SPECTRA: A MULTI-SCENARIO BENCHMARK FOR GRANULAR RAG EVALUATION IN CHEMISTRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite rapid progress in large language models (LLMs) and agentic systems, retrieval-augmented generation (RAG) remains indispensable in any setting where authoritative answers must come from a controlled, often proprietary corpus such as scientific literature, internal company documentation, or regulatory archives. We introduce **SPECTRA** (**S**cenario-**P**artitioned **E**xpert **C**hemistry **T**estbed for **R**AG **A**ssessment), a benchmark for fine-grained evaluation of RAG systems on expert-level questions grounded in open-access scientific publications. The benchmark organizes 2007 questions across 8 chemistry domains into a two-dimensional taxonomy: 6 question scenarios (single-hop extraction, multi-hop reasoning, conditional filtering, comparison, aggregation, and multi-turn dialogue) that isolate distinct system capabilities and 3 orthogonal modification types (multimodal, constrained, and negative queries). We also propose an evaluation framework to assess LLMs and RAG configurations across several metrics and apply it to 7 RAG configurations and 3 standalone LLM baselines, with the primary Answer Accuracy judge validated against two experts on 200 usable examples. The benchmark evaluation yields two key findings. First, scenario- and modification-level evaluation on one domain reveals performance trade-offs invisible to aggregate scoring: configurations that appear equivalent by overall accuracy differ substantially across question types. This allows practitioners to optimize RAG configurations according to their primary use case. Second, the dedicated multi-turn scenario illustrates how per-turn analysis supports fault localization: document-level DOI Recall remains stable across dialogue turns for hybrid configurations, while evidence-level Source Support declines modestly and Answer Accuracy declines more strongly. This shows that degradation is not explained by complete failure to retrieve the source article. While the corpus is chemistry-focused, the scenario taxonomy and evaluation protocol are not intrinsically chemistry-specific and can be adapted to other scientific corpora. Beyond evaluation, the per-scenario performance vectors produced by our framework can support data-driven RAG configuration design, and the semi-automated question generation pipeline can suggest a route toward periodic question refresh to reduce contamination. The dataset and evaluation code are publicly available through anonymized repositories: dataset collection and evaluation code.

1 INTRODUCTION

Large language models (LLMs) and agentic systems are increasingly used in scientific workflows, including tool use, planning, and literature analysis (Yehudai et al., 2025; M. Bran et al., 2024; Boiko et al., 2023; Lála et al., 2023). However, in settings where answers must come from a fixed and auditable document collection, retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2023; Huang et al., 2025) remains necessary. This is especially true in chemistry, where useful answers often depend on exact experimental values, units, tables, figures, and source-specific claims scattered across full-text articles.

Prior work spans general RAG benchmarks, multi-hop and multi-turn evaluation, diagnostic frameworks, and chemistry-oriented QA (Tang & Yang, 2024; Katsis et al., 2025; Ru et al., 2024; Zhong et al., 2025). Section 2 reviews these strands; the gap we address is scenario-structured comparison of RAG configurations over scientific literature.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

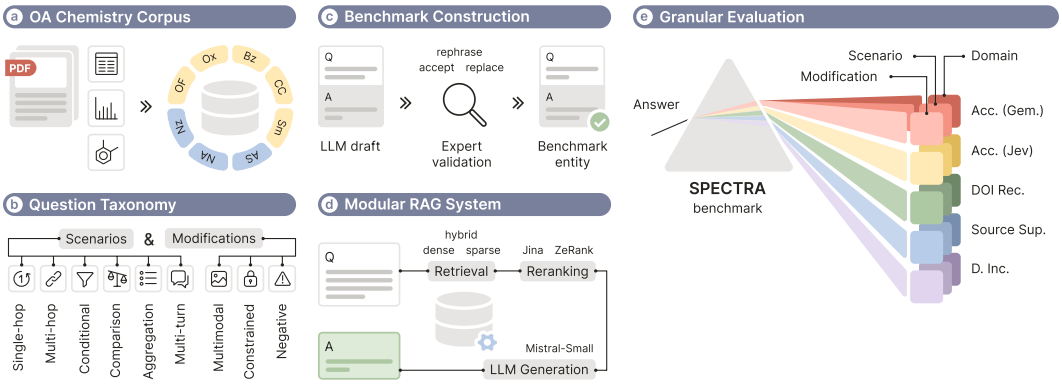

Figure 1: SPECTRA workflow. (a) Open-access chemistry corpus curation across 8 domains. (b) Scenario–modification question taxonomy. (c) Semi-automated question drafting followed by expert validation. (d) Modular RAG evaluation with configurable retrieval and reranking. (e) Granular evaluation across domains, scenarios, and modifications using answer, retrieval, grounding, and diagnostic metrics to compare RAG configurations.

A single aggregate score can hide different failure modes: the system may miss the relevant article, retrieve the right article but rank it too low, fail on tables or figures, lose information across dialogue turns, or answer when the corpus contains no answer. To study these differences, the benchmark itself must separate question types that stress different parts of the RAG pipeline.

We introduce SPECTRA, a scenario-partitioned benchmark for granular RAG evaluation in chemistry. The benchmark organizes expert-validated questions along two axes: six scenarios covering single-hop extraction, multi-hop reasoning, conditional filtering, comparison, aggregation, and multi-turn dialogue; and three modification types covering multimodal, constrained, and negative queries. This design makes it possible to compare RAG configurations by their scenario-level performance rather than only by overall accuracy.

SPECTRA pairs this taxonomy with a modular evaluation protocol that separates source discovery, evidence localization, answer correctness, and diagnostic failure signals. This makes it possible to connect performance differences to specific question types and pipeline stages rather than relying on a single aggregate score.

The benchmark workflow is illustrated in Figure 1. Our contributions are as follows:

1. We present a benchmark of 2007 expert-validated chemistry questions across 8 domains, grounded in 224 open-access publications and organized into a 6-scenario $\times$ 3-modification taxonomy. The questions are produced through a semi-automated pipeline with LLM drafting and expert validation, and we report correction statistics for the generated questions.
2. We introduce a RAG evaluation framework with retrieval, generation, and diagnostic metrics, compatible with standard tools such as RAGAS (Es et al., 2024), ARES (Saad-Falcon et al., 2024), and DeepEval (Confident AI, 2024). The RAG pipeline uses general-purpose ingestion, chunking, retrieval, and reranking components rather than chemistry-specific modules, so the same configuration-comparison protocol can be adapted to other scientific or domain-specific document collections.
3. We evaluate 7 RAG configurations and 3 standalone LLM baselines. The results show that configurations with similar aggregate accuracy can differ substantially across scenarios and modifications, supporting scenario-aware selection of RAG configurations.

2 Related Work

Table 5 in Appendix A provides a feature-level comparison of SPECTRA with existing RAG and chemistry benchmarks.

RAG benchmarks and evaluation frameworks. Several benchmarks have shaped the evaluation of RAG and knowledge-intensive QA. KILT (Petroni et al., 2021) unified several tasks over a fixed Wikipedia snapshot. CRAG (Yang et al., 2024) evaluates factual QA with web and knowledge-graph search APIs. RAGBench (Friel et al., 2024) provides large-scale explainable RAG examples across applied domains. MultiHop-RAG (Tang & Yang, 2024) focuses on multi-hop retrieval over news articles, while DRAGOn (Chernogorskii et al., 2026) studies benchmark updating to reduce contamination. RAGAS (Es et al., 2024), ARES (Saad-Falcon et al., 2024), DeepEval (Confident AI, 2024), and RAGChecker (Ru et al., 2024) provide reusable metrics and diagnostics for RAG evaluation. These works provide important datasets and metrics, but they are not designed to compare modular RAG configurations by scenario-specific performance.

Multi-hop and multi-turn RAG. HotpotQA (Yang et al., 2018) established multi-hop reasoning over linked Wikipedia evidence, and MultiHop-RAG (Tang & Yang, 2024) adapted this direction to RAG over news corpora. MTRAG (Katsis et al., 2025) evaluates multi-turn RAG conversations with answerability labels, and RAD-Bench (Kuo et al., 2025) studies changing constraints across dialogue turns. SPECTRA uses multi-hop and multi-turn questions as two scenarios within a broader taxonomy. This allows direct comparison with single-hop, conditional, comparison, and aggregation questions under the same corpus, metrics, and RAG configurations.

Scientific and chemistry RAG. PaperQA (Lála et al., 2023), ChemCrow (M. Bran et al., 2024), and TextReact (Qian et al., 2023) study scientific retrieval or tool-augmented reasoning, while ChemLit-QA (Wellawatte et al., 2025), ChemRAG-Bench (Zhong et al., 2025), and ChemPaperBench (Dubrovsky et al., 2025) benchmark chemistry QA or literature-grounded reasoning. ChemX (Vepreva et al., 2025) focuses on multimodal chemical information extraction, while SciRerankBench (Chen et al., 2025) and MRAG-Bench (Hu et al., 2024) isolate reranking or multimodal retrieval. SPECTRA instead targets scenario-specific comparison of RAG configurations over full-text chemistry literature. Several domains build on chemistry areas represented in ChemX; SPECTRA reconstructs the collections under an open-access-only requirement and contributes separate QA pairs, evidence annotations, distractors, taxonomy, and evaluation protocol.

Retrieval, reranking, and multimodal evidence. Other benchmarks study narrower components of scientific RAG. SciRerankBench (Chen et al., 2025) evaluates rerankers with difficult scientific distractors, Matter-of-Fact (Jansen et al., 2025) studies temporally grounded materials-science claims, and MRAG-Bench (Hu et al., 2024) focuses on multimodal retrieval. These directions are complementary to SPECTRA. Our benchmark combines source retrieval, reranking sensitivity, multimodal evidence, constrained queries, negative queries, and dialogue in a single chemistry literature testbed.

Positioning. Prior benchmarks answer whether retrieval helps, whether a model can solve multi-hop or multi-turn queries, whether a chemistry agent can reason over literature, or whether an extraction system can recover structured chemical data. SPECTRA targets a different question: which RAG configuration should be used for a given distribution of chemistry literature questions. Its scenario taxonomy, orthogonal modifications, and diagnostic metrics make it useful for ranking systems as well as for choosing and improving RAG configurations.

3 BENCHMARK DESIGN

SPECTRA is designed to attribute RAG performance differences to specific query types and pipeline stages rather than to a single aggregate score. The benchmark combines three components (Figure 1): an open-access corpus spanning eight chemistry domains, a taxonomy of six scenarios and three orthogonal modifications, and a modular evaluation framework with retrieval, generation, and diagnostic metrics. Although the corpus is chemistry-focused, the scenario taxonomy and evaluation protocol are not intrinsically chemistry-specific and can be adapted to other scientific corpora.

3.1 DOCUMENT CORPUS

All non-negative questions in SPECTRA are grounded in facts from open-access scientific publications (Gold or Diamond access based on OpenAlex (Priem et al., 2022)), ensuring full reproducibility

Table 1: Chemical domains in SPECTRA and key corpus statistics. Avg. tok. is mean document length in thousands; Inter-doc sim. is mean pairwise embedding similarity; table/image density are percentages of articles containing the respective modality.

Domain	Art.	Q.	Avg. tok.	Inter-doc sim.	Table dens., %	Image dens., %
Nanoparticle Antibacterials	41	245	4.25	0.74	81	91
Antibiotic-NP Synergy	40	254	3.73	0.70	75	81
Nanozymes	40	240	3.52	0.64	67	71
Benzimidazole Antibiotics	40	239	8.61	0.81	91	76
Oxazolidinone Antibiotics	40	246	6.49	0.76	96	61
Chelate Metal Complexes	62	275	9.22	0.75	76	81
Ophthalmic Formulations	30	263	5.08	0.55	73	73
Supramolecular Drug Systems	40	245	4.40	0.64	93	74

and enabling the inclusion of complete PDF documents in the retrieval corpus. SPECTRA spans eight chemical domains, each representing a distinct area of chemistry with characteristic data types and retrieval challenges (Table 1). The domains were selected to maximize thematic diversity: from nanomaterial characterization and antibiotic activity to thermodynamic constants and photostability of drug co-crystals (see Appendix B.1).

We characterize each domain along four retrieval-oriented features. Average document length (**Avg. tok.**) affects chunking and evidence localization, while inter-document similarity (**Inter-doc sim.**), the mean pairwise cosine similarity between document embeddings, measures how hard it is to discriminate relevant from irrelevant passages within a domain. Table and image densities (**Table dens.**, **Image dens.**) quantify the share of articles with structured visual content, which text-based pipelines often fail to preserve. Together, these statistics provide a RAG-oriented profile of each domain and motivate domain-aware retrieval design.

The eight domains span a wide range of corpus characteristics relevant to RAG: mean document length varies more than 2.5× (3.5–9.2k tokens), and inter-document similarity ranges from 0.55 (Ophthalmic Formulations) to 0.81 (Benzimidazole Antibiotics), implying substantially different retrieval difficulty. Table and image densities exceed 60% across all domains. This makes the corpus a useful stress test for pipelines that discard tables or figures, especially on multimodal questions. Together, these statistics show that SPECTRA is not a homogeneous corpus but a stress-test along multiple structural dimensions of scientific literature.

The retrieval database contains 333 articles: 224 source articles and 109 topically related distractors. This forms a controlled retrieval setting focused on discrimination among similar scientific sources rather than production-scale corpus search.

3.2 Question scenarios

Questions are organized into six scenarios targeting distinct reasoning capabilities, from simple fact extraction to multi-turn dialogue.

1. **Direct Fact Retrieval (Single-hop).** Extraction of a specific value or statement from a single source. Tests basic retrieval and information extraction.
2. **Sequential Retrieval (Multi-hop).** The answer to a first sub-question serves as the search key for a second source. Tests query decomposition and chained reasoning.
3. **Conditional Retrieval.** Multiple conditions from different sources must be satisfied simultaneously. Tests cross-document reasoning and set intersection.
4. **Comparison.** Entities extracted from separate sources must be compared quantitatively or qualitatively. Tests parallel retrieval and comparative reasoning.
5. **Aggregation (Set Query).** All entities satisfying a condition must be collected from across the corpus. Tests retrieval completeness and aggregation.

6. **Multi-turn Dialogue.** A sequence of thematically linked questions that may draw on different source articles within the same domain. Consecutive questions may share a topic, shift to a related subtopic, or address the same entity from a different angle. Tests context-dependent retrieval and coherent reasoning across topic transitions.

3.3 Question modifications

Orthogonally to scenarios, each question may carry one or more of three modification types. These modifications introduce additional complexity without changing the underlying reasoning pattern.

Multimodal. The answer is obtainable only from a non-textual source and is not duplicated in the article body. We distinguish two subtypes: *table*, where the answer must be read from a structured data table, and *image*, which covers all visual evidence (spectra, microscopy images, reaction schemes, plots).

Constrained. The question includes metadata-level conditions that restrict which sources are relevant: *theme* constraints specify a topical focus (e.g., "in studies on silver nanoparticles for wound healing"), *temporal* constraints reference publication dates or time periods of reported experiments, and *source* constraints target specific journals, authors, or article titles.

Negative (Null). Questions designed to probe robustness: *conflicting sources* present contradictory answers across papers; *incorrect queries* contain factually invalid premises; *unanswerable queries* have no answer in the corpus. This tests hallucination resistance and appropriate abstention.

The full distribution of questions across the scenario-modification matrix is reported in Appendix B.2 Table 6.

3.4 Question generation pipeline

Questions were generated through a semi-automated pipeline combining LLM-based drafting with expert validation. Candidate questions were drafted with GPT-5.2 between February 18 and March 26, 2026; this drafting model is distinct from both the RAG backbone and the evaluation judges. For Scenario 1, we retain each LLM draft together with its expert-reviewed version. After whitespace normalization, unchanged pairs are labeled *accepted*; for changed pairs, semantic similarity between the pre- and post-expert questions is used to distinguish *rephrased* from *replaced* examples (Appendix C).

Six first- and second-year PhD researchers participated in validation; two covered two domains each and the others one domain each. Every question underwent two-stage review: a domain expert checked scientific correctness and supporting evidence and accepted, revised, or replaced the candidate, after which a second expert checked answerability, evidence sufficiency, scenario compliance, format, and consistency. Uncertain cases were resolved jointly. Because this was iterative review and adjudication rather than independent labeling, conventional inter-annotator agreement is not applicable to the construction process. Each question record carries scenario and modification labels with subtypes, source DOIs, a verbatim quote or evidence reference (e.g. "Table 12" or "Figure 4a"), and for Scenario 1 the LLM-generated and expert-corrected versions side by side; a representative example and a catalog of expert correction patterns are in Appendix B.2 and Appendix C. All questions are phrased as plausible queries from a domain researcher and are grounded in the source literature; some can nevertheless be answered from parametric model knowledge, which we analyze separately in Appendix I. All questions avoid direct table, figure, and DOI references. Source metadata such as journal, year, or author may appear only in constrained questions where it is part of the intended modification.

