LAKEQUEST: A Three-Domain Benchmark for Grounded Question Answering across Data Lakes

Michael Solodko[*] **Steven Gong** **Guangwei Yu**[⑥] **Satya Krishna Gorti**[⑥]
Jesse C. Cresswell[⑥] **Victor Zhong**
University of Waterloo [⑥]Layer 6 AI

Code, data, and starter kit: https://github.com/michael0402/LakeQuest
LakeQuest Homepage: https://michael0402.github.io/LakeQuest/

Abstract

While modern question answering (QA) systems excel on clean, schema-aligned corpora, real-world knowledge is rarely so neatly packaged. Answering questions over enterprise and scientific data lakes requires systems to navigate heterogeneous, weakly structured collections of tables, passages, and linked metadata. Current benchmarks abstract away this noisy discovery process, failing to evaluate end-to-end performance. To bridge this gap, we introduce LAKEQUEST, a human-validated benchmark of 9,846 QA pairs designed to evaluate the end-to-end retrieve-and-synthesize pipeline over realistic data lakes. LAKEQUEST spans three diverse domains—AI/ML metadata, retail banking, and multimodal biomedical drug information—and pairs every question with exact, modality-aware evidence pointers. By isolating source discovery from cross-modal synthesis, LAKEQUEST exposes critical failure modes in modern QA systems. Our baseline evaluations, including standard Retrieval-Augmented Generation (RAG) and agentic tool-use methods, reveal that high-quality retrieval does not guarantee correct reasoning. Systems consistently struggle with relation chaining in metadata graphs, policy grounding in bank ledgers, and joint tabular QA in biomedical contexts, highlighting the need for robust discovery and faithful cross-file composition mechanisms in future agentic QA systems.

1 Introduction

Question answering (QA) systems have achieved strong performance on curated, homogeneous corpora, ranging from reading comprehension (Rajpurkar et al., 2016) to open-domain text retrieval (Kwiatkowski et al., 2019) and multi-hop reasoning (Yang et al., 2018). However, real-world enterprise and scientific knowledge is rarely stored as flat text. Instead, it often resides in **data lakes**: heterogeneous, weakly structured collections of tables, passages, and linked metadata.

Answering natural language questions over data lakes requires navigating siloed sources, combining cross-modal evidence, and maintaining explicit provenance. However, modern retrieve-and-synthesize architectures (Lewis et al., 2020; Yao et al., 2023) lack evaluation environments that simultaneously stress discovery, cross-source composition, and faithful attribution over heterogeneous corpora. Existing benchmarks bypass the noise, scale, and weak metadata characteristic of real data lakes, evaluating either pure data discovery without natural language synthesis, or reasoning over already-clean, schema-aligned text.

To bridge this gap, we introduce LAKEQUEST, a benchmark designed to evaluate the end-to-end QA pipeline over realistic data lakes. LAKEQUEST provides a shared construction and evaluation framework instantiated across three diverse domains: AI/ML metadata, a realistic retail bank, and multimodal biomedical drug information. Rather than flattening

