ID8 (monoT5) and ID4 (no reranker) shows that reranking is essential in technical domains. FlagReranker distinguishes between closely related explanations and ensuring the most relevant diagnostic procedure is ranked highest.

• **HyDE is detrimental for factual, precise retrieval.** The underperformance of ID5 (Recall@5: 0.7168) demonstrates that *hypothetical* generation is ill-suited for a field governed by exact parameters. A HyDE-generated pseudo-document might invent a plausible but incorrect solution. This hallucination misdirects retrieval toward irrelevant chunks, delaying resolution for a time-sensitive experiment.

• **Advanced sparse retrieval (SPLADE) is highly effective for scientific terminology.** The strong standalone performance of ID2 (Recall@5: 0.8849) shows that SPLADE can expand short queries into the full set of terms used in technical manuals.. A concise query like "LC-MS sensitivity drop" can be effectively expanded to include terms like ["electrospray ionization", "capillary voltage",



(a) BERTScore (b) Recall@5 (c) nDCG@5

(d) mAP (e) Radar plot comparison

Figure 3: Comparison of retrieval metrics between the baseline best-practice pipeline (BP) and configuration ID7. (a–d) show the distributions of BERTScore, Recall@5, nDCG@5, and mAP, respectively.(e) provides a radar plot comparing the average performance of the two models across these metrics

"source contamination"], directly mirroring the language in technical manuals and retrieving chunks that address the exact instrument module involved.

• **Dense retrieval alone captures semantic intent but lacks precision.** The performance of ID3 (Dense only, Recall@5: 0.8673) shows that while semantic search can interpret the general meaning of a query, it often fails to distinguish between closely related technical issues. Its scope is too broad to consistently support precise diagnosis, highlighting the need to pair dense retrieval with a sparse retriever for reliable performance.

• **Simple, consistent chunking outperforms complex hierarchies.** The outperformance of ID7 (Fixed512) over ID6 (Small2Big) suggests that for complex procedures, retrieving the entire context is essen tial. The Small2Big strategy is fragile; it might retrieve a small child chunk containing a common error code (e.g., "Pressure Error"), but this code could be shared across many different instruments and root causes. This can easily lead to retrieving the wrong "parent" chunk from a different system. Fixed 512-token chunking, in contrast, often encapsulates an entire, self-contained troubleshooting procedure, ensuring the user receives the complete context needed for resolution.

5.3 Limitations of Current RAG

Our analysis reveals fundamental limitations in existing text-based RAG systems that render them inadequate for real-world laboratory troubleshooting scenarios. Unlike conventional information retrieval datasets [14], our source documents are rich with non-textual information, containing critical information encoded in instrument photographs, diagnostic diagrams, error screenshots, and complex parameter tables. Current RAG pipelines systematically ignore this non-textual content, creating significant gaps in retrievable knowledge.

Furthermore, practical troubleshooting is often a human-interactive, multi-turn process. A technician may need to answer clarifying questions or provide additional details based on initial results. To address these gaps, our findings identify two primary extensions:

1. **Multi-modal Information Integration:** Text-based retrieval cannot access the substantial technical knowledge embedded in visual content. Critical troubleshooting information requires visual analysis capabilities. For instance, questions like “How do I decrease my autotune window m/z width?” demand interpretation of instrument interface diagrams, while queries such as “I need to run my analysis at pH 10.5 using a phosphate buffer, but

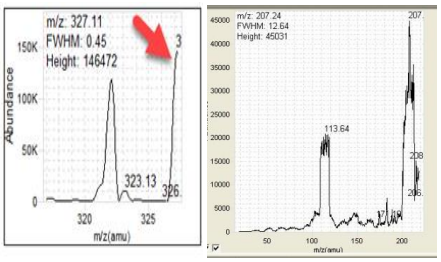


Figure 1: Two Examples of Contamination Which is Taller than the Actual Calibrant Peaks. These Taller Contaminants get Picked by the Tune Algorithm Instead of the Correct m/z, Causing Autotune to Fail

(a) A Figure Example

InfinityLab Poroshell Family		Pore Size	Temp. Limits	pH Range	Endcapped	Carbon Load	Surface Area
Best all around	EC-C18	120Å	60°C	2.0-8.0	Double	10%	130 m2/g
	EC-C8	120Å	60°C	2.0-8.0	Double	5%	130 m2/g
Best for low-pH mobile phases	SB-C18	120Å	90°C	1.0-8.0	No	9%	130 m2/g
	SB-C8	120Å	80°C	1.0-8.0	No	5.5%	130 m2/g
Best for high-pH mobile phases	HPH-C18	100Å	60°C	3.0-11.0	Double	Proprietary	95 m2/g
	HPH-C8	100Å	60°C	3.0-11.0	Double	Proprietary	95 m2/g
Best for basic compounds at low pH	CS-C18	120Å	60°C	2.0-8.0	Double	Proprietary	95 m2/g
Best for polar compounds (HILIC)	HILIC	120Å	60°C	0.0-8.0	N/A	N/A	130 m2/g
	HILIC-Z	120Å	80°C	3.0-11.0	Proprietary	Proprietary	130 m2/g
	HILIC-OH5	120Å	45°C	1.0-7.0	Double	Proprietary	130 m2/g
Best for alternative selectivity	Bonus-RP	120Å	60°C	2.0-9.0	Triple	9.5%	130 m2/g
	PFP	120Å	60°C	2.0-8.0	Double	5.1%	130 m2/g
	Phenyl-Hexyl	120Å	60°C	2.0-8.0	Double	9%	130 m2/g
	SB-Aq	120Å	80°C	1.0-8.0	No	Proprietary	130 m2/g
	EC-CN	120Å	60°C	2.0-8.0	Double	3.5%	130 m2/g
Best for Chiral separations	Chiral-T	120Å	45°C	2.5-7.0	Proprietary	Proprietary	130 m2/g
	Chiral-V	120Å	45°C	2.5-7.0	Proprietary	Proprietary	130 m2/g
	Chiral-CD	120Å	45°C	3.0-7.0	Proprietary	Proprietary	130 m2/g
	Chiral-CF	120Å	45°C	3.0-7.0	Proprietary	Proprietary	130 m2/g

Specifications represent typical values only

(b) A Table Example

Figure 4: Multi-modal examples in the source document.