4 Experimental results

We run the benchmark across all 8 domains, 6 scenarios, and 3 modification types using a single hybrid RAG configuration with the Jina reranker; full cross-domain results are reported in Appendix F. We use Chelate Metal Complexes as a configuration-study (Section 4.1) testbed because it has the

Table 2: Overall performance on Chelate Metal Complexes (mean ± SEM over seeds 42 and 43). Higher is better except for D. Incomplete; bold marks the best result.

System	Acc. (Gem.)↑	Acc. (Jev)↑	DOI Rec.↑	Source Sup. (%)↑	D. Inc.↓
Jina · Dense	$0.56_{\pm 0.01}$	$0.55_{\pm 0.00}$	$0.83_{\pm 0.02}$	$72.54_{\pm 0.53}$	$0.47_{\pm 0.01}$
Jina · Sparse	$0.48_{\pm 0.01}$	$0.46_{\pm 0.01}$	$0.62_{\pm 0.00}$	$64.72_{\pm 0.12}$	$0.56_{\pm 0.02}$
Jina · Hybrid	$0.64_{\pm 0.00}$	$0.61_{\pm 0.00}$	$0.86_{\pm 0.00}$	$72.60_{\pm 0.04}$	$0.48_{\pm 0.00}$
ZeRank · Dense	$0.56_{\pm 0.00}$	$0.54_{\pm 0.01}$	$0.84_{\pm 0.01}$	$\mathbf{72.99_{\pm 0.05}}$	$0.48_{\pm 0.01}$
ZeRank · Sparse	$0.48_{\pm 0.02}$	$0.45_{\pm 0.01}$	$0.61_{\pm 0.00}$	$64.73_{\pm 0.11}$	$0.57_{\pm 0.00}$
ZeRank · Hybrid	$0.65_{\pm 0.01}$	$0.61_{\pm 0.01}$	$0.86_{\pm 0.00}$	$72.55_{\pm 0.14}$	$\mathbf{0.44_{\pm 0.00}}$
W/o · Hybrid	$\mathbf{0.66_{\pm 0.01}}$	$\mathbf{0.63_{\pm 0.01}}$	$\mathbf{0.87_{\pm 0.00}}$	$72.70_{\pm 0.14}$	$\mathbf{0.44_{\pm 0.01}}$

largest article count, broad coverage of negative subtypes, and stable metric estimates, reducing noise in configuration comparisons. For headline comparative analyses we rely on Answer Accuracy, document-level DOI Recall, the evidence-level Source Support Max Similarity signal, and D. Incomplete. The cross-domain results show that the retained diagnostics capture scenario difficulty more sensitively than accuracy alone; in particular, D. Incomplete rises sharply from single-hop to aggregation questions. Full eight-domain results are reported in Appendix F. Appendix I further reports matched three-domain comparisons of all seven retrieval configurations, chunk-size/top-k sensitivity, backbone and end-to-end system comparisons, bootstrap uncertainty, and parametric answerability.

We additionally validate the automatic judge on 200 usable examples scored by two experts. Gemini-2.5-Flash reaches quadratic weighted $\kappa = 0.793$ and 0.763 for Answer Accuracy, compared with 0.812 between experts. As an independent check, Jev-1.13 (Almeida, 2026) reaches 0.809 and 0.736 on the same items. We also re-score the answers from the full Chelate configuration study with Jev (Table 2); the system ordering is nearly unchanged, with W/o · Hybrid remaining highest and the two reranked hybrid systems forming the next group. We retain Gemini as the primary judge because Jev is comparable on Answer Accuracy but agrees substantially less with experts on D. Incomplete when the SPECTRA diagnostic prompt is ported to Jev. Full agreement statistics and implementation details are reported in Appendix D. A closed-book analysis on the 254 Chelate questions shared across all systems finds 40.2% parametrically answerable and 32.7% closed-book-hard; on the latter subset, Jina-Hybrid reaches 0.52 Accuracy versus 0.20 for the strongest standalone model (Table 17; Appendix I.5).

We compare 7 retrieval/reranking configurations of the same RAG pipeline and 3 standalone LLM baselines on Chelate Metal Complexes. Because the LLM baselines operate without retrieval, DOI Recall, Source Support, and retrieval-conditioned diagnostics such as D. Incomplete are not directly comparable for them. They therefore appear in the scenario-level table, where Answer Accuracy is the common signal, while aggregate and modification-level tables focus on the seven RAG configurations. All reported uncertainty values are standard errors of the mean (SEM) over two random seeds (42, 43), computed as $s/\sqrt{n}$ across seeds.

4.1 Configuration study on Chelate Metal Complexes

Aggregate comparison. Table 2 reports overall performance of seven RAG configurations on the Chelate Metal Complexes domain using the validated headline metrics available for every run. W/o · Hybrid achieves the highest Gemini Answer Accuracy (0.66) and document-level DOI Recall (0.87), while W/o · Hybrid and ZeRank · Hybrid have the lowest D. Incomplete rate (0.44). Jev re-scoring is slightly stricter in absolute Accuracy but leaves the aggregate ordering nearly unchanged. Source Support Max Similarity adds a finer-grained retrieval view: dense and hybrid configurations cluster at 72.5–73.0%, whereas both sparse-only configurations are near 64.7%. Thus, DOI Recall and Source Support distinguish source-article discovery from localization of answer-bearing evidence. The two reranked hybrid configurations remain close behind on aggregate Accuracy, while sparse-only configurations trail on both Accuracy and retrieval coverage. These aggregates may, however, obscure configuration-specific trade-offs, so we analyze Answer Accuracy separately across scenarios and modification types.

Table 3: Answer Accuracy by scenario on Chelate Metal Complexes. Values are mean ± SEM over two random seeds (42, 43). Bold: best result per scenario within each block.

System	S1	S2	S3	S4	S5	S6
Jina · Dense	$0.69_{\pm 0.01}$	$0.26_{\pm 0.07}$	$0.56_{\pm 0.11}$	$\mathbf{0.67}_{\pm 0.01}$	$0.49_{\pm 0.03}$	$0.51_{\pm 0.01}$
Jina · Sparse	$0.45_{\pm 0.01}$	$0.46_{\pm 0.02}$	$0.67_{\pm 0.07}$	$0.22_{\pm 0.08}$	$0.36_{\pm 0.01}$	$0.55_{\pm 0.01}$
Jina · Hybrid	$0.67_{\pm 0.00}$	$0.59_{\pm 0.01}$	$0.79_{\pm 0.03}$	$0.62_{\pm 0.03}$	$0.42_{\pm 0.00}$	$0.64_{\pm 0.01}$
ZeRank · Dense	$\mathbf{0.70}_{\pm 0.00}$	$0.31_{\pm 0.03}$	$0.44_{\pm 0.03}$	$0.50_{\pm 0.00}$	$\mathbf{0.53}_{\pm 0.01}$	$0.53_{\pm 0.01}$
ZeRank · Sparse	$0.46_{\pm 0.00}$	$0.45_{\pm 0.02}$	$0.57_{\pm 0.19}$	$0.34_{\pm 0.01}$	$0.38_{\pm 0.00}$	$0.53_{\pm 0.01}$
ZeRank · Hybrid	$0.68_{\pm 0.01}$	$0.61_{\pm 0.00}$	$\mathbf{0.82}_{\pm 0.03}$	$0.59_{\pm 0.01}$	$0.49_{\pm 0.01}$	$0.64_{\pm 0.03}$
W/o · Hybrid	$0.68_{\pm 0.00}$	$\mathbf{0.63}_{\pm 0.01}$	$0.78_{\pm 0.01}$	$0.59_{\pm 0.01}$	$0.49_{\pm 0.00}$	$\mathbf{0.68}_{\pm 0.03}$
GPT-5.4	$\mathbf{0.57}_{\pm 0.00}$	$0.50_{\pm 0.00}$	$\mathbf{0.76}_{\pm 0.06}$	$\mathbf{0.52}_{\pm 0.01}$	$\mathbf{0.87}_{\pm 0.01}$	$\mathbf{0.58}_{\pm 0.00}$
GPT-5.4 no reasoning	$0.52_{\pm 0.02}$	$\mathbf{0.54}_{\pm 0.01}$	$0.67_{\pm 0.03}$	$0.34_{\pm 0.06}$	$0.80_{\pm 0.02}$	$0.51_{\pm 0.01}$
Qwen-3.6-27B	$0.30_{\pm 0.03}$	$0.27_{\pm 0.03}$	$0.61_{\pm 0.03}$	$0.43_{\pm 0.03}$	$0.71_{\pm 0.03}$	$0.39_{\pm 0.02}$

Table 4: Answer Accuracy by modification on Chelate Metal Complexes. Mult. · Table/Image denotes table- or figure-grounded questions. Values are mean ± SEM over seeds 42 and 43; bold indicates the best result per column.

System	Base	Negative	Constrained	Mult · Table	Mult · Image
Jina · Dense	$0.66_{\pm 0.01}$	$0.43_{\pm 0.00}$	$0.46_{\pm 0.01}$	$0.36_{\pm 0.02}$	$0.25_{\pm 0.05}$
Jina · Sparse	$0.54_{\pm 0.00}$	$0.49_{\pm 0.05}$	$0.42_{\pm 0.02}$	$0.36_{\pm 0.03}$	$0.20_{\pm 0.00}$
Jina · Hybrid	$0.75_{\pm 0.00}$	$0.43_{\pm 0.05}$	$\mathbf{0.54}_{\pm 0.00}$	$0.42_{\pm 0.02}$	$\mathbf{0.38}_{\pm 0.01}$
ZeRank · Dense	$0.69_{\pm 0.00}$	$0.38_{\pm 0.01}$	$0.41_{\pm 0.00}$	$0.31_{\pm 0.01}$	$0.33_{\pm 0.00}$
ZeRank · Sparse	$0.53_{\pm 0.01}$	$\mathbf{0.61}_{\pm 0.03}$	$0.43_{\pm 0.02}$	$0.37_{\pm 0.04}$	$0.21_{\pm 0.01}$
ZeRank · Hybrid	$0.76_{\pm 0.01}$	$0.58_{\pm 0.05}$	$\mathbf{0.54}_{\pm 0.01}$	$0.45_{\pm 0.02}$	$0.35_{\pm 0.02}$
W/o · Hybrid	$\mathbf{0.77}_{\pm 0.01}$	$0.55_{\pm 0.01}$	$\mathbf{0.54}_{\pm 0.00}$	$\mathbf{0.46}_{\pm 0.01}$	$0.30_{\pm 0.02}$

Scenario-level results. Table 3 breaks down answer accuracy by scenario. RAG configurations outperform LLM baselines on five of the six scenarios. Among RAG systems, ZeRank · Hybrid achieves the highest accuracy on Scenario S3 (Conditional Retrieval, 0.82), while W/o · Hybrid leads on S2 (Multi-hop, 0.63) and S6 (Multi-turn, 0.68). Dense-based configurations perform best on S1 (Direct Fact Retrieval) and S4 (Comparison). The only consistent reversal appears in Scenario S5 (Aggregation Query), where all three LLM baselines surpass every RAG configuration: GPT-5.4 with reasoning reaches 0.87 against a RAG maximum of 0.53. Aggregation queries require recovering a complete set of entities or properties satisfying a condition, and parametric models appear to benefit when this set is a canonical, frequently cited inventory in the literature, which they recall holistically; RAG retrieves top-k chunks and systematically undercovers such sets. Concrete examples are given in Appendix G. Table 3 shows that no single configuration achieves the highest accuracy on every scenario, and the per-scenario spread for individual configurations is comparable to or larger than the aggregate spread between configurations. We return to the implications of this pattern in Section 5.

Modification-level results. Table 4 reports Answer Accuracy by modification type and subtype. Questions without modifications are the easiest across all configurations, with the best systems reaching 0.75–0.77 Accuracy. Negative questions form the next tier: performance drops relative to unmodified questions, but the strongest configurations still reach 0.55–0.61, indicating that several systems can handle unanswerable, incorrect, or conflicting queries reasonably well. Constrained questions are harder, with the best Accuracy limited to 0.54. This suggests that metadata-level restrictions, such as source, theme, or temporal constraints, create additional retrieval and grounding difficulty. Multimodal questions are the most challenging. Table-based questions reach at most 0.46, while image-based questions reach at most 0.38. The drop persists despite structured table ingestion and a separate OpenCLIP figure index in our pipeline, indicating that retrieving, localizing, and using non-textual evidence remains a bottleneck for the evaluated configurations.

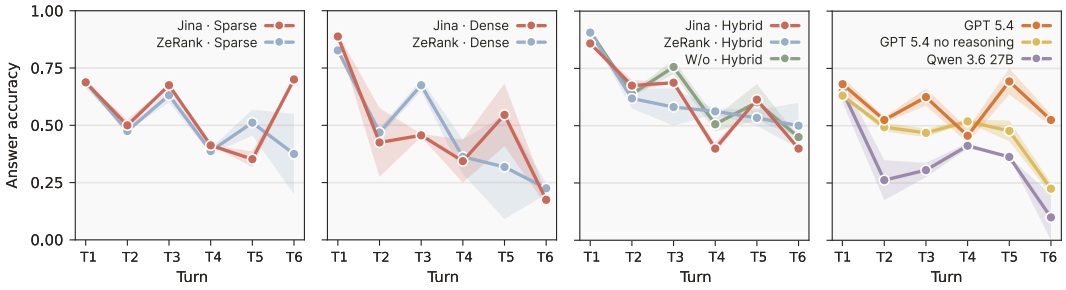

Figure 2: Answer accuracy by dialogue turn for the S6 multi-turn scenario on Chelate Metal Complexes. Systems are grouped by retrieval setup: sparse, dense, hybrid, and standalone LLM baselines. Points show mean accuracy over two random seeds; shaded bands denote SEM.

Multi-turn scenario. Figure 2 reports mean answer accuracy per turn for all seven RAG configurations and three LLM baselines. Two patterns are visible. First, accuracy declines with dialogue depth for six of the seven RAG systems, while the LLM baselines fluctuate around their means without a comparable monotonic drop; nevertheless, RAG configurations on average outperform LLMs at every turn, so the degradation does not erase the RAG advantage. Second, the per-metric breakdown (Table 10, Appendix H) shows that document-level DOI Recall stays stable or rises slightly across turns for all hybrid configurations, while Source Support Max Similarity declines modestly (about -0.03/turn) and Answer Accuracy declines more strongly. The degradation is therefore not explained by complete failure to find the source article; weaker evidence localization contributes, while query reformulation, context use, and generation remain plausible downstream causes.

5 Discussion

Scenario- and modification-level evaluation enables profile-driven configuration selection. Existing benchmarks either organize questions into scenario types or compare RAG configurations, but none does both on scientific corpora. Without a per-capability profile it is impossible to tell whether a configuration fails on a specific query type or uniformly, or to match a configuration to a target question distribution. SPECTRA fills this gap by isolating distinct capabilities in separate scenarios and layering orthogonal modifications.

A first analysis on the Chelate Metal Complexes domain suggests that no single configuration dominates across capability types, and that the performance gap between configurations widens with task complexity: differences modest on single-fact extraction become substantial on cross-document chaining or simultaneous constraint satisfaction. Within the seven SPECTRA retrieval/reranking configurations, the only consistent reversal appears in Scenario S5 (Aggregation), where all three standalone LLM baselines outperform the evaluated configurations: GPT-5.4 reaches 0.87 against a RAG maximum of 0.53, suggesting that parametric memory handles inventory-style queries more completely than top-k retrieval. This behavior is pipeline-dependent rather than inherent to RAG: on the matched GPT-OSS-20B Chelate sample, RAGU-Naive (Komarov et al., 2026) reaches 0.93 on the 10 aggregation questions versus 0.55 for the SPECTRA Jina-Hybrid pipeline (Appendix I.6). For negative queries with conflicting or unanswerable inputs, hybrid and sparse configurations with a ZeRank reranker offer better robustness against confident hallucination. Multimodal questions remain uniformly difficult (0.20–0.46), pointing to table- and figure-grounded retrieval as an open engineering problem. The three-domain studies further show that both scenario difficulty and the leading retrieval configuration vary by domain (Appendix I), reinforcing the value of per-domain performance profiles.