[*]Correspondence to: msolodko@uwaterloo.ca

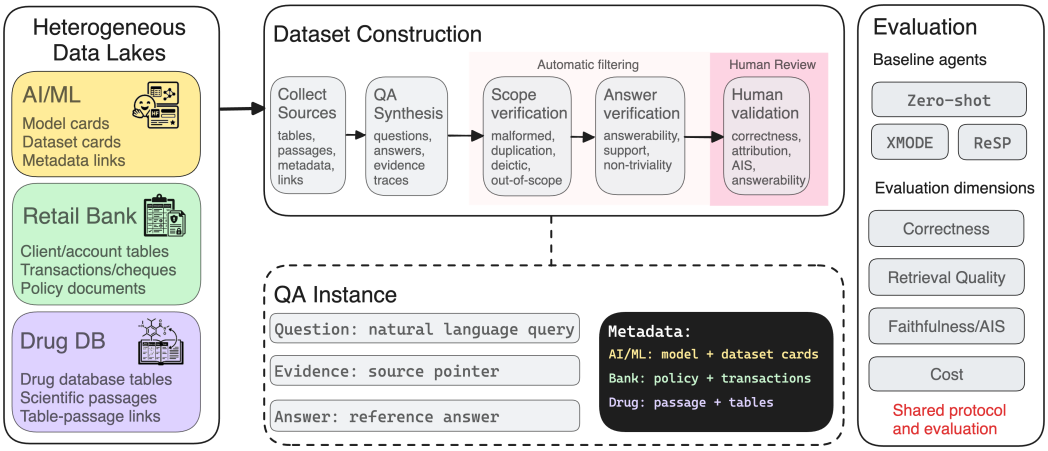

Figure 1: Overview of LAKEQUEST. The benchmark spans three heterogeneous data lakes built under a shared construction and evaluation protocol for grounded question answering.

these domains into a uniform representation, LAKEQUEST preserves their native structural heterogeneity as data lakes. Each benchmark instance pairs a natural language question with a reference answer and exact, modality-aware evidence pointers (e.g., specific table rows or passage spans). Furthermore, because LAKEQUEST preserves the native structure of the data lakes, it inherently captures factual redundancy. A system might successfully synthesize a grounded answer by discovering valid *alternative provenance* (e.g., a redundant table or densely informative passage) even if the target gold artifact is missed. This explicit attribution enables precise isolation of system failures into discovery errors (failing to retrieve evidence) versus reasoning errors (failing to synthesize an answer from the retrieved evidence) (Bryan et al., 2025; Leung et al., 2026).

We construct LAKEQUEST using a reproducible five-stage pipeline encompassing domain instantiation, language-model-driven QA synthesis, structural filtering, semantic validation, and rigorous human review. The resulting benchmark contains 9,846 human-validated QA pairs that entail single-source lookup, cross-source composition, and multi-hop reasoning.

Evaluating a diverse suite of baselines—from zero-shot reasoning (Kojima et al., 2022) and standard retrieval-augmented generation (RAG) (Lewis et al., 2020) to structured pipelines and agentic tool-use methods such as XMODE (Nooralahzadeh et al., 2025) and RESP (Jiang et al., 2025)—we find that the bottlenecks in data lake QA remain fundamentally unsolved and are strongly domain-dependent. Specifically, accurate retrieval does not guarantee faithful reasoning when cross-modal composition is required, and even agentic strategies struggle with multi-hop discovery over weakly structured metadata.

Our primary contributions are: **(1) The LAKEQUEST Benchmark:** A human-validated dataset of over 9.8k questions evaluating grounded reasoning across three heterogeneous data lakes (AI/ML, Retail Banking, and Biomedical Drug). **(2) A Reproducible Construction Framework:** A five-stage protocol for synthesizing, filtering, and validating QA pairs with explicit evidence traces over weakly structured corpora. **(3) Empirical Analysis of Modern QA:** Extensive baseline evaluations revealing that high retrieval quality does not consistently translate into accurate, faithful answers—especially when multi-step or cross-source composition is required.

2 Related Work

Table 1 provides a structural comparison between LAKEQUEST and existing QA and data discovery benchmarks across modalities, evidence tracing, and evaluation scope.

Data Discovery and Integration. Data management literature extensively studies how to discover, link, and integrate datasets (Fernandez et al., 2018; Fan et al., 2023). Accordingly, benchmarking efforts in this space, such as LakeBench (Deng et al., 2024) and CMDL (Eltabakh et al., 2023), evaluate the retrieval of joinable tables or cross-modal artifacts. While

Table 1: Comparison of LAKEQUEST with existing QA and data discovery benchmarks. **Modality** describes the data formats involved. **Ev** indicates the presence of explicit evidence traces linking answers to sources. **Md** denotes the inclusion of weak metadata and distractor documents to test retrieval robustness. **Val** indicates datasets that underwent human validation. **Objects** denotes approximate source-side evidence containers (e.g., paragraphs, tables, passages, or hybrid contexts).