I'm currently using an EC-C18 column which only goes up to pH 8.0. Why do I get poor peak shape?" require analysis of tabular specification data, as illustrated in Figure 4.

2. **Multi-turn Refinement:** Laboratory technicians rarely provide complete problem descriptions in initial queries, while effective troubleshooting requires systematic information gathering for effective troubleshooting. This diagnostic process inherently requires multi-turn interactions to progressively refine solutions. Consider the diagnostic gap between complete and typical queries:

   Diagnostic Information Gap in Real-World Queries

   Complete diagnostic query: Q: *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?*

   Typical incomplete query: Q: *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? Required diagnostic clarifications:* – Are environmental factors like room temperature stable? – Is there a trend in the fluctuations (upward, downward, random)? – Which type of data is fluctuating—signal intensity, retention time, or m/z?

   The inability to conduct such diagnostic dialogues fundamentally limits current systems to providing generic rather than targeted troubleshooting guidance.

These limitations explain why generic RAG configurations underperform in technical domains: they operate within architectural constraints that preclude access to essential multimodal information and interactive diagnostic processes that define expert-level troubleshooting.

6 Conclusion

In this work, we introduced `TroubleRAG`, a comprehensive framework for evaluating RAG pipelines on the challenging domain of laboratory instrument troubleshooting. We constructed a new, expert-validated dataset, and conducted a systematic evaluation of six retrieval configurations. Our central finding is that generic RAG "best practices" are insufficient for this specialized task. We presented a domain-tailored recipe—combining hybrid search with a strong reranker that significantly outperforms a robust baseline, boosting Recall@5 by nearly 8 percentage points. However, this work also highlights critical limitations in current retrieval techniques that present clear avenues for future research. Our source documents are rich with non-textual information, such as instrument photographs, diagrams, and complex data tables, which are largely ignored by current text-based RAG pipelines. Furthermore, practical troubleshooting is often an interactive, multi-turn process. A technician may need to answer clarifying questions or provide additional details based on initial results. To address these gaps, our future work will focus on two primary directions: the enhancement of multi-modal retrieval and interactive, multi-turn dialogue.

References

[1] Andres M Bran et al. "Chemcrow: Augmenting large-language models with chemistry tools". In: *arXiv preprint arXiv:2304.05376* (2023).

[2] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. "Reciprocal rank fusion outperforms condorcet and individual rank learning methods". In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 2009, pp. 758–759.

[3] Thibault Formal et al. "SPLADE v2: Sparse lexical and expansion model for information retrieval". In: *arXiv preprint arXiv:2109.10086* (2021).

[4] Luyu Gao et al. "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)*. 2023, pp. 1762–1777.

[5] Taicheng Guo et al. "What can large language models do in chemistry? a comprehensive benchmark on eight tasks". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 59662–59688.

[6] Ziwei Ji et al. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12 (2023), pp. 1–38.

[7] Qiao Jin 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)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 2567–2577. DOI: 10.18653/v1/D19-1259. URL: https://aclanthology.org/D19-1259.

[8] Jeff Johnson, Matthijs Douze, and Hervé Jégou. "Billion-scale similarity search with GPUs". In: *IEEE Transactions on Big Data* 7.3 (2019), pp. 535–547.

[9] Tom Kwiatkowski et al. "Natural questions: a benchmark for question answering research". In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 453–466.

[10] Patrick Lewis et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[11] A Mirza et al. "Are large language models superhuman chemists?, arXiv, 2024". In: *arXiv preprint arXiv:2404.01475* 10 ().

[12] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. "Document ranking with a pretrained sequence-to-sequence model". In: *arXiv preprint arXiv:2003.06713* (2020).

[13] Stephen Robertson, Hugo Zaragoza, et al. "The probabilistic relevance framework: BM25 and beyond". In: *Foundations and Trends® in Information Retrieval* 3.4 (2009), pp. 333–389.

[14] Nandan Thakur et al. "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021. URL: https://openreview.net/forum?id=wCu6T5xFjeJ.

[15] "Troubleshooting Liquid Chromatography-Tandem Mass Spectrometry in the Clinical Laboratory". In: *Clinical Laboratory News* (Aug. 2015). URL: https://myadlm.org/cln/articles/2015/august/troubleshooting-liquid-chromatography-tandem-mass-spectrometry-in-the-clinical-laboratory.

[16] Xiaohua Wang et al. "Searching for best practices in retrieval-augmented generation". In: *arXiv preprint arXiv:2407.01219* (2024).

[17] Peitian Zhang et al. "Retrieve anything to augment large language models". In: *arXiv preprint arXiv:2310.07554* (2023).

[18] Tianyi Zhang et al. "Bertscore: Evaluating text generation with bert". In: *arXiv preprint arXiv:1904.09675* (2019).

[19] Xianrui Zhong et al. "Benchmarking retrieval-augmented generation for chemistry". In: *arXiv preprint arXiv:2505.07671* (2025).