Multi-turn evaluation. The per-turn decomposition in Section 4.1 complements prior multi-turn benchmarks (Katsis et al., 2025; Kuo et al., 2025) by jointly reporting document-, evidence-, and answer-level signals within the same scientific RAG setting; Appendix H.2 illustrates a concrete coreference-related failure. Multi-agent architectures with a dedicated coreference resolution step would be a natural next system class to evaluate; SPECTRA's scenario structure and per-turn metrics would apply directly, though such a comparison is beyond the scope of this work.

5.1 FUTURE DIRECTIONS

Our objective with SPECTRA extends beyond a single chemistry-focused evaluation. We highlight three directions. *Data-driven configuration design.* Existing automated tuning approaches such as AutoRAG-HP (Fu et al., 2024) optimize a single aggregate metric on general-domain corpora, whereas our benchmark provides a richer per-scenario, per-modification performance vector for learning-based design. Several SPECTRA domains have already served as evaluation benchmarks for a separate study of machine learning-driven RAG system design (see Appendix M). Extending this to the full scenario-modification taxonomy is a natural next step. *Session-aware evaluation primitives.* Our multi-turn results expose a gap in existing frameworks: RAGAS (Es et al., 2024) and DeepEval (Confident AI, 2024) assess questions independently and aggregate linearly, missing session-level structure. Metrics such as per-turn accuracy degradation, evidence-localization trajectories across coreference chains, and retrieval–answer asymmetries motivate additional session-aware primitives. *Dynamic question generation.* Correction statistics from question generation provide a catalogue of failure patterns in LLM-drafted scientific questions. These can drive a dynamic benchmark that periodically produces new question sets, similar to DRAGOn (Chernogorskii et al., 2026), while preserving the scenario structure.

6 LIMITATIONS

Several limitations bound the scope of our conclusions. First, the primary judge is an LLM and may carry systematic biases on open-ended scientific text (Zheng et al., 2023; Liu et al., 2023; Kim et al., 2024); its agreement with experts and with an independent second judge is quantified in Appendix D. Second, the 333-article corpus is a controlled retrieval environment rather than a production-scale collection, and the eight domains are concentrated in biomedical, pharmaceutical, coordination, and nanomaterial chemistry. Third, the main configuration study varies retrieval and reranking with a fixed embedder and backbone; the backbone and RAGU analyses broaden this scope, but we provide no controlled embedding ablation, and the ingestion pipeline uses no chemistry-specific parsers or representations. Fourth, the cross-domain, chunk-size/top-k, backbone, and cross-system analyses use 50-question scenario-stratified samples and a single seed, and serve as sensitivity studies rather than high-precision rankings. Finally, multi-turn sessions are thematically linked question sequences rather than free-form conversations with clarification exchanges.

7 CONCLUSION

We introduced SPECTRA, a benchmark of 2007 expert-validated questions across eight chemistry domains, organized into a two-dimensional taxonomy of six scenarios and three orthogonal modification types, and applied it to seven RAG configurations and three standalone LLMs. Our central finding is that aggregate accuracy is an insufficient basis for RAG configuration decisions in scientific settings. Scenario- and modification-level decomposition in the configuration study on Chelate Metal Complexes reveals that configurations indistinguishable by overall accuracy diverge substantially across question types, and that the optimal choice depends on the capability the system is expected to exercise most frequently. RAG configuration selection is a matching problem between pipeline properties and a target query distribution, and SPECTRA provides the structure needed to perform that matching in a principled way. Cross-domain analyses further show that both scenario difficulty and the leading retrieval configuration vary by domain, reinforcing profile-based rather than aggregate evaluation.

We believe the value of this work extends beyond its immediate results. The per-scenario performance vectors produced by our framework constitute a structured signal for data-driven RAG configuration design; the semi-automated generation pipeline provides a foundation for dynamic benchmark versioning; and the session-level metric structure motivates additional dialogue-aware evaluation primitives. Human and alternative-judge validation, cross-domain and retrieval-parameter studies, backbone and end-to-end system comparisons, and parametric-answerability analysis further characterize the reliability and scope of the protocol. We see SPECTRA not as a fixed leaderboard for a narrow domain, but as a reusable methodology for systematic, diagnostic evaluation of retrieval-augmented systems wherever authoritative answers must be grounded in a controlled corpus.

AI USE STATEMENT

Large language models were used to draft candidate benchmark questions, implement the LLM-based evaluation studied in this work, assist with literature discovery, and support manuscript drafting and language editing. Candidate questions were reviewed and revised or replaced by domain experts, and all experimental claims and manuscript content were reviewed by the authors.

REPRODUCIBILITY STATEMENT

The benchmark schema, question-construction procedure, expert-review workflow, model and judge settings, retrieval architecture, and evaluation prompts are documented in the main text and appendices. The anonymized dataset and evaluation code are provided with the submission; the appendix also reports the additional judge, configuration, chunk-size/top-k, backbone, parametric-answerability, and RAGU analyses.

REFERENCES

Diogo Almeida. Introducing system one models & jev. TypeSafe AI Blog, September 2026. URL https://typesafe.ai/blog/introducing-system-one-models-and-jev. Accessed September 2026.

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

Haotian Chen, Qingqing Long, Meng Xiao, Xiao Luo, Wei Ju, Chengrui Wang, Xuezhi Wang, Yuanchun Zhou, and Hengshu Zhu. Scirerankbench: Benchmarking rerankers towards scientific retrieval-augmented generated llms. *arXiv preprint arXiv:2508.08742*, 2025.

Fedor Chernogorskii, Sergei Averkiev, Liliya Kudraleeva, Zaven Martirosian, Maria Tikhonova, Valentin Malykh, and Alena Fenogenova. Dragon: Designing rag on periodically updated corpus. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 622–638, 2026.

Confident AI. DeepEval: The open-source evaluation framework for large language model applications. https://github.com/confident-ai/deepeval, 2024. Accessed: 2026-05-05.

Ivan Dubrovsky, Kamil Fatkhiev, Anastasia Vepreva, Anastasia Orlova, Nina Gubina, Dina Shakirova, Rodion Golovinskii, Al Abbass Mohamed, Andrei Dmitrenko, Nikolay O Nikitin, et al. Chemparperbench: A multi-domain benchmark for literature-grounded chemical reasoning of llm-based multi-agent systems. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 68–75. IEEE, 2025.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pp. 150–158, 2024.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*, 2024.

Jia Fu, Xiaoting Qin, Fangkai Yang, Lu Wang, Jue Zhang, Qingwei Lin, Yubo Chen, Dongmei Zhang, Saravan Rajmohan, and Qi Zhang. Autorag-hp: Automatic online hyper-parameter tuning for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3875–3891, 2024.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, et al. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32, 2023.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182*, 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Peter Jansen, Samiah Hassan, and Ruoyao Wang. Matter-of-fact: A benchmark for verifying the feasibility of literature-supported claims in materials science. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4090–4102, 2025.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808, 2025.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024.

Mikhail Komarov, Ivan Bondarenko, Stanislav Shtuka, Oleg Sedukhin, Roman Shuvalov, Yana Dementyeva, Matvey Solovyov, and Nikolay O Nikitin. Ragu: A multi-step graphrag engine with a compact domain-adapted llm. *arXiv preprint arXiv:2607.11683*, 2026.

Tzu-Lin Kuo, FengTing Liao, Mu-Wei Hsieh, Fu-Chieh Chang, Po-Chun Hsu, and Da-Shan Shiu. Rad-bench: Evaluating large language models' capabilities in retrieval augmented dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 868–902, 2025.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

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

Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, 2021.

Nicholas Pipitone, Ghita Houir Alami, Advaith Avadhanam, Anton Kaminskyi, and Ashley Khoo. zelo: Elo-inspired training method for rerankers and embedding models. *arXiv preprint arXiv:2509.12541*, 2025.

Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

Yujie Qian, Zhening Li, Zhengkai Tu, Connor Coley, and Regina Barzilay. Predictive chemistry augmented with text retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12731–12745, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pp. 3982–3992, 2019.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027, 2024.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, 2024.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*, 2024.

Yixuan Tang and Yi Yang. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*, 2024.

Anastasia Vepreva, Julia Razlivina, Mariia Eremeyeva, Nina Gubina, Anastasia Orlova, Aleksei Dmitrenko, Xenia Kapranova, Susan Jyakhwo, Nikita Vasilev, Arsen Sarkisyan, Ivan Chernyshov, Vladimir Vinogradov, and Andrei Dmitrenko. Chemx: A collection of chemistry datasets for benchmarking automated information extraction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9e08a1db869a9646418e3371b24c6ae6-Paper-Datasets_and_Benchmarks_Track.pdf.

Feng Wang, Yuqing Li, and Han Xiao. Jina-reranker-v3: Last but not late interaction for listwise document reranking. *arXiv preprint arXiv:2509.25085*, 2025.

Geemi P Wellawatte, Huixuan Guo, Magdalena Lederbauer, Anna Borisova, Matthew Hart, Marta Brucka, and Philippe Schwaller. Chemlit-qa: a human evaluated dataset for chemistry rag tasks. *Machine Learning: Science and Technology*, 6(2):020601, 2025.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze D Gui, Ziran W Jiang, Ziyu Jiang, et al. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*, 2025.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Xianrui Zhong, Bowen Jin, Siru Ouyang, Yanzhen Shen, Qiao Jin, Yin Fang, Zhiyong Lu, and Jiawei Han. Benchmarking retrieval-augmented generation for chemistry. *arXiv preprint arXiv:2505.07671*, 2025.

A Related works

Table 5 provides a feature-level comparison of SPECTRA with existing RAG benchmarks across six dimensions: target domain, presence of an explicit question-type taxonomy, support for multimodal evidence (tables, figures, spectra), multi-turn dialogue, negative or unanswerable queries, and diagnostic metrics that attribute errors to specific pipeline stages. Each row reflects the design choices made by the corresponding benchmark; partial support is indicated where a feature is present in a limited form rather than as a first-class capability. Within the dimensions compared in Table 5, SPECTRA combines all six within a single chemistry-literature benchmark, enabling granular diagnosis of modular RAG systems across multiple axes simultaneously. The diagnostic decomposition itself is not unique to SPECTRA; frameworks such as RAGChecker, RAGAS, and ARES provide complementary retrieval/generation diagnostics.

Table 5: Comparison of SPECTRA with existing RAG benchmarks. ✓ = supported, (✓) = partial support, — = not supported. "Scenarios" indicates explicit question-type taxonomy; "Diagnostic" indicates metrics attributing errors to pipeline stages.

Benchmark	Domain	Scenarios	Multi-modal	Multi-turn	Negative	Diagnostic
MultiHop-RAG	News	4 types	—	—	✓	—
DRAGOn	News	4 types	—	—	—	—
HotpotQA	Wikipedia	2 types	—	—	—	—
GaRAGe	General	—	—	—	✓	(✓)
MTRAG	General	(✓)	—	✓	✓	—
RAD-Bench	General	(✓)	—	✓	(✓)	(✓)
RAGChecker	General	—	—	—	—	✓
RAGAS / ARES	General	—	—	—	—	✓
ChemRAG-Bench	Chemistry	—	—	—	—	—
ChemPaperBench	Chemistry	3 levels	✓	—	—	—
Ours	Chemistry	6 types	✓	✓	✓	✓

B Dataset Details

B.1 Description of benchmarks

1. **Nanoparticle Antibacterials:** antibacterial activity of inorganic nanoparticles against standard and multidrug-resistant bacterial strains.
2. **Antibiotic-NP Synergy:** combined antibacterial effects of nanoparticles and conventional antibiotics.
3. **Nanozymes:** catalytic properties of enzyme-mimicking nanoparticles, covering composition, morphology, substrate specificity, and kinetic parameters.
4. **Benzimidazole Antibiotics:** structures and minimum inhibitory concentrations of benzimidazole-based antibiotics against various bacterial species.
5. **Oxazolidinone Antibiotics:** oxazolidinone-class antibiotic structures and inhibitory concentrations against various bacterial species, with extraction provenance metadata.
6. **Chelate Metal Complexes:** thermodynamic stability constants (log K) of metal-ligand chelate complexes (Ga, Gd, Tc, Lu) for MRI contrast agent design.
7. **Ophthalmic formulations:** ophthalmic therapy of the corneal surface and tissues.
8. **Supramolecular Drug Systems:** photostability of pharmaceutical co-crystals and cyclodextrin inclusion complexes, relating crystal form and host-guest chemistry to degradation resistance.

B.2 QUESTION DISTRIBUTION AND DATA FORMAT.

Question distribution. Table 6 reports the distribution of questions across the scenario-modification matrix.

Table 6: Scenario–modification distribution (% within each scenario). Modifications are not mutually exclusive; Base denotes no modification. Total is the absolute question count.

Scenario	Base (%)	Multimodal (%)	Constrained (%)	Negative (%)	Total
1. Single-hop	47.5	29.9	25.1	8.4	646
2. Multi-hop	20.3	42.4	53.8	9.5	158
3. Conditional	39.1	33.7	39.1	9.5	169
4. Comparison	35.5	45.8	21.7	13.3	166
5. Aggregation	49.1	17.4	41.9	4.2	167
6. Multi-turn	42.2	29.5	31.8	6.4	701
Total	42.0	31.3	32.0	7.9	2007

Data format. The benchmark is distributed as JSON with one entry per multi-turn session (single-turn scenarios use a single-element `turns` list). Each session contains the scenario type, and each turn carries a question, golden answer, list of evidence entries (with DOI, source quotes or table/figure references, applied modifications, and contrast DOIs for conflicting-source questions). A representative session is shown below.

```
{
  "session_id": "oxazolidinones_s1_session_68",
  "scenario_type": "Direct Fact Retrieval",
  "turns": [{
    "question_id": "oxazolidinones_s1_q68",
    "question": "At which specific codon does the ribosome stall when
                treated with chloramphenicol or linezolid in the
                absence of the Cfr modification?",
    "golden_answer": "UUC",
    "evidence": [{
      "doi": "10.1101/2021.08.10.455846",
      "sources": ["Figure 4a"],
      "modifications": {"multimodal": ["image"]},
      "contrast_doi": []
    }]
  }]
}
```

C QUESTION GENERATION AND VALIDATION

To quantify the extent of expert intervention, each LLM-generated question was systematically compared to its expert-validated counterpart. Table 7 reports correction statistics for Scenario 1 questions. Pairs that were identical after whitespace normalization were labeled as *accepted*. For all remaining pairs, semantic similarity was estimated using a pretrained cross-encoder model (`cross-encoder/stsb-roberta-large`) from the Sentence-Transformers framework Reimers & Gurevych (2019). The model directly predicts a similarity score for each sentence pair.

A conservative similarity threshold (≥ 0.85) was applied to distinguish *rephrased* from *replaced* questions. This threshold was selected based on manual calibration: lower thresholds systematically overestimated semantic equivalence by classifying overly general LLM-generated questions as acceptable paraphrases. In contrast, expert revisions frequently introduced source or temporal constraints, as well as explicit reference to experimental conditions. Increasing the threshold improved alignment with expert judgment by treating such constraints as substantive modifications rather than superficial rewordings. Some domains with highly structured tabular data yielded more directly usable LLM outputs, while those requiring interpretation of spectroscopic or other visual data required more expert intervention.

Table 7: Correction statistics for the 646 Scenario 1 questions. Percentages are computed within each domain. *Accepted*: LLM output used without changes. *Rephrased*: factual content preserved, wording improved by expert. *Replaced*: LLM output rejected, question rewritten by expert.

Domain	Accepted (%)	Rephrased (%)	Replaced (%)
Nanoparticle Antibacterials	61.6	7.0	31.4
Antibiotic-NP Synergy	65.0	7.5	27.5
Nanozymes	77.8	2.5	19.8
Benzimidazole Antibiotics	32.5	8.8	58.8
Oxazolidinone Antibiotics	32.5	13.8	53.8
Chelate Metal Complexes	28.7	23.8	47.5
Ophthalmic Formulations	47.5	10.0	42.5
Supramolecular Drug Syst.	64.6	17.7	17.7
Overall	**51.4**	**11.3**	**37.3**

Table 7 shows that overall 37.3% of questions were replaced and 11.3% rephrased, with substantial variation across domains. Expert corrections cluster into three recurring patterns.

Ambiguous phrasing. LLM-generated questions often omitted essential experimental context, making them answerable by multiple sources or unresolvable without the specific article. Experts added reagent identities, experimental conditions, and clinical context to enforce unambiguous grounding. For example:

> *Before:* "What reagent served as the in situ CO source in the kit-like preparation of the fac-$[^{99m}Tc(CO)_3(OH_2)_3]^+$ precursor?"
> *After:* "Which compound serves as the carbon monoxide source for forming the $[^{99m}Tc(CO)_3]^+$ core from $^{99m}TcO_4{}^-$?"