Benchmark	Modality	Ev	Md	Val	# Qs	Objects	Task	Domain
HotpotQA	Text	✓	×	✓	113k	5M+ paragraphs	Text Reasoning	Wikipedia
HybridQA	Tables Text	×	×	✓	69k+	13k tables + 293k passages	Hybrid Reasoning	Wikipedia
TAT-QA	Tables Text	×	×	✓	16k+	2.5k+ hybrid contexts	Numerical QA	Finance
LakeBench	Tables Metadata	×	✓	✓	10k+	16M+ tables	Table Discovery	Open Data
LAKEQUEST	Tables, Text Metadata	✓	✓	✓	9.8k	1M+ model/data cards + 10k+ bank records + 500k+ drug texts/tables	Grounded QA	AI/ML Banking Drug

essential for data discovery, these benchmarks remain purely retrieval-focused and do not evaluate natural language synthesis.

Text-centric and Multi-hop QA. In the NLP community, QA benchmarks have evolved from single-document reading comprehension (Rajpurkar et al., 2016) to open-domain retrieval (Kwiatkowski et al., 2019) and multi-hop reasoning with HotpotQA (Yang et al., 2018). While foundational, these datasets largely assume text-centric evidence within curated, relatively homogeneous corpora (e.g., Wikipedia). Real data lakes, however, are characterized by weak metadata, fragmented entity references, and a mixture of semi-structured objects, making reasoning pathways noisier than those found in standard text-only multi-hop settings.

Hybrid and Multimodal QA. A closer line of work evaluates QA over multiple evidence types. HybridQA (Chen et al., 2020) and OTT-QA (Chen et al., 2021) require reasoning across tables and text, while TAT-QA (Zhu et al., 2021) emphasizes numerical reasoning over financial reports. MultiModalQA (Talmor et al., 2021) expands the scope to joint reasoning over text, tables, and images. While these benchmarks require cross-modal synthesis, they generally operate over highly curated, schema-aligned, or task-specific corpora. They do not model the core challenges of enterprise data lakes: navigating siloed, weakly linked sources, resolving entities across heterogeneous files, and maintaining strict, modality-aware provenance for auditability.

Retrieval-Grounded and Agentic Systems. Methodologically, LAKEQUEST provides a necessary testbed for modern retrieve-and-synthesize pipelines. RAG (Lewis et al., 2020) and fusion-in-decoder (FiD) (Izacard & Grave, 2021) established paradigms for grounding generation in retrieved contexts, while recent methods propose structured decomposition (Zhou et al., 2023), iterative tool use (Yao et al., 2023; Schick et al., 2023), and multi-agent orchestration like Symphony (Tang et al., 2024). By providing a heterogeneous environment with explicit evidence requirements and realistic discovery bottlenecks, LAKEQUEST enables the community to rigorously evaluate whether these control strategies can handle the noise and scale of real-world data lakes.

3 The LAKEQUEST Framework

To rigorously evaluate QA over heterogeneous data lakes, we require a framework that isolates source discovery from cross-modal synthesis, scales across domains, and maintains strict provenance. LAKEQUEST operationalizes this through a formalized task definition and a reproducible construction pipeline.

Domain: Retail banking **Family:** Cross-source composition (Policy + Table + Metadata)

Question (q_i): Based on standard routing policies, does the wire transfer initiated by Acme Corp on Oct 14th require manual manager approval?
Gold Evidence (E_i^*):
- `Policy_Manual_v3`: Span [1124-1180] *("...Tier 2 business accounts require manual manager approval for all outbound international wire transfers exceeding $15,000...")*
- `Transaction_Ledger_72` (Acme Corp): Row 84, Columns `Type`, `Amount` *("Int_Wire $18,500.00")*
- `Client_Profile_7742`: Key `Account_Tier` *("Tier 2")*

Reference Answer (a_i^*): Yes, the $18,500 transfer exceeds the $15,000 limit for Tier 2 accounts.

Figure 2: Anatomy of a LAKEQUEST instance — this requires identifying and synthesizing specific, sub-document evidence units across multiple heterogeneous sources to derive the correct answer.