Reference to specific source. LLMs frequently anchored questions to figures, tables, or workflow diagrams visible in the PDF, producing questions that explicitly referenced the source document. Experts rewrote these as standalone scientific queries grounded in domain knowledge rather than document structure. For example:

> *Before:* "According to the apoptosis quantification graph of transplanted cells (Figure 5B), what is the approximate number of apoptotic UMSC-derived cells measured at 24 hours after transplantation?"
> *After:* "In the murine model of congenital corneal disease treated with umbilical mesenchymal stem cell transplantation, what is the approximate number of apoptotic UMSC-derived cells measured at 24 hours after transplantation?"

Overly narrow or source-dependent scope. Some questions were too narrow to be retrievable without knowing the exact article (publication year, journal). Experts added document-specific anchors while retaining the chemical context. For example:

> *Before:* "What is the conditional stability constant (log K) at pH 7.4 given for the Gd(III) complex of 3,4,3-LI(1,2-HOPO)?"
> *After:* "In a 2018 *Scientific Reports* paper, what is the conditional stability constant (log K) at pH 7.4 for the Gd(III) complex of 3,4,3-LI(1,2-HOPO)?"

These patterns explain the domain-level variation in Table 7: domains with highly structured tabular data (Nanozymes: 77.8% accepted) yielded more directly usable LLM outputs, while domains requiring interpretation of complex experimental protocols or spectroscopic data (Benzimidazole Antibiotics: 58.8% replaced, Oxazolidinone Antibiotics: 53.8% replaced) required substantially more expert intervention.

Scaling and versioning. The first release prioritizes expert validation, but future versioned releases can reduce the manual burden through schema and evidence-consistency checks, duplicate detection,

Table 8: Judge–human agreement on 200 usable items. Brackets are 95% bootstrap CIs from 10,000 item-level resamples.

Statistic	Gem.–E1	Gem.–E2	Jev–E1	Jev–E2	E1–E2
Answer Accuracy					
QWK	0.793 [0.719,0.852]	0.763 [0.678,0.832]	0.809 [0.736,0.868]	0.736 [0.646,0.812]	0.812 [0.739,0.872]
Spearman	0.800 [0.727,0.858]	0.772 [0.689,0.840]	0.812 [0.739,0.874]	0.743 [0.655,0.817]	0.811 [0.741,0.871]
Exact agreement	0.580 [0.510,0.650]	0.560 [0.490,0.630]	0.665 [0.600,0.730]	0.590 [0.520,0.660]	0.650 [0.585,0.715]
Within-one agreement	0.900	0.880	0.870	0.840	0.885
D. Incomplete					
Cohen κ	0.511 [0.386,0.627]	0.468 [0.353,0.577]	0.227 [0.125,0.327]	0.227 [0.099,0.349]	0.574 [0.458,0.680]
F1	0.791 [0.727,0.846]	0.729 [0.656,0.794]	0.456 [0.354,0.551]	0.439 [0.328,0.547]	0.773 [0.701,0.834]

and confidence-based triage of newly drafted questions from recent open-access papers. LLM verification can be used as one verifier alongside deterministic checks rather than as the sole quality gate; experts can then focus on flagged or high-complexity items and random quality-control samples. This preserves expert oversight while making periodic benchmark refreshes more practical.

D JUDGE VALIDATION

The submitted benchmark used Gemini-2.5-Flash as the automatic judge. We validate it against two independent experts on the same 200 usable items and additionally rerun the validation with Jev-1.13 (Almeida, 2026), a structured decision model. The sample is stratified by scenario and Answer Accuracy level, with all eight domains and seven retrieval configurations represented. Table 8 reports the complete agreement statistics used in the revised analysis. Jev uses the same Answer Accuracy rubric and a direct port of the SPECTRA D. Incomplete diagnostic prompt. We retain Gemini as the primary judge for all benchmark results because the two judges are comparable on Answer Accuracy, whereas Gemini agrees substantially better with experts on D. Incomplete. Jev is therefore used as an independent validation check rather than as a second score averaged into the main experiments.

D.1 JEV-1.13 AS AN ALTERNATIVE JUDGE

We re-ran the judge validation with Jev-1.13 (TypeSafe), a decision model that does not generate free-form text. It takes a *state*, meaning the material to be judged, and typed questions. For each question it returns a probability distribution over the supplied options: a *Score* for ordered levels or a *Noul*, the probability that a yes/no statement holds. Because the answer is always one of the supplied options, no output parsing is needed and no parse failures can occur. Jev scored the same 200 usable items as the experts and the Gemini judge and was likewise blind to configuration, domain, and generator.

Run configuration. Table 9 summarizes the two automatic judges used in the validation study. The OpenRouter endpoint for Jev exposes no sampling parameters, so, unlike for the Gemini judge, no temperature or seed could be set. To check stability, we re-queried the first 40 Answer Accuracy items. The assigned level was identical for 39 of 40, and the expected level moved by 0.055 on average (maximum 0.30).

Answer Accuracy: differences from the Gemini prompt. Jev receives the same three inputs as the Gemini judge, represented as structured state fields `question`, `reference_answer`, and `model_response`. The rubric is carried over verbatim: the five evaluation criteria form the question instructions and the five scale descriptions (0 INCORRECT to 4 CORRECT) form the Score levels. Three implementation differences are forced by the model type: (i) the role sentence names

Table 9: Automatic judge configurations used in the validation study. Both judges use the same five-level Answer Accuracy rubric and the SPECTRA D. Incomplete diagnostic definition.

Setting	Gemini-2.5-Flash	Jev-1.13
Role	Primary benchmark judge	Alternative validation judge
Provider / interface	OpenRouter; chat-completions interface	TypeSafe via OpenRouter; native System One interface
Sampling	Temperature 0; seed 42	No sampling parameters exposed
Answer Accuracy	Five ordered levels (0–4); parsed score	Same five levels; most-probable Score level
D. Incomplete	SPECTRA `G_INCOMPLETE` binary prompt	Direct port of the same prompt; Noul probability > 0.5
Output	Free-form analysis followed by score / binary label	Typed probability distribution; no output parsing

the state fields; (ii) the instruction to write a brief analysis before scoring is dropped, so Jev produces no reasoning text; and (iii) the `Score: N` output format and parsing rule are dropped. The same request also contained a Choice question with the five levels as unordered options as a sensitivity check. That variant gave QWK 0.806 with E1 and 0.747 with E2, close to the Score variant reported in Table 8.

D. Incomplete: differences from the Gemini prompt. The SPECTRA diagnostic template in Appendix L.4 was ported with `failure_code` fixed to `G_INCOMPLETE`. The role text, the G_INCOMPLETE description, and decision rule 1 (assess only this condition) were kept verbatim. The four inputs (user query, golden answer, retrieved contexts, and model response) become state fields. Rules 2 and 3 use "answer yes" and "answer no" instead of "respond with 1" and "respond with 0", while rule 4, which requires a single-digit reply, is dropped because a Noul returns a probability. Retrieved contexts were rebuilt from the pipeline result files in the annotation-workbook format: numbered chunks with DOI and title. They are identical to what the experts saw for 194 items and complete for the 6 items whose workbook cells had been truncated at the spreadsheet cell limit.

At the fixed 0.5 threshold, Jev flagged 20.5% of answers as incomplete, compared with 54.0% for E1, 41.2% for E2, and 61.0% for Gemini. Most of its lower agreement comes from this under-flagging. Its probabilities nevertheless rank items in line with the experts (AUC 0.827 against E1 and 0.787 against E2); we do not tune a threshold on these same evaluation items. We therefore retain Gemini-2.5-Flash as the primary benchmark judge: Jev provides comparable independent validation for Answer Accuracy, but substantially weaker agreement on D. Incomplete under the matched SPECTRA diagnostic prompt.

E Evaluation Framework

E.1 RAG system and configurations

We evaluate SPECTRA using a modular RAG pipeline that processes the 333-article corpus into chunked text and figure indices for hybrid dense+sparse retrieval, neural reranking, and LLM generation (full architecture in Appendix E.3). To study the effect of pipeline choices, we vary three retrieval strategies (*dense*, *sparse*, *hybrid*) and two rerankers (`jinaai/jina-reranker-m0` (Wang et al., 2025) and `zeroentropy/zerank-1-reranker` (Pipitone et al., 2025)), plus a hybrid configuration with no reranker, yielding 7 RAG configurations. The same generation backbone (the 4-bit quantized `Mistral-Small-3.2-24B-Instruct-2506` checkpoint) and system prompt are used across configurations to isolate the effect of retrieval choices. We additionally evaluate 3 standalone LLMs without retrieval (GPT-5.4 with reasoning, GPT-5.4 no reasoning, Qwen-3.6-27B with reasoning) as baselines. The LLMs are evaluated zero-shot: they receive each question with a generic chemistry-assistant system prompt (Appendix L) but no retrieved context, no article text, and no figures, and must answer entirely from parametric knowledge.

The benchmark is not tied to this specific pipeline; any generator can be evaluated on answer-level metrics, while systems that expose retrieved contexts and source identifiers can be evaluated on the full retrieval, generation, and diagnostic suite. The evaluation framework supports custom retrieval and generation metrics alongside standard implementations from RAGAS (Es et al., 2024) and DeepEval (Confident AI, 2024).

E.2 METRICS

Metrics are organized into three groups that collectively support diagnosis of where errors originate. *Retrieval metrics* (deterministic) measure whether expected source articles and chunks were retrieved and how highly they were ranked, including DOI-level recall and MRR and chunk-level semantic similarity to ground-truth quotes. *Generation metrics* include both deterministic measures (numeric recall, scientific-fact recall via NLI entailment) and LLM-judged scores (answer accuracy on a 5-point scale, faithfulness to retrieved context, context relevance and recall, answer relevancy). *Diagnostic metrics* (LLM-judged) are binary flags for specific failure modes: hallucination without context, retrieval of contradicting sources, scientific errors, incomplete or imprecise answers, unjustified refusal, and inappropriate clarification requests. Combined, these allow attribution of any observed accuracy drop to a specific pipeline stage. Per-metric definitions are provided in Appendix E.4; LLM-as-a-judge prompts are in Appendix L.

E.3 RAG SYSTEM ARCHITECTURE AND INFERENCE DETAILS

Document ingestion. Articles enter the pipeline as PDF files referenced by a CSV manifest containing DOI, title, publisher, and publication year. PDF-to-Markdown conversion is performed by a Marker PDF server, which simultaneously extracts embedded images. The Markdown output passes through (i) a rule-based paragraph filtrator that removes non-scientific sections (preamble, references, acknowledgements, author biographies, ethics statements, funding disclosures, supplementary), (ii) a dedicated table grabber that isolates chemical data tables together with their captions, and (iii) a Markdown cleaner that normalizes whitespace and symbol encoding. A hash-based change-detection mechanism enables incremental reprocessing on subsequent ingestion runs.

Chunking. Cleaned text is segmented into fixed-size overlapping chunks of 300 tokens with a 75-token overlap, while respecting sentence and paragraph boundaries. The later chunk-size ablation in Appendix I varies this token-based chunk size explicitly.

Indexing and storage. Text chunks are encoded with JinaEmbV3 (1024-dim) and stored in PostgreSQL extended with pgvector. A Hierarchical Navigable Small World (HNSW) index supports approximate KNN retrieval. The same chunks are pushed to a Meilisearch full-text index for lexical retrieval. Extracted figures are encoded with OpenCLIP ViT-B/32 (512-dim) and stored in a parallel image table with its own HNSW index. Original PDFs and figure images are persisted in an S3-compatible MinIO object store.

Retrieval. The text-retrieval budget is 20 candidates per question. Dense-only configurations retrieve the top 20 chunks with JinaEmbV3 (Sturua et al., 2024) and HNSW (Malkov & Yashunin, 2018); sparse-only configurations retrieve the top 20 chunks with Meilisearch. Hybrid configurations retrieve the top 10 candidates from each branch, merge the two lists by score averaging (treating an absent branch score as zero), and deduplicate them, yielding up to 20 candidates before reranking. In parallel, the query is encoded by the OpenCLIP (Ilharco et al., 2021) text encoder and a separate image KNN search retrieves the top-$k = 5$ most relevant scientific figures.

Reranking. We test `jinaai/jina-reranker-m0` (Wang et al., 2025) and `zeroentropy/zerank-1-reranker` (Pipitone et al., 2025) on the same retrieved text candidates, together with a no-reranker hybrid baseline. The models were selected as contemporary neural rerankers with publicly documented checkpoints, while keeping the retriever inputs and candidate budgets fixed; this is a controlled reranking comparison rather than an exhaustive reranker benchmark. Reranked chunks are passed to the LLM without further truncation.

Generation. The 4-bit quantized `Mistral-Small-3.2-24B-Instruct-2506` checkpoint served via Ollama is used as the generation backbone across all original RAG configurations, with the same system prompt across runs. This is a vision-language checkpoint: figures retrieved from the separate OpenCLIP index are passed as pixel inputs together with text metadata containing the source, caption or figure reference, and URL. The prompt establishes a chemistry-research-assistant role and enforces strict grounding: queries beyond basic chemistry knowledge trigger a single retrieval call; if retrieval returns no relevant context, the model is instructed to reply "I did not find relevant information in the knowledge base" rather than answer from parametric knowledge; for partial retrieval the model answers within the returned content and explicitly states what is unknown. The full system prompt is reproduced in Appendix L. Standalone LLM baselines (GPT-5.4, GPT-5.4 no reasoning, Qwen-3.6-27B) are queried via OpenRouter with their respective default decoding parameters. They receive each question with a chemistry-assistant system prompt (Appendix L) but no retrieved context or attached files, so all answers come from parametric knowledge alone.

Multilingual handling. Non-English queries are auto-detected and translated to English before retrieval to maintain alignment with the embedded corpus. The original query language is preserved for the user-facing response.

Evaluation backend. Gemini-2.5-flash served via OpenRouter is used as the primary LLM judge for all reported benchmark runs, with seed 42, temperature 0, and 4024 max output tokens. We retain this backend after the human and alternative-judge validation in Appendix D; Jev-1.13 is used only as an independent validation check, not to rescore the full benchmark. The same Gemini backend is used for RAGAS and DeepEval implementations to ensure cross-framework comparability. Full prompt templates for each metric and each diagnostic failure mode are provided in Appendix L.

Deployment. The system is fully containerised via Docker Compose. Components run as isolated services: a Python/FastAPI application server (port 8080) hosting the RAG pipeline and SSE streaming endpoint; PostgreSQL 15 with pgvector (port 5436); Meilisearch (port 7700); MinIO (port 9000); and the Marker PDF conversion server (port 8000). Generation is served either by a local Ollama instance (open-weight models) or by OpenRouter for hosted LLMs.

Hardware. All experiments were conducted on a single machine equipped with one GPU (24 GB VRAM), 32 GB system RAM, an 8-core CPU, and 100 GB of SSD storage. The GPU runs the local generation model (Mistral-Small-3.2 4-bit quantized served via Ollama, $\sim$14 GB VRAM at 4-bit quantization), while embedding and reranking calls are dispatched to external APIs and do not contribute to local GPU usage. The remaining services (PostgreSQL with pgvector, Meilisearch, MinIO, the Marker PDF server, and the application server) run as Docker containers on CPU and consume roughly 6 GB of system RAM in aggregate. A full evaluation run on one of the 8 domains takes approximately 5,600 seconds on average, of which the majority is consumed by LLM-judged metrics; deterministic retrieval and generation metrics complete within seconds.

E.4 METRIC DETAILS

The benchmark framework computes a larger set of metrics than is reported in the main paper. This appendix documents each reported metric and explains the exclusions.

Reported retrieval metrics (deterministic). *Golden DOI Recall* is the fraction of expected source articles whose chunks appear in the retrieved context, treating each expected DOI as a binary hit. *Golden DOI MRR* is the reciprocal rank of the first chunk from an expected DOI, capturing whether the right document is at the top of the list rather than buried; a high recall with a low MRR indicates that retrieval finds the source but the reranker fails to surface it. *Source Support Max Similarity* is the maximum cosine similarity (using JinaEmbV3) between the annotated gold evidence and any retrieved chunk. It provides an evidence-level localization signal complementary to document-level DOI Recall: a low value can reflect failure to retrieve the answer-bearing passage or chunking that separates the relevant evidence.

Reported generation metrics. *Numeric Recall* (deterministic) extracts every numeric value from the reference answer and reports the fraction recovered in the model output, ignoring surrounding

text; it returns null for non-numeric questions to avoid biasing the average. *Scientific Fact Recall* (deterministic) decomposes the reference answer into atomic claims and uses a local DeBERTa-v3-large NLI model to score each claim's entailment by the model output, providing a strict factual measure that is insensitive to paraphrase but penalizes numeric or unit changes. *Answer Accuracy* (LLM-judged, Gemini-2.5-flash) grades the response on a 0–4 scale that explicitly weighs unit and chemical-name correctness, normalized to [0,1]. *Context Relevance* is computed per retrieved chunk as a YES/NO judgment of topical relevance and aggregated as the fraction of relevant chunks; it is the LLM-judged precision of retrieval. *Context Recall* is two-stage: the judge decomposes the reference answer into atomic facts and checks each one against the retrieved context, measuring whether retrieval surfaced enough information for a complete answer. *Faithfulness* mirrors this procedure but applied to the model's own response, flagging facts that are not supported by retrieved context (i.e., parametric-knowledge leakage); after the human validation reported in Appendix D, we retain it only as an exploratory output and do not use it for headline claims. *Answer Relevancy* grades on-topic-ness on a 0–4 scale ignoring factual correctness, distinguishing models that evade questions from those that answer them incorrectly.

Reported diagnostic metrics. D. Incomplete is retained in the headline analyses because it shows moderate judge–expert agreement (Appendix D); D. Scientific Error remains part of the released raw outputs but is not used for headline comparative claims after validation. Each diagnostic is a binary flag returned by a dedicated single-condition judge (template and failure-mode catalogue in Appendix L). *D_NO_SOURCE* fires when retrieval returns nothing yet the model produces a factual answer (pure hallucination). *D_WRONG_SOURCE* fires when retrieved context directly contradicts the reference, signalling a false positive at retrieval. *D_SCIENTIFIC_ERROR* fires on numeric or formula contradictions in the response. *D_INCOMPLETE* fires when the model omits facts that were available in the retrieved context. *D_IMPRECISE* fires on factual distortions short of outright contradiction. *D_REFUSE* fires on explicit refusals. *D_CLARIFY* fires when the model asks a follow-up question instead of answering. The diagnostics are designed to be summed for per-domain, per-scenario, or per-modification breakdowns, attributing accuracy drops to specific pipeline stages.

Selection rationale. The metrics in Figure 3 were selected on three criteria: (i) discriminative power across scenarios and modifications (Token-F1, for example, varied only between 0.05 and 0.18 with no clear pattern), (ii) non-redundancy with other reported metrics, and (iii) interpretability for pipeline-level diagnosis.

Excluded metrics. The framework also computes the following, which we omit from the main paper. Lexical overlap measures *Token Recall*, *Token-F1*, and *ROUGE-L Recall* have low discriminative power on open-ended chemistry answers and double-count what *Scientific Fact Recall* captures more strictly. *BERTScore Recall* compresses scores into a narrow band (typically 0.24–0.44) that obscures real differences, and the same factual signal is carried by *Scientific Fact Recall*. *Cosine Similarity* between full reference and generated answers is too coarse to separate factual from stylistic differences. *Source Support Hit*, the binarised version of *Source Support Max Similarity*, is dominated by its continuous counterpart. The DeepEval implementations of faithfulness and answer relevancy stayed near 0.92–0.97 across all conditions in our runs, providing no usable signal. The RAGAS implementation of *Answer Correctness* correlates strongly with our custom *Answer Accuracy* while being substantially slower; we report it in Figure 3 for cross-framework comparability but base our analysis on the custom implementation. *Image Source Hit*, which checks whether the correct figure was retrieved for multimodal questions, has sparse coverage (only multimodal-image cells) and remains under active development.

Computational cost. Deterministic retrieval and generation metrics compute in milliseconds per question; the NLI-based *Scientific Fact Recall* runs locally on a small DeBERTa model and adds a few hundred milliseconds. LLM-judged metrics require 1–3 seconds per metric per question for binary diagnostic judges and 6–9 seconds for multi-stage metrics (*Answer Accuracy*, *Faithfulness*, *Context Relevance*, *Context Recall*). Per-metric prompts are listed in Appendix L

F CROSS-DOMAIN EVALUATION: FULL RESULTS

Figure 3 reports mean metric values across all 8 domains, 6 scenarios, and 3 modification types for the hybrid RAG configuration with the Jina reranker. We highlight the patterns most relevant to benchmark interpretation.



Figure 3: Cross-domain benchmark behavior for the Jina·Hybrid configuration. Each cell shows the mean metric value aggregated by domain, scenario, or modification type. Generation/retrieval metrics use the score scale where higher is better; diagnostic rows report failure rates where lower is better. Domain abbreviations: Bz = Benzimidazole Antibiotics, CC = Chelate Metal Complexes, OP = Ophthalmic Formulations, Ox = Oxazolidinone Antibiotics, Sm = Supramolecular Drug Systems, NA = Nanoparticle Antibacterials, AS = Antibiotic-NP Synergy, Nz = Nanozymes. Scenario labels S1–S6 follow Section 3.2; Base, Mul, Neg, and Con denote unmodified, multimodal, negative, and constrained questions.

Domain effects. Answer Accuracy spans a 0.23 range across domains (0.41 on Benzimidazole Antibiotics to 0.64 on Chelate Metal Complexes), compared to a 0.24 range across scenarios and 0.28 range across modifications. The two highest-accuracy domains, Chelate Metal Complexes (0.64) and Nanozymes (0.60), are characterized by structured numerical answers (stability constants, kinetic parameters), which are easier to extract and verify. The weakest domain, Benzimidazole Antibiotics (0.41), contains many ground-truth answers consisting of MIC values reported across multiple bacterial strains; the model retrieves these only partially, as reflected by elevated D. Refusal (0.24, the highest among domains). Numeric Recall reaches 0.63 on Nanozymes but drops to 0.42–0.43 on Antibiotic-NP Synergy and Nanoparticle Antibacterials, where quantitative values are more often embedded in tables than in free text.

Scenario effects. Answer Accuracy varies modestly across scenarios (0.43–0.67), but D. Incomplete and D. Imprecise rise monotonically with scenario complexity: from 0.38 and 0.44 on single-hop queries (S1) to 0.89 and 0.90 on aggregation queries (S5) respectively. This confirms that the headline accuracy metric understates the actual failure rate on complex questions. Numeric Recall on conditional queries (S3) reaches 0.76, the highest of any scenario, reflecting that conditional queries

reward precise number extraction; on multi-hop queries (S2) it drops to 0.38, where numerical fidelity is lost along the reasoning chain.

Modification effects. Multimodal questions cause sharp drops in Numeric Recall ($0.62 \rightarrow 0.34$) and Scientific Fact Recall ($0.62 \rightarrow 0.38$) relative to unmodified questions, reflecting limited ability to extract values from tables and figures. Notably, DOI Recall remains high across all modifications including multimodal (0.78–0.88), indicating that source-article discovery remains strong. Because DOI Recall is document-level, however, this alone does not establish that the answer-bearing chunk was retrieved; evidence localization is measured separately with Source Support Max Similarity. D. Refusal is concentrated on multimodal queries (0.19) and multi-hop scenarios (S2, 0.20), suggesting cautious abstention when evidence is non-textual or distributed. Constrained queries are the most benign modification, with most retained metrics changing by less than 0.05 from baseline.

Cross-cutting observation. The retained metrics show complementary failure patterns across axes: multimodal and multi-hop questions more often produce abstention, while aggregation questions are dominated by incompleteness. These patterns are examined further in the configuration study in Section 4.1.

G Examples of Aggregation Query (S5)

We illustrate the S5 scenario (Table 3) with two representative aggregation questions from the Chelate Metal Complexes domain. In both cases, the reference answer is a curated list of entities or properties collected across multiple sources, and the standalone LLM outperforms RAG by recalling the canonical list from parametric memory.

Example 1. Property aggregation for a well-studied isotope.

> *Which core physical or practical properties can be considered as reasons for the dominant diagnostic role of ^{99m}Tc?*
>
> Golden answer: short half-life ($\sim$6 h), $\sim$140 keV gamma emission, ^{99}Mo/^{99m}Tc generator availability, kit-based preparation, and low patient dose.

`jina_hybrid` retrieves chunks discussing ^{99m}Tc applications but does not surface all five properties together (accuracy = 0.0). GPT-5.4 with reasoning produces the full canonical list (accuracy = 1.0). The list is well represented across nuclear medicine textbooks and likely encoded in pretraining; retrieval surfaces fragments but cannot assemble the complete set from the top-k chunks.

Example 2. Approved radiopharmaceutical inventory.

> *Which somatostatin receptor-targeting radiopharmaceuticals have approval, orphan designation, or registration for neuroendocrine tumour imaging or therapy up to 2025?*
>
> Golden answer: [^{111}In]In-pentetreotide, [^{64}Cu]Cu-DOTA-TATE, [^{68}Ga]Ga-DOTA-TATE, [^{68}Ga]Ga-DOTA-TOC, [^{99m}Tc]Tc-HYNIC-TOC, [^{177}Lu]Lu-DOTA-TATE.

`jina_hybrid` recovers part of the list (accuracy = 0.5); GPT-5.4 with reasoning returns the full set (accuracy = 1.0). Aggregation queries of this type favour parametric models when the answer is a stable, frequently cited inventory: the LLM may benefit from pretraining, while RAG must reconstruct it chunk-by-chunk from the corpus.

Summary. Both examples expose the same underlying limitation: aggregation queries reward completeness of recall over a stable set, and RAG retrieval surfaces top-k chunks rather than exhaustively scanning the corpus. When the answer set is well represented in the LLM pretraining data, parametric memory provides an inventory-like advantage that retrieval-based systems cannot match without dedicated coverage mechanisms such as iterative retrieval or set-based re-querying.

Table 10: Multi-turn performance on Chelate Metal Complexes. Δ/turn is the mean consecutive turn change; higher is better except for D. Incomplete.

System	Accuracy ↑		DOI Recall ↑		Source Support ↑		D. Incomplete ↓	
	Mean	Δ/turn	Mean	Δ/turn	Mean	Δ/turn	Mean	Δ/turn
Jina · Hybrid	0.64	−0.09	0.87	+0.02	0.74	−0.03	0.39	−0.02
Jina · Dense	0.51	−0.14	0.80	−0.02	0.73	−0.04	0.43	+0.00
Jina · Sparse	0.55	+0.00	0.68	−0.02	0.68	−0.01	0.51	−0.02
ZeRank · Hybrid	0.64	−0.08	0.86	+0.00	0.74	−0.03	0.36	+0.07
ZeRank · Dense	0.53	−0.12	0.83	+0.00	**0.75**	−0.05	0.42	+0.08
ZeRank · Sparse	0.53	−0.06	0.65	−0.06	0.68	−0.01	0.49	−0.01
W/o · Hybrid	**0.68**	−0.09	**0.90**	+0.02	**0.75**	−0.03	**0.32**	+0.05
GPT-5.4	0.58	−0.03	—	—	—	—	—	—
GPT-5.4 no reasoning	0.51	−0.08	—	—	—	—	—	—
Qwen-3.6-27B	0.40	−0.08	—	—	—	—	—	—

H Multi-turn Scenario. Detailed analysis

H.1 Per-metric results for the multi-turn scenario

Table 10 reports mean values and per-turn trends (Δ/turn) for the validated headline metrics available across the S6 scenario. RAG rows and the two GPT-5.4 baselines average seeds 42 and 43.

DOI Recall stays broadly stable across turns for the hybrid configurations, whereas Source Support declines by about 0.03 per turn and Answer Accuracy declines more strongly. Thus, the multi-turn degradation is not caused simply by losing the source article, but evidence localization does weaken with dialogue depth. D. Incomplete varies less systematically and partly reflects question breadth: later turns often ask for a single value or term, whereas earlier turns can require broader answers spanning several facts. The combined pattern separates document discovery, evidence localization, and answer quality.

H.2 Case study

We focus on three turns that expose complementary failure modes: a terminology turn (Turn 2) where dense vs sparse retrieval differ, a numeric turn (Turn 4) where standalone LLMs produce plausible but ungrounded values, and a competing-evidence failure on an implicit-reference turn from a different session.

Turn 2. Terminology precision.

> *Which chelator does this radiopharmaceutical contain?* Golden answer: **HBED-CC**

Table 11 summarizes generated answers across configurations. Dense-only retrievers (`jina_dense`, `zerank_dense`) return *HBED* without the `-CC` suffix, because the dense embedding averages over semantic similarity and retrieves passages where the full acronym is not the salient token. Configurations with a sparse component (`jina_sparse`, `jina_hybrid`, `zerank_sparse`) correctly recover *HBED-CC* through exact lexical matching. GPT-5.4 also returns *HBED-CC*, but this reflects parametric memorisation of a widely-cited clinical tracer rather than document grounding — the same model fails on less prominent compounds within the same benchmark.

Turn 4. Numeric fact retrieval.

> *What positivity rate of this radiopharmaceutical is reported for PET-CT in biochemical recurrence?* Golden answer: **64.2%**

This turn requires locating a specific percentage from a specific publication. Table 12 shows the results. Sparse-based retrievers (`jina_sparse`, `zerank_sparse`) correctly retrieve *64.2%* through

Table 11: Generated answers for Turn 2 (chelator name) across configurations.

Configuration	Answer	Accuracy
jina_hybrid	HBED-CC	1.0
hybrid_no_rerank	HBED-CC	1.0
jina_sparse	HBED-CC	1.0
zerank_sparse	HBED-CC	1.0
zerank_hybrid	HBED-CC	1.0
jina_dense	HBED (no `-CC`)	0.5
zerank_dense	HBED (no `-CC`)	0.5
GPT-5.4	HBED-CC	1.0
GPT-5.4 no reasoning	HBED-CC	1.0
Qwen-3.6-27B	DOTA	0.0

token-level matching of the exact figure. Hybrid and dense-only systems retrieve a semantically similar passage reporting *68.8%* from a different study within the same corpus — a plausible but incorrect answer. All LLM baselines fail in a characteristic way: GPT-5.4 with reasoning produces ≈*63.9%*, close but unsourced; GPT-5.4 without reasoning reports ≈*74%*; Qwen-3.6-27B provides a PSA-stratified detection-rate table spanning 10–95% without a single reported figure. This pattern confirms that parametric memory generates contextually plausible but document-ungrounded responses for domain-specific numeric facts.

Table 12: Generated answers for Turn 4 (positivity rate) across configurations.

Configuration	Answer	Accuracy
jina_sparse	64.2%	1.0
zerank_sparse	64.2%	1.0
zerank_dense	58–74% (range)	0.5
zerank_hybrid	meta-analysis, no figure	0.25
jina_hybrid	68.8% (different study)	0.0
hybrid_no_rerank	68.8% (different study)	0.0
jina_dense	68.8% (different study)	0.0
GPT-5.4	≈63.9%	0.75
GPT-5.4 no reasoning	≈74%	0.25
Qwen-3.6-27B	PSA-stratified table	0.25

Turn 2 (different session). Competing evidence under an implicit reference.

> *Session 4, Turn 1: "Which technetium precursor is produced in one step from generator pertechnetate for His-tag radiolabelling?"*
> Golden answer (Turn 1): ***fac*-$[^{99m}Tc(CO)_3(H_2O)_3]^+$**
> *Turn 2: "Which amino-acid residue is favored for binding **this core**?"*
> Golden answer (Turn 2): **Histidine**

For Jina·Hybrid, Turn 1 is answered correctly in both seeds (Accuracy = 1.0, DOI Recall = 1.0). On Turn 2, the model answers *lysine* rather than histidine (Accuracy = 0.0), while the expected source article is still present in the retrieved set (DOI Recall = 1.0; Source Support ≈ 0.67). The retrieved context also contains passages about lysine-based chelation, and the generated answer follows this competing evidence. The trace therefore does not support attributing this error to loss of the source document or, by itself, to a query-reformulation failure. It instead shows that document-level retrieval success can coexist with incorrect selection or use of evidence when an implicit reference must be resolved in a multi-turn context. The same Jina·Hybrid run returns to Accuracy = 1.0 and DOI Recall = 1.0 on Turn 3.

Summary. The three turns illustrate complementary failure modes spanning the RAG pipeline. At Turn 2, dense-only retrieval recovers semantic context but misses the exact chemical token (*HBED* vs

HBED-CC), while sparse retrieval recovers it through lexical matching. At Turn 4, only sparse-based configurations retrieve the precise figure (64.2%) from the correct article, whereas hybrid and dense-only systems anchor their answers in a numerically plausible but document-incorrect passage (68.8%), and standalone LLMs produce values from parametric memory with no source attribution. The Turn 2 example from a different session exposes a third failure mode: for Jina·Hybrid the expected source remains retrieved (DOI Recall = 1.0), yet the answer follows competing retrieved evidence and is incorrect. This illustrates why document-level DOI Recall alone is insufficient and why evidence-localization and context-use signals are useful in multi-turn evaluation. The sparse-vs-hybrid gap on numeric facts (Turn 4) mirrors the pattern observed in single-turn Conditional Retrieval (S3, Table 3), where sparse-based configurations also lead on tasks requiring exact value recovery. Hybrid search without reranking achieves the highest mean multi-turn Accuracy in Table 10, but the case studies show that different turns can fail for different retrieval and generation reasons.