3.1 Problem Formulation and Task Taxonomy

We formulate grounded data lake QA: given a data lake $\mathcal{D} = \{d_1, \ldots, d_n\}$ (where d_i is a table, passage, or metadata record) and a natural language question q, a system f must return a textual answer $\hat{a}$ and a precise supporting evidence set $\hat{E} \subseteq \mathcal{D}$: $f(q, \mathcal{D}) \rightarrow (\hat{a}, \hat{E})$. Modern agents decompose this operation into retrieve-and-synthesize workflows: $\hat{E} = R(q, \mathcal{D})$ and $\hat{a} = G(q, \hat{E})$. By requiring systems to output $\hat{E}$, LAKEQUEST allows us to independently evaluate the retrieval mechanism R (discovery) and the generator G (synthesis).

To systematically test this pipeline, LAKEQUEST is organized into four distinct reasoning families: **single-source lookups** requiring fact extraction from isolated objects; **cross-source composition** necessitating the combination of facts across multiple heterogeneous objects (e.g., joining a policy with a transaction); **multi-hop reasoning** requiring chained inference over intermediate entities or linked metadata; and **anchor-based retrieval**, which conditions questions on a selected entity, forcing the agent to reason over related entities amidst semantically close distractors.

Unlike standard text QA benchmarks where evidence is simply a document ID, LAKEQUEST enforces modality-aware provenance. Each benchmark instance is defined by the tuple $x_i = (q_i, a_i^*, E_i^*, m_i)$, where E_i^* resolves to concrete sub-document units (e.g., specific table rows or exact passage spans) and m_i tracks the question family and domain metadata. Figure 2 illustrates a concrete cross-source instance from the Retail Bank data lake.

3.2 The Shared Five-Stage Pipeline

Manually authoring tens of thousands of complex, cross-modal QA pairs is unscalable, while purely synthetic datasets often suffer from hallucinations and trivial reasoning pathways. To balance scale with rigor, we construct LAKEQUEST using an LLM-in-the-loop pipeline consisting of five shared stages across all domains (a visual schematic of the Retail Bank data lake question-answer synthesis process is provided in App G).

The five shared stages include: **(1) Lake Instantiation**, preserving the domain's native schema rather than homogenizing documents; **(2) Candidate Synthesis**, where a strong language model (OpenAI GPT-5, Feb. 1, 2026) synthesizes a candidate question q, answer a^*, and exact evidence pointers E^* from a sampled evidence subgraph; **(3) Structural Filtering**, which programmatically removes malformed outputs or deictic questions; **(4) Semantic Filtering**, utilizing an LLM-as-a-judge to reject trivial candidates answerable via parametric memory alone; and **(5) Human Validation**, the definitive gate for benchmark inclusion, assessing answerability and attribution. Exact prompts and rubrics are detailed in App B.

3.3 Human Validation and Quality Control

To guarantee evaluation fidelity, all items passing the automated filters are subjected to rigorous human validation. Annotators were tasked with independently verifying two criteria: *Answerability* (can the question be answered exclusively using the cited evidence?) and *Attribution* (is the generated answer completely faithful to the cited text/table?). The use

Table 2: Abstraction map to unify task design while preserving domain-native structure.

Domain	Lake Representation	Task Abstraction	Primary Reasoning Stress
AI/ML	Linked metadata entities (models, datasets, relations)	Entity/relation QA with lookup, composition, hop, and anchor families	Entity disambiguation and relation chaining
Retail Bank	Structured operational records with policy context	Persona-workflow tasks with client/group scopes	Decision grounding under workflow/policy constraints
Drug	Hybrid tables + scientific passages with linkage artifacts	Evidence-bundle QA across single-modality and mixed-modality settings	Cross-modality evidence integration and multi-hop synthesis

of human annotators was approved by the ethics review board at the authors' institution. Prolific annotators (paid $12 USD/hr) independently verified each item under this protocol. Inter-annotator agreement (Fleiss' κ) demonstrates strong consistency across raters: 0.82 for Answerability and 0.85 for Attribution. Full annotator guidelines, including a screenshot of the custom validation interface (Figure 3), are provided in App C. Only instances that achieved strict human consensus were retained in the final LAKEQUEST release.

4 The LAKEQUEST Dataset — Instantiating the Three Data Lakes

To evaluate the framework across diverse reasoning requirements, LAKEQUEST instantiates