I Additional configuration and robustness analyses

The following controlled studies test whether the scenario-based protocol remains informative when the domain, retrieval parameters, backbone, or end-to-end RAG system changes. Unless stated otherwise, they use the same three 50-question samples from Chelate Metal Complexes, Nanozymes, and Oxazolidinone Antibiotics, with 10 questions from each of Scenarios 1–5 and one seed. The cross-domain and backbone studies use 300-token chunks, while the chunk-size study varies this setting over 300, 500, and 800 tokens. Identical stored Mistral/Jina-Hybrid outputs were judged in more than one experimental batch; repeated Gemini-2.5-Flash scoring differed by at most 0.01 in domain-level mean Answer Accuracy, so each table reports the scoring pass associated with that experiment.

I.1 Cross-domain configuration study

Table 13 reports Answer Accuracy, document-level DOI Recall, evidence-level Source Support, and D. Incomplete for all seven retrieval/reranking configurations on the three matched domain samples. The domain-level ordering changes across both aggregate and scenario-specific views. For example, Jina Sparse has the lowest aggregate Accuracy on Nanozymes (0.36) but the second-highest on Oxazolidinone Antibiotics (0.44). The scenario profile changes even more strongly: Aggregation is led by Jina Dense on Chelate (0.60), ZeRank Dense on Nanozymes (0.55), and Jina Sparse on Oxazolidinone Antibiotics (0.60). Multi-hop is led by Jina Hybrid on Chelate and Nanozymes, but by Jina Sparse on Oxazolidinone Antibiotics. Scenario difficulty also changes by domain: averaged across configurations, Aggregation is 0.28 on Nanozymes versus 0.50 on Oxazolidinone Antibiotics, while Direct Fact Retrieval is 0.76 on Chelate versus 0.40 on Oxazolidinone Antibiotics.

I.2 Chunk size and retrieval depth

Table 14 reports all four headline signals for the 3×3 token chunk-size/top-k grid. Increasing retrieval depth generally raises DOI Recall, but its effect on Accuracy and D. Incomplete depends on domain and chunk size. The largest Accuracy range is on Oxazolidinone Antibiotics (0.35–0.61), followed by Chelate Metal Complexes (0.52–0.69) and Nanozymes (0.51–0.61). The best Accuracy setting is 800/20 for Chelate and Oxazolidinone, while Nanozymes reaches 0.61 at both 300/20 and 800/10. These results show that retrieval depth and chunking affect source coverage and answer quality differently rather than defining one universally optimal setting.

I.3 Backbone and end-to-end system comparison

Table 15 compares the three SPECTRA backbones and adds RAGU-Naive as an end-to-end system comparison on the same matched questions. Within SPECTRA, Mistral-Small-3.2 and GPT-OSS-20B have nearly identical mean Accuracy (0.52) but different domain profiles: GPT-OSS-20B is strongest on Oxazolidinone Antibiotics (0.64), while Mistral is stronger on Chelate and Nanozymes. Across the three SPECTRA backbone runs, GPT-OSS-20B also has the lowest mean D. Incomplete (0.61), whereas Mistral has the highest Numeric Recall (0.53). RAGU-Naive (Komarov et al., 2026) uses the same GPT-OSS-20B answering backbone as the corresponding SPECTRA row but changes the retrieval stack, chunking, prompting, and embedding model (Qwen3-Embedding-8B (Zhang et al.,

Table 13: Cross-domain results for seven retrieval/reranking configurations on the same 50 scenario-stratified questions per domain (one seed). Source Support is shown as a percentage to retain resolution; other metrics use the 0–1 scale.

Configuration	Accuracy↑	DOI Recall↑	Source Support (%)↑	D. Incomplete↓
Chelate Metal Complexes				
Jina · Hybrid	0.56	0.80	66.82	0.64
Jina · Dense	0.56	**0.81**	67.69	0.69
Jina · Sparse	0.45	0.52	58.09	0.71
ZeRank · Hybrid	**0.57**	0.79	66.84	**0.62**
ZeRank · Dense	0.53	**0.81**	**67.77**	0.69
ZeRank · Sparse	0.50	0.53	58.57	0.69
Hybrid, no rerank	0.51	0.80	66.61	0.69
Nanozymes				
Jina · Hybrid	0.55	0.80	67.48	0.56
Jina · Dense	0.57	0.83	69.11	**0.52**
Jina · Sparse	0.36	0.54	60.36	0.62
ZeRank · Hybrid	0.53	0.80	67.31	0.63
ZeRank · Dense	**0.57**	**0.85**	**69.13**	0.56
ZeRank · Sparse	0.41	0.54	60.02	0.68
Hybrid, no rerank	0.53	0.79	67.81	0.63
Oxazolidinone Antibiotics				
Jina · Hybrid	**0.47**	0.80	69.71	0.74
Jina · Dense	0.43	**0.84**	69.84	0.74
Jina · Sparse	0.44	0.72	66.10	**0.70**
ZeRank · Hybrid	0.39	0.81	70.05	0.77
ZeRank · Dense	0.39	0.82	69.40	0.77
ZeRank · Sparse	0.42	0.72	66.35	0.72
Hybrid, no rerank	0.37	0.81	**70.05**	0.79

2025)); its 0.79 mean Accuracy therefore represents an end-to-end system difference rather than a generator-only effect.

I.4 Question-level uncertainty on the full Chelate study

The paired question bootstrap uses the 254 questions shared by all configurations and averages seeds 42 and 43 per item before resampling. The leading hybrid systems form a tight group: W/o Hybrid reaches 0.66 [0.61, 0.71], ZeRank Hybrid 0.65 [0.60, 0.70], and Jina Hybrid 0.64 [0.59, 0.69]. Their intervals overlap substantially, whereas the hybrid group remains clearly separated from the sparse-only configurations. This bootstrap captures question-sample uncertainty; it does not replace multi-seed uncertainty.

I.5 Parametric answerability

A question is counted as parametrically answerable when at least two of the three standalone models obtain Answer Accuracy ≥ 0.75, and as closed-book-hard when all three score ≤ 0.50. The analysis measures parametric answerability rather than article-level contamination: a correct closed-book response cannot establish whether a specific source paper was present in pretraining.

I.6 RAGU scenario profile

To inspect whether the cross-system gain is uniform across question types, Table 18 aggregates exactly matched questions across Chelate Metal Complexes, Nanozymes, and Oxazolidinone Antibiotics. Both systems use GPT-OSS-20B as the answering backbone and the same Answer Accuracy judge. The supplied RAGU result artifacts contain the generated answers and judge scores but do not retain retrieved contexts or source identifiers, so this comparison reports Answer Accuracy only and does not infer DOI Recall or Source Support for RAGU. RAGU-Naive (Komarov et al., 2026) improves

Table 14: Token chunk-size × top-k sensitivity for Jina-Hybrid with Mistral-Small-3.2 (50 questions per domain, one seed). Top-k is per dense/sparse branch; Source Support is shown as a percentage.

Chunk	top-k	Accuracy↑	DOI Recall↑	Source Support (%)↑	D. Incomplete↓
Chelate Metal Complexes					
300	5	0.57	0.69	64.18	0.58
300	10	0.56	0.80	66.82	0.64
300	20	0.67	**0.89**	**69.25**	**0.56**
500	5	0.52	0.75	62.83	0.73
500	10	0.58	0.81	65.43	0.67
500	20	0.64	0.87	66.62	0.58
800	5	0.55	0.71	60.59	0.67
800	10	0.62	0.82	63.35	0.64
800	20	**0.69**	0.87	64.91	0.58
Nanozymes					
300	5	0.51	0.68	65.16	0.63
300	10	0.55	0.80	67.48	0.56
300	20	**0.61**	0.86	**69.98**	0.50
500	5	0.53	0.70	62.79	0.54
500	10	0.57	0.82	66.29	0.54
500	20	0.59	0.90	68.40	**0.48**
800	5	0.58	0.72	60.89	0.60
800	10	**0.61**	0.80	62.55	0.60
800	20	0.61	**0.91**	65.42	0.54
Oxazolidinone Antibiotics					
300	5	0.35	0.67	67.83	0.79
300	10	0.47	0.80	69.71	0.74
300	20	0.53	0.92	**72.13**	0.66
500	5	0.36	0.69	65.02	0.79
500	10	0.42	0.83	67.52	0.74
500	20	0.59	**0.95**	69.72	**0.64**
800	5	0.44	0.72	64.39	0.81
800	10	0.56	0.86	66.89	0.68
800	20	**0.61**	0.93	67.76	0.66

Table 15: Answer Accuracy on the same three 50-question domain samples. The first three rows vary the SPECTRA backbone with Jina-Hybrid retrieval; RAGU-Naive uses GPT-OSS-20B with its own retrieval stack.

System / backbone	Chelate	Nanozymes	Oxazolidinone	Mean
SPECTRA / Mistral-Small-3.2	0.55	0.55	0.47	0.52
SPECTRA / GPT-OSS-20B	0.51	0.42	0.64	0.52
SPECTRA / Qwen3-8B	0.35	0.51	0.47	0.44
RAGU-Naive / GPT-OSS-20B	**0.85**	**0.78**	**0.73**	**0.79**

the pooled mean from 0.52 to 0.79; the largest gains occur on Conditional Retrieval (+0.42) and Multi-hop (+0.35), whereas Direct Fact Retrieval improves by +0.10. A paired question bootstrap over all 150 matched items (10,000 resamples, seed 42) gives a mean difference of 0.27 with a 95% percentile interval of [0.19, 0.34].

J Additional examples and judge spot-check

Note. The examples below preserve the raw metric outputs from the original evaluation artifact. Faithfulness and D. Scientific Error are shown only for traceability and are not used in the revised headline claims after the human validation in Appendix D.

Table 16: Full Chelate configuration study on 254 shared questions after averaging seeds 42 and 43 per item. CIs are paired question-bootstrap intervals with 10,000 resamples.

Configuration	Answer Accuracy	95% CI
Hybrid (no rerank)	0.66	[0.61, 0.71]
ZeRank · Hybrid	0.65	[0.60, 0.70]
Jina · Hybrid	0.64	[0.59, 0.69]
Jina · Dense	0.56	[0.51, 0.61]
ZeRank · Dense	0.56	[0.50, 0.61]
Jina · Sparse	0.48	[0.42, 0.53]
ZeRank · Sparse	0.48	[0.42, 0.53]

Table 17: Parametric-answerability analysis on 254 Chelate questions shared by all standalone and RAG systems.

Quantity	Result
Parametrically answerable ($\geq$2/3 standalone models ≥ 0.75)	40.2%
Closed-book-hard (all standalone models ≤ 0.50)	32.7% (83/254)
Jina-Hybrid Accuracy on closed-book-hard subset	0.52
Best standalone Accuracy on closed-book-hard subset	0.20
Aggregation: best standalone / Jina-Hybrid	0.87 / 0.43

We spot-checked LLM-judged metrics on 18 question–answer pairs from the Jina·Hybrid configuration on Chelate Metal Complexes, covering three correctness levels for each scenario. Each example preserves five metrics from the original evaluation artifact and includes a short expert note; Faithfulness and D. Scientific Error in this section are retained only for traceability and are not part of the revised headline metric set. The goal is not to validate every individual judge decision, but to show how judged diagnostics should be interpreted together with deterministic metrics such as DOI Recall, Numeric Recall, and NLI-based Scientific Fact Recall.

S1 · DIRECT FACT RETRIEVAL

Correct (Acc. 1.0 | DOI Rec. 1.0 | Faith. 0.75 | D.Inc. 0 | D.Sci. 0). *Q:* In the μ-CO bridged Tc(I) dimer containing two Cp* ligands and three bridging carbonyls, what Tc–Tc separation is given as reported by Abram et al.?
Golden: 2.413 Å.
Generated: The Tc–Tc separation . . . is 2.413 Å. This information is derived from the structure determination of compound **11**, which is formed by irradiation of compound **10b** with UV light.
Note: All metrics agree with the expert assessment. Faithfulness below 1.0 reflects the additional unsupported context about compound **11**.

Partial (Acc. 0.5 | DOI Rec. 1.0 | Faith. — | D.Inc. — | D.Sci. 1). *Q:* For the 1:1 Gd(III)/KEMPPr(3,4-HP)$_3$ system at $C_L = 3.2 \times 10^{-4}$ M in 50% w/w DMSO/water, at what pH do GdHL and GdL occur at equal relative formation?
Golden: pH 4.7–4.8.
Generated: From the image provided, we can see th[e species distribution curves]. . . [answer truncated; the model could not extract the value from the figure].
Note: Faithfulness and D.Incomplete are not applicable because the model produced no concrete answer. D. Sci. Error is questionable: an expert would likely treat this as a failed extraction from a figure rather than a scientific contradiction.

Incorrect (Acc. 0.0 | DOI Rec. 1.0 | Faith. 0.83 | D.Inc. 1 | D.Sci. 1). *Q:* Which compound serves as the carbon monoxide source for forming the $[^{99m}Tc(CO)_3]^+$ core from $^{99m}TcO_4^-$?
Golden: $Na_2[H_3BCOO]$.
Generated: The compound serving as the CO source is fac-$[^{99m}Tc(CO)_3(H_2O)_3]^+$.

Table 18: Matched cross-system Answer Accuracy by scenario across Chelate Metal Complexes, Nanozymes, and Oxazolidinone Antibiotics. Both systems use GPT-OSS-20B as the answer backbone.

System	Direct	Multi-hop	Conditional	Comparison	Aggregation	Overall
SPECTRA Jina · Hybrid	0.78	0.35	0.40	0.48	0.59	0.52
RAGU-Naive	**0.88**	**0.70**	**0.82**	**0.76**	**0.77**	**0.79**
Δ	+0.10	+0.35	+0.42	+0.28	+0.18	+0.27

Note: The model confuses the product with the reagent. DOI Recall is 1.0, so the diagnostic flags reflect a generation-side error.

S2 · MULTI-HOP

Correct (Acc. 1.0 | DOI Rec. 1.0 | Faith. 0.67 | D.Inc. 0 | D.Sci. 0). *Q:* What is the log P value of the CP04 complex of the radionuclide whose NOTA complex is reported in post-2020 literature to reach >95% radiochemical yield at room temperature, pH 3, within 10 min?
Golden: -2.68 ± 0.03.
Generated: The log P value ... is -2.68 ± 0.03 for ^{68}Ga-CP04.
Note: All metrics agree with the expert assessment.

Partial (Acc. 0.5 | DOI Rec. 1.0 | Faith. 0.20 | D.Inc. 0 | D.Sci. 0). *Q:* What percentage of transchelation after 7 days in 50 mM DTPA was observed for the ^{89}Zr complex of the chelator clinically developed as an Fe^{3+} chelator for iron overload?
Golden: 0.59.
Generated: The chelator is DFO. According to the information retrieved, 59% of the ^{89}Zr-DFO complex underwent transchelation after 7 days.
Note: The answer is penalized for a formal mismatch between fraction and percentage. D. Sci. Error remains 0, which is reasonable because the two forms are scientifically equivalent under a lenient reading.

Incorrect (Acc. 0.0 | DOI Rec. 1.0 | Faith. — | D.Inc. — | D.Sci. 1). *Q:* What is the 4–5 h SUV_{max} for [^{64}Cu]Cu-14 in the organ showing the second-highest uptake at 4 h in the biodistribution profile of [^{64}Cu][Cu(DO2A2S)]?
Golden: 0.72 ± 0.24.
Generated: The organ ... is the lungs. The 4–5 h SUV_{max} for [^{64}Cu]Cu-14 in the lungs is not explicitly provided.
Note: The first hop is incorrect, after which the model abstains rather than guessing a numeric value. D. Sci. Error fires on the wrong intermediate conclusion, but this is a borderline case: an expert may treat it as a multi-hop reasoning error followed by refusal rather than as a direct scientific contradiction.

S3 · CONDITIONAL QUERY

Correct (Acc. 1.0 | DOI Rec. 0.5 | Faith. 0.14 | D.Inc. 0 | D.Sci. 0). *Q:* What chelator is discussed in the *Journal of Cancer Research and Therapeutics* review as part of renal-cell-carcinoma xenograft pretargeting with G250×DTIn1?
Golden: DTPA.
Generated: The chelator ... is DTPA.
Note: The answer is correct despite low Faithfulness and DOI Recall. This illustrates why answer accuracy and grounding metrics are reported separately.

Partial (Acc. 0.75 | DOI Rec. 1.0 | Faith. 0.38 | D.Inc. 1 | D.Sci. 1). *Q:* What radionuclide, whose NOTA complex reaches >95% yield at room temperature pH 3 within 10 min, forms a CP04 complex with log $P = -2.68$?
Golden: ^{68}Ga.

Generated: The radionuclide is ^{68}Ga . . . [followed by a lengthy explanation that adds unsupported claims].
Note: The core answer is correct, but the explanation adds unsupported claims; the diagnostic flags reflect this.

Incorrect (Acc. 0.0 | DOI Rec. 0.5 | Faith. 0.5 | D.Inc. 1 | D.Sci. 0). *Q:* What chelator bound all added $^{68}Ga^{3+}$ at pH 3.5 but only 60–80% at pH 6.5, and whose peptide conjugates correspond to 0.11 μg cold mass for a 185-MBq dose?
Golden: TRAP.
Generated: The chelator is TRAP . . . [the explanation that follows confuses TRAP with a different compound].
Note: Accuracy is 0 because the justification contradicts the answer. D.Incomplete correctly signals the dropped cold-mass condition.

S4 · Comparison

Correct (Acc. 1.0 | DOI Rec. 1.0 | Faith. 0.88 | D.Inc. 0 | D.Sci. 0). *Q:* Which synthetic transformation required the longer reaction time: formation of (3-phthalimidylpropyl)triphenylphosphonium from bromopropyl phthalimide and triphenylphosphine, or hydrolysis of the triphenylphosphonium ethyl ester to the Boc-protected acid?
Golden: Formation of the phosphonium salt required longer (24 h vs. 2 h).
Generated: The formation . . . required 24 hours under reflux in acetonitrile. In contrast, the hydrolysis . . . required 2 hours.
Note: Both reaction times are correctly extracted and compared.

Partial (Acc. 0.5 | DOI Rec. 1.0 | Faith. 0.38 | D.Inc. 1 | D.Sci. 1). *Q:* Which conjugate gave the higher radiochemical yield: $(HE)_3$-ZHER3-DOTAGA after 15 min at 85 °C, or Ec1-DOTA after 25 min at 60 °C?
Golden: $(HE)_3$-ZHER3-DOTAGA ($81 \pm 16\%$ vs. $39 \pm 14\%$).
Generated: $(HE)_3$-ZHER3-DOTAGA gave higher yield . . . reported to be over 95% after purification.
Note: The comparison direction is correct, but the numeric value is wrong. D. Sci. Error correctly fires on the numeric contradiction.

Incorrect (Acc. 0.0 | DOI Rec. 0.5 | Faith. 0.60 | D.Inc. 1 | D.Sci. 1). *Q:* How did replacing DOTA with THP in the TATE system compare with replacing DOTA with DATA in the TOC system in terms of tumour-to-liver contrast?
Golden: Directions were opposite: THP lowered contrast ($27.2 \rightarrow 10.5$), DATA raised it ($3.1 \rightarrow 5.2$).
Generated: Both replacements resulted in improved tumour-to-liver contrast.
Note: The model misses the opposing directions; D.Incomplete and D. Sci. Error both fire correctly.

S5 · Aggregation

Correct (Acc. 1.0 | DOI Rec. 0.5 | Faith. 0.79 | D.Inc. 1 | D.Sci. 0). *Q:* Which structural elements, approval status, and clinical role define [^{68}Ga]Ga-PSMA-11 in recurrent prostate cancer imaging?
Golden: Glu-Urea-Lys inhibitor linked via Ahx to HBED-CC; first and only FDA-approved metal-based PSMA-i radiopharmaceutical for biochemical recurrence.
Generated: [^{68}Ga]Ga-PSMA-11 . . . is a radiopharmaceutical used for PET imaging . . . [correct but incomplete].
Note: Accuracy 1.0 with D.Incomplete 1 is possible because the model gives the main answer but omits structural and approval details. This supports reporting diagnostic flags alongside accuracy.

Partial (Acc. 0.5 | DOI Rec. 1.0 | Faith. 0.14 | D.Inc. 1 | D.Sci. 1). *Q:* Which chelator system offers the highest specific activity and lowest cold-mass burden in ^{68}Ga labeling studies published before 2020?
Golden: $TRAP(RGD)_3$; $\sim$5000 GBq/μmol, 0.11 μg cold mass per 185-MBq dose.
Generated: The chelator system is TRAP . . . [identifies the correct chelator class but gives wrong specific activity and omits the cold-mass figure].
Note: D.Incomplete and D. Sci. Error correctly capture the missing and incorrect quantitative details.

Incorrect (Acc. 0.0 | DOI Rec. 0.4 | Faith. 0.88 | D.Inc. 1 | D.Sci. 0). *Q:* Which core physical or practical properties explain the dominant diagnostic role of ^{99m}Tc?
Golden: Short half-life (∼6 h), 140 keV gamma, generator availability, kit-based preparation, low patient dose.
Generated: . . . short half-life, optimal nuclear properties . . . [generic answer not grounded in the corpus].
Note: DOI Recall is 0.4, and D.Incomplete fires because several key properties are missing. The failure is mainly due to incomplete retrieval coverage.

S6 · MULTI-TURN

Correct (Acc. 1.0 | DOI Rec. 1.0 | Faith. 0.67 | D.Inc. 0 | D.Sci. 0). *Q:* What log $D_{7.4}$ value was measured for the radioligand bearing this chelator?
Golden: -3.00.
Generated: The log $D_{7.4}$ value . . . for the radioligand bearing the NODA-GA chelator is -3.00.
Note: All metrics agree with the expert assessment.

Partial (Acc. 0.5 | DOI Rec. 1.0 | Faith. 0.70 | D.Inc. 0 | D.Sci. 1). *Q:* What coordination-sphere limitation is associated with this chelator for ^{89}Zr?
Golden: Two empty coordination sites.
Generated: The limitation . . . is related to the unsaturated coordination sphere of DFO . . . [correct concept but expressed at higher level than the reference].
Note: The model captures the right concept but does not specify the number of empty sites. D. Sci. Error reflects the stricter interpretation of the answer.

Incorrect (Acc. 0.0 | DOI Rec. 1.0 | Faith. 0.29 | D.Inc. 0 | D.Sci. 0). *Q:* What follow-up imaging strategy does the journal *Cancer Imaging* recommend to improve the detection of recurrence?
Golden: 43.4%.
Generated: The journal recommends hybrid imaging modalities, specifically [^{68}Ga]PSMA-11 PET-CT and PET-MRI . . .
Note: The reference answer is a numeric detection rate, while the question asks for a strategy. This is a benchmark-side labelling inconsistency rather than a clear model error; the diagnostic flags do not surface it.

SUMMARY OF THE SPOT-CHECK

This qualitative spot-check is retained as illustrative material; quantitative judge validity is established by the 200-item study in Appendix D. Fifteen of the 18 examples agree with expert reading, while the remaining three illustrate borderline behavior of fixed-prompt LLM judging and benchmark annotation. In S1 Partial, a failed figure extraction is marked as a scientific error despite the absence of a concrete numeric claim. In S2 Incorrect, D. Sci. Error fires on a wrong intermediate conclusion in a multi-hop question, although the model ultimately refuses to provide the requested numeric value. In S6 Incorrect, the benchmark item has a mismatch between question phrasing and the reference answer. These examples support the use of diagnostic metrics as scalable signals rather than as replacements for expert review. In the main analysis, we therefore read the judged metrics together with deterministic measures such as DOI Recall, Numeric Recall, and NLI-based Scientific Fact Recall.

The benchmark is intended for evaluation of literature-grounded RAG systems, not for laboratory, clinical, or regulatory decision-making; results should not be interpreted as certification that a system is scientifically reliable in deployment.

K QUESTION GENERATION PROMPTS

K.1 SCENARIO 1 PROMPT TEMPLATE

```
You are an expert in chemistry and scientific benchmark construction.
Create high-quality question--answer pairs for evaluating
retrieval-augmented generation (RAG) systems in chemistry.

Scenario: Direct Fact Retrieval (Single-hop).

Each question must require retrieving one specific fact from the
attached article: a value, term, entity, phrase, experimental
condition, observation, or briefly paraphrased finding. The source
must be unambiguously identifiable, even if the same information may
also appear elsewhere.

Example:
Q: What experimental value of rabbit corneal permeability is commonly
used as a reference for acetate forms of glucocorticoids, such as
dexamethasone, when optimizing ophthalmic formulations?
A: logP = -4.43.

Create 10 question-answer pairs for the attached article.

Requirements:
1. Questions must be standalone, clear, and unambiguous.
2. Do not mention the article, title, authors, journal, DOI, tables,
figures, or phrases such as "in the article", "as reported", or
"according to the study".
3. Questions must require the attached article and must not be
answerable from general chemical knowledge or a simple web search
alone.
4. Questions should be suitable for researchers in chemistry,
especially coordination chemistry, chelate complexes, ligands, metal
complexes, materials, catalysis, or related areas, depending on the
article.
5. Include enough chemical or experimental context to identify the
relevant entity, system, property, method, or condition
unambiguously.
6. Avoid overly broad, opinion-based, speculative, or purely
interpretive questions.
7. Avoid questions that are so narrow that they merely point to a
single experiment without testing meaningful retrieval.
8. Avoid ambiguous rankings such as "the best", "the most effective",
or "the most stable" unless the criterion is explicitly stated and
directly supported.
9. Do not present article-specific findings as universal chemical
truths; tie them to the relevant system, compound, method, or
conditions.
10. Make the 10 questions diverse in topic and answer type.

For each item, provide the question, answer, source evidence from the
article, and DOI. Source evidence may be a short quote, table,
figure, scheme, or section.

Return a table with the columns:
Question | Answer | Source evidence | DOI.
```

K.2 SCENARIO 2 PROMPT TEMPLATE

First, article pairs sharing a chemically meaningful key were identified to define possible multi-hop links.

```
You are an expert in chemistry and scientific benchmark construction.
I have provided a set of articles. Identify pairs of articles
suitable for multi-hop question generation.

For each suitable pair, find a shared chemically meaningful key. The
key may be a compound, metal complex, ligand, catalyst, solvent,
material, reaction type, property, analytical method, experimental
condition, application, or another shared concept.

For each pair, provide the two DOIs, the shared key, evidence that
the key appears in each article, and a brief explanation of how the
articles can be linked in a multi-hop question.

Return a table with the columns:
Article 1 DOI | Article 2 DOI | Shared key | Evidence from article 1
| Evidence from article 2 | Possible bridge.
```

The identified links were then used to generate bridge questions in which one article identifies the shared key and the other provides the final answer.

```
You are an expert in chemistry and scientific benchmark construction.
Create high-quality question--answer pairs for evaluating
retrieval-augmented generation (RAG) systems in chemistry.

Scenario: Sequential Retrieval (Multi-hop / Bridge).

Each question must require two different attached articles. A clue
from one article must identify a shared key, and this key must then
be used to retrieve the final answer from the other article. The
shared key may be a compound, metal complex, ligand, catalyst,
solvent, material, reaction type, property, analytical method,
experimental condition, application, or another chemically meaningful
concept.

General pattern:
What is property X of the entity or concept identified by property Y?

Here, property Y identifies the shared key using one article, while
property X is retrieved from the other article.

Create 10 question-answer pairs for the attached article pairs.

Requirements:
1. Questions must be standalone, clear, and unambiguous.
2. Do not mention the articles, titles, authors, journals, DOIs,
tables, figures, or phrases such as "in the first article", "in the
second article", "as reported", or "according to the study".
3. Both articles must be necessary: one must identify the shared key,
and the other must provide the final answer.
4. The clue for identifying the shared key must come from one
article, and the final answer must come from the other.
5. Do not create questions answerable from one article alone.
6. The shared key must be explicit and identifiable without
guesswork.
7. The final answer must be explicitly supported by the second
article.
8. Avoid questions based on general chemical knowledge, broad
background information, or facts likely to be found by a simple web
search.
9. Avoid opinion-based, speculative, or purely interpretive
questions.
```

```
10. Avoid ambiguous rankings such as "the best", "the most
effective", or "the most stable" unless the criterion is explicitly
stated and directly supported.
11. Do not present article-specific findings as universal chemical
truths; tie them to the relevant system, compound, method, or
conditions.
12. Make the 10 questions diverse in topic, shared key type, and
final answer type.

For each item, provide the question, shared key, final answer,
evidence for identifying the key, evidence for retrieving the answer,
and DOIs of both articles.

Return a table with the columns:
Question | Shared key | Answer | Evidence for key identification |
Evidence for final answer | DOIs.
```

K.3 SCENARIO 3 PROMPT TEMPLATE.

```
You are an expert in chemistry and scientific benchmark construction.
Rewrite all questions in the "question" column so that each question
follows the Conditional Query from Multiple Sources scenario.

Scenario: Conditional Query from Multiple Sources.

Each question must define several conditions coming from different
sources, and the answer must be the central entity that satisfies all
these conditions. The central entity may be a compound, material,
ligand, metal complex, catalyst, drug, solvent, method,
property-bearing system, or another chemically meaningful entity.

Example:
Original question:
What coordination geometry is formed by the 99mTc renal imaging
radiopharmaceutical that has an adult injected activity of 58 MBq and
an estimated effective dose of 0.4 mSv?

Central concept:
99mTc renal imaging radiopharmaceutical

Rewritten question:
Which 99mTc renal imaging radiopharmaceutical is administered at 58
MBq in adults, has an estimated effective dose of 0.4 mSv, and forms
the five-coordinate, distorted square-pyramidal complex?

Answer:
99mTc-MAG3

Task:
Rewrite each question so that the answer is the central entity
linking all stated conditions.

Requirements:
1. The rewritten question must be standalone, clear, and unambiguous.
2. The answer must be the entity satisfying all conditions, not a
property value.
3. Conditions must come from multiple sources and must all be
necessary to identify the answer.
4. Do not mention articles, titles, authors, journals, DOIs, tables,
figures, or phrases such as "in the article", "as reported", or
"according to the study".
```

```
5. Preserve the original scientific meaning and all essential
conditions.
6. Do not add unsupported information or change numerical values,
units, compounds, properties, or experimental conditions.
7. Avoid broad, opinion-based, speculative, or purely interpretive
formulations.
8. Do not present source-specific findings as universal chemical
truths; keep the conditions tied to the relevant entity, system,
property, or context.
9. Make the rewritten question natural and grammatically correct.

Return a table with the columns:
Original question | Central entity type | Rewritten question |
Answer.
```

K.4 SCENARIO 4 PROMPT TEMPLATE

```
You are an expert in chemistry and scientific benchmark construction.
Create high-quality question--answer pairs for evaluating
retrieval-augmented generation (RAG) systems in chemistry.

Scenario: Comparison of Entities (Comparison Query).

Each question must compare one entity from the first PDF with one
entity from the second PDF using the same clearly defined property,
metric, outcome, or criterion. Both PDFs must be necessary: each must
provide one comparable entity and its value or result, and the answer
must follow from comparing them.

You may receive optional Direct Fact Retrieval questions for each PDF.
Use them only as hints for candidate entities, properties, and fact
types; do not simply restate or merge them.

General pattern:
Which of entity X and entity Y has the
higher/lower/better/worse/larger/smaller value according to property
or criterion Z?

Example:
Q: How did changing the chelator from DOTA to NOTA affect kidney
uptake for the A1-His and Ec1 tracers?
A: The direction of the effect was opposite in the two systems. For
A1-His, changing from DOTA to NOTA increased kidney uptake from 29.10
± 2.24 to 50.42 ± 8.02 %ID/g, so DOTA gave lower kidney retention.
For Ec1, changing from DOTA to NOTA decreased kidney uptake from 233
± 27 to 174 ± 17 %ID/g, so NOTA gave lower kidney retention.

Create at least 10 question-answer pairs.

Requirements:
1. Questions must be in English, standalone, clear, and unambiguous.
2. Do not mention PDFs, articles, titles, authors, tables, figures,
supplementary information, or phrases such as "in the first PDF" or
"in the second PDF".
3. The compared entities must come from different PDFs, and both PDFs
must be necessary to answer the question.
4. The comparison must use the same explicit property, metric,
outcome, or criterion in both PDFs.
```

```
5. Compared values must be scientifically comparable in meaning,
units, and context; do not compare different properties, incompatible
conditions, unclear contexts, or similar terms with different
meanings.
6. Prefer explicit, measurable, PDF-supported comparisons: numerical
values, categorical results with clear ordering, or directly stated
performance metrics.
7. If several values are reported for the same entity, use only
values clearly matched in context and meaning.
8. Avoid vague, opinion-based, speculative, overly easy, or
one-PDF-answerable questions.
9. Avoid questions requiring guesswork, hidden assumptions,
unsupported domain knowledge, averaging, heavy calculation, or
reconstruction from many scattered statements.
10. If no strong valid comparison exists, output exactly:
NO_VALID_QUESTION.

For each item, provide the question, answer, evidence from both PDFs,
and DOIs. Evidence should come from PDF text, tables, figure
captions, or clearly readable labels; include exact quotes when
possible, page numbers, and relevant table or figure numbers.

Return a markdown table with the columns:
Question | Answer | Evidence from PDF 1 (exact quote +
page/table/figure) | Evidence from PDF 2 (exact quote +
page/table/figure) | DOIs of both articles.

Final self-check:
Does the question require both PDFs? Is the same comparison criterion
used on both sides? Are the entities from different PDFs? Is the
comparison scientifically fair and explicitly supported?
```

K.5 SCENARIO 5 PROMPT TEMPLATE.

```
You are an expert in chemistry and scientific benchmark construction.
Create high-quality question--answer pairs for evaluating
retrieval-augmented generation (RAG) systems in chemistry.

Scenario: Full Query (Set / Aggregation Query).

Each question must require retrieving the complete set of entities,
facts, properties, compounds, tissues, mechanisms, geometries, ions,
approvals, or related items that satisfy the query by aggregating
evidence across 2--5 provided articles.

General pattern:
Which entities or facts across the provided sources satisfy condition
or criterion X?

Example:
Q: Which tissues or organs remain above 1% of the administered dose
at the stated post-administration timepoints for untreated 153Gd and
for 67Ga-KEMPPr(3,4-HP)3?

Create 10 new question-answer pairs in the same style.

Requirements:
1. Questions must be standalone, precise, scientific, and
unambiguous.
```

```
2. Each question must require combining evidence from at least two
different articles.
3. Do not create questions answerable from a single article alone.
4. The answer must be a complete set of all relevant items supported
by the sources.
5. Use only the provided articles; do not add outside knowledge.
6. Prefer questions testing cross-document synthesis rather than
simple lookup.
7. Avoid yes/no questions, vague formulations, and unsupported
generalizations.
8. If sources partially overlap, merge them into one unified answer
without duplicates.
9. For each item, include exact supporting evidence from each source:
short quote, table number, figure number, scheme, or section.
10. Every row must use at least two sources.

Return a table with the columns:
Question | Answer | Sources from the articles (quote, table number,
or figure) | DOIs of all articles.

Formatting rules:
One row must contain one benchmark item. In the evidence column,
group evidence by article. In the DOI column, list all DOIs used for
that row.

Final self-check:
Does the question require aggregation across multiple articles? Is
the answer a full set rather than a partial subset? Are all claims
directly supported by the cited sources?
```

K.6 SCENARIO 6 PROMPT TEMPLATE

```
You are an expert in chemistry and scientific benchmark construction.
Create high-quality benchmark items for evaluating
retrieval-augmented generation (RAG) systems in chemistry.

Scenario: Multi-turn Query.

Each benchmark item must be a sequence of 3-5 connected questions.
The first question should start from a relatively broad query, and
each following question must refine, narrow, compare, or redirect the
search using the answer to the previous question. It should be
impossible to answer a later question correctly without knowing the
previous answer.

General pattern:
Q1 identifies entity or concept X. Q2 asks about a property,
condition, mechanism, value, or related entity of X. Q3 then uses the
answer to Q2 to ask for a comparison, consequence, synthesis detail,
biological result, coordination feature, or further linked fact.

Example:
Q1: Which zirconium-89-labeled octreotide radiopharmaceutical was
developed for PET imaging of somatostatin receptor-positive
neuroendocrine tumors?
A1: [89Zr]Zr-DFO-TOC.

Q2: Among AR42J, PC-3, and PANC-1 cells, which showed the highest
uptake of this tracer?
A2: AR42J.
```

```
Q3: Among [68Ga]Ga-THP-TATE and [68Ga]Ga-DOTA-TATE, which tracer gave
a tumor-to-liver ratio of 27.2 in SST2-positive tumours of this cell
line?
A3: [68Ga]Ga-DOTA-TATE.

Q4: Among DATA-TOC and the TOC tracer containing this chelator, which
tracer had the higher target-to-blood ratio in the first-in-human
comparison, and what were the two ratios?
A4: DOTA-TOC had the higher target-to-blood ratio: the ratios were
17.4 for DATA-TOC and 35.6 for DOTA-TOC.

Create 10 multi-turn benchmark items.

Requirements:
1. Each item must contain 3-5 sequential questions with answers.
2. Questions must be in English, standalone within the sequence,
clear, and unambiguous.
3. Each later question must explicitly depend on the answer to the
previous question, for example by using "this compound", "this
complex", "this ligand", "this value", or another clear reference.
4. Use at least two source articles across each sequence. If several
turns use the same article, their evidence must come from different
parts of the text, tables, figures, or schemes.
5. Include questions with different answer types: single values,
single concepts, lists, and short reasoning-based answers.
6. Prefer questions that require cross-document linking, comparison,
refinement, or synthesis rather than independent single-hop lookup.
7. Use information from tables where appropriate; table-based
evidence is especially encouraged.
8. Do not use outside knowledge or unsupported assumptions.
9. Avoid yes/no questions, vague formulations, and questions
answerable without the previous turn.
10. Do not mention articles, titles, authors, journals, DOIs, tables,
figures, or phrases such as "in the article" in the question text.

For each item, provide the full sequence of questions and answers,
supporting evidence for each turn, and DOIs of all articles used.
Evidence should include short quotes, table numbers, figure numbers,
schemes, sections, or page numbers where available.

Return a table with the columns:
Item ID | Multi-turn question sequence | Answer sequence | Evidence
for each turn | DOIs of all articles.

Final self-check:
Does each later question depend on the previous answer? Does the
sequence use at least two sources or distinct evidence locations? Are
answer types diverse? Are all answers directly supported by the cited
evidence?
```

L EVALUATION PROMPTS

This appendix documents all prompts used in the evaluation pipeline, organized into four groups: the system prompt that controls generation behavior during retrieval-augmented inference (Section L.1), standalone LLM system prompt (Section L.2), the LLM-as-a-judge prompts for continuous-valued metrics (Section L.3), and the binary diagnostic judges that flag specific failure modes (Section L.4). All production benchmark judges in these prompts use the same backend (Gemini-2.5-flash, served via OpenRouter) at temperature 0 and seed 42; the alternative Jev validation is reported separately in Appendix D. The same Gemini backend is used for RAGAS and DeepEval implementations to ensure cross-framework comparability.

L.1 GENERATION SYSTEM PROMPT

The following system prompt is used by all RAG configurations. It enforces strict grounding for in-domain queries and explicit refusal when retrieval returns no usable context, which is what the *G_REFUSE* and *C_NO_SOURCE* diagnostics measure.

```
You are a Chemistry Research Assistant with deep expertise.

For advanced chemistry questions, you must rephrase the query into
clear English, call the `rag_search` tool, and synthesise your answer
strictly from the tool output.
For basic school-level chemistry questions, answer directly without
calling the tool.
For questions unrelated to chemistry, politely decline.

When `rag_search` is used, base the answer exclusively on the
retrieved content.
If the tool returns no results or no sufficiently relevant passages,
reply:
"I did not find relevant information in the knowledge base"
and do not attempt to answer.
If the tool returns partial information, answer only within those
boundaries
and state explicitly what is unknown.

Always respond in the user's language.
You are permitted to call `rag_search` at most once per user message.
```

L.2 STANDALONE LLM SYSTEM PROMPT

The three standalone LLM baselines (GPT-5.4, GPT-5.4 no reasoning, Qwen-3.6-27B) receive each question with the system prompt below and no other context. The prompt establishes a generic scientific-assistant role without any retrieval tool or domain-specific instructions, so all responses come from parametric knowledge alone.

```
You are a scientific AI assistant with expertise in chemistry,
materials science,
and related fields.
TASK
Answer scientific questions. Provide accurate, detailed answers based
on established
scientific principles and facts.
GUIDELINES
Give precise, factual responses
Include specific values, formulas, and technical details when
relevant
Explain concepts clearly and scientifically
If uncertain about specific data, acknowledge limitations
Use standard scientific terminology and notation
```

L.3 CONTINUOUS-VALUED METRIC JUDGES

The five prompts below produce continuous scores (0–4 scale or fractional). *Answer accuracy* and *answer relevancy* grade the model's response against the question and reference answer. *Context relevance* is applied to each retrieved chunk independently and aggregated. *Faithfulness* is computed in two stages: the prompt below decomposes the model's answer into atomic statements, and a separate entailment check (using the same judge) verifies each statement against the retrieved context. *Context recall* mirrors faithfulness but checks whether claims of the reference answer can be attributed

to the retrieved context, measuring whether retrieval surfaced enough information for a complete answer.

L.3.1 ANSWER_ACCURACY

```
You are a scientific accuracy evaluator.  Your task is to assess how
correctly a Model Response answers a Question compared to a Reference
Answer.
EVALUATION CRITERIA:
     - Focus on factual correctness, not style or verbosity
     - Numerical values must match exactly (2 µg/mL ≠ 20 µg/mL)
     - Units must be correct (µg/mL ≠ mg/mL)
     - Chemical names and formulas must be accurate
     - For multi-part answers, assess what fraction of key facts are
       correct
SCORING SCALE:
       0 - INCORRECT: The response is factually wrong, contradicts
       the reference, or does not answer the question at all.
       1 - MINIMAL: Only minor overlap with reference.  Most key
       information is missing or incorrect.
       2 - PARTIAL: Some key facts are correct, but significant
       information is missing or wrong.
       3 - MOSTLY CORRECT: Almost all facts are correct.  Only minor
       details are missing or slightly imprecise.
       4 - CORRECT: All key facts from the reference are present and
       accurate in the response.
INPUT:
Question:  {question}
Reference Answer:  {golden_answer}
Model Response:  {generated_answer}
First, briefly analyze what key facts are in the Reference Answer and
whether they appear correctly in the Model Response.  Then provide
your score.
Output format:
Analysis:  <your brief analysis>
Score:  <0, 1, 2, 3, or 4>
```

L.3.2 CONTEXT_RELEVANCE

```
You are evaluating whether a retrieved text chunk is relevant to a
question.

A chunk is RELEVANT if it contains information related to the
question topic, even if it doesn't fully answer it.
A chunk is NOT RELEVANT if it is off-topic or contains unrelated
information.

Question: {question}

Chunk: {ctx}
```

```
Is this chunk relevant to the question? Answer only YES or NO.
```

L.3.3 FAITHFULNESS

```
Split the following answer into separate factual statements. Each
statement should contain exactly one claim or fact.

Answer: {generated_answer}

Output each statement on a new line, numbered:
1. <first statement>
2. <second statement>
```

L.3.4 ANSWER_RELEVANCY

```
You are evaluating whether an answer is on-topic for a given question.
Ignore correctness, focus only on relevance.

SCORING SCALE:
0 - IRRELEVANT: The answer is completely off-topic or does not
address the question at all.
1 - MINIMALLY RELEVANT: The answer mentions the topic but does not
actually address what was asked.
2 - PARTIALLY RELEVANT: The answer addresses part of the question but
ignores other parts or is too vague.
3 - MOSTLY RELEVANT: The answer addresses the question well but may
include some off-topic information.
4 - FULLY RELEVANT: The answer directly and completely addresses
exactly what was asked.

Question: {question}

Answer: {generated_answer}

Respond with ONLY the number 0, 1, 2, 3, or 4.
```

L.3.5 CONTEXT_RECALL

```
Given a context and answer, analyze each sentence in the answer and
classify if the sentence can be attributed to the given context or
not.

INSTRUCTIONS:
- For each sentence, determine if it can be attributed to (supported
by) the provided context
- A sentence is ATTRIBUTED if the context contains information that
supports or mentions the claim
- Focus on factual content, ignore minor formatting differences
- Output verdict for each sentence: 1 if attributed to context, 0
otherwise

CONTEXT:
{context_str}

ANSWER TO ANALYZE:
{golden_answer}
```

```
SENTENCES TO EVALUATE:
{claims}

For each sentence, provide verdict (1 if attributed to context, 0 if
not):
OUTPUT FORMAT: [verdict1, verdict2, verdict3, ...]
EXAMPLE OUTPUT: [1, 0, 1]
```

L.4 DIAGNOSTIC JUDGES

The diagnostic judges share a single template (below) and differ only in the failure mode they target. Each judge takes the user query, reference answer, retrieved contexts, and model response, and returns a binary flag indicating whether the specific failure mode is present. The fields `{failure_code}` and `{failure_description}` are filled from the catalogue listed afterwards. Codes prefixed with `C_` target context/retrieval-side failures; codes prefixed with `G_` target generation-side failures. The single-condition design is deliberate: by isolating each failure mode in a separate call, the judge cannot trade off one error against another, and the resulting flags can be summed to produce per-domain or per-scenario error breakdowns.

L.4.1 DIAGNOSTIC PROMPT TEMPLATE

```
You are a highly specialized binary diagnostic judge for RAG systems.
Your SOLE objective is to evaluate the Model Response and Contexts
against ONE specific FAILURE MODE.

FAILURE MODE: {failure_code}: {failure_description}

    ---
    1. User Query: "{question}"
    2. Reference Answer (Golden Truth): "{golden_answer}"
    3. Retrieved Contexts: {context_str}
    4. Model Response: "{generated_answer}"
    ---

    STRICT DECISION RULES:
    1. Assess only the condition defined by the FAILURE MODE
    description. IGNORE all other potential errors.
    2. If the condition is met, respond with 1.
    3. If the condition is NOT explicitly and clearly met, respond
    with 0.
    4. Your response MUST be ONLY the single digit 1 or 0.
```

L.4.2 FAILURE MODE CATALOGUE

```
"C_EXTRA_SOURCE":

Retrieved Contexts contain blocks of information that are clearly
irrelevant to the User Query, AND the Model Response includes
non-essential, irrelevant statements derived from these extraneous
contexts.

"C_WRONG_SOURCE":

At least one of the Retrieved Contexts contains information that
directly contradicts the Golden Answer, indicating the retrieval
system returned a false source.
```

```
"C_NO_SOURCE":

The Retrieved Contexts list is empty ('NONE' or zero chunks), but the
Model Response still attempts to provide a factual answer, indicating
an LLM hallucination.

"G_INCOMPLETE":

The Model Response is missing one or more critical facts from the
Golden Answer, AND these specific missing facts were clearly present
and available in the Retrieved Contexts.

"G_IMPRECISE":

The Model Response contains factual distortions, incorrect numbers,
or imprecise statements compared to the Golden Answer, making the
final information potentially misleading or unusable.

"G_EXTRA_INFO":

The Model Response includes significant factual details that were NOT
explicitly requested by the User Query, even though these extra
details are supported by the Retrieved Contexts.

"G_CLARIFY":

The Model Response includes an explicit question to the user or
requests further clarification or input, instead of providing the
requested factual information.

"G_REFUSE":

The Model Response explicitly states an inability or refusal to
answer the query (e.g., due to ethical policies, lack of internal
knowledge, or limitations).

"G_SCIENTIFIC_ERROR":

The Model Response contains numerical values, chemical formulas, or
statistical data that CONTRADICT the Golden Answer. This includes
incorrect temperatures, concentrations, lattice parameters, or molar
ratios. Even a small numerical deviation is a failure.
```

M USE OF SPECTRA FOR ML-DRIVEN RAG SYSTEM DESIGN

Three SPECTRA domains (Chelate Metal Complexes, Nanozymes, and Oxazolidinone Antibiotics), with the same 50-question samples as Appendix I, have been used in a separate study that treats RAG configuration as a supervised prediction problem. The study evaluates 648 configurations per domain, varying search mode, reranker, generation backbone, chunk size, and retrieval depth, and trains surrogate models that predict retrieval and generation metrics from configuration parameters with $R^2 \geq 0.84$ in every domain. Without target-domain data, the surrogates transfer across domains for Golden DOI Recall and Scientific Fact Recall; with 10% of target-domain runs, they transfer on all metrics for two of the tested domain pairs. Used as objectives in a search over about 61,000 candidate configurations, they recover configurations on or near the Pareto front of the exhaustive grid, including in a fully out-of-domain setting. These results show that SPECTRA supports learned configuration design. Extending it from scores averaged over Scenarios 1–5 to full per-scenario and per-modification profiles remains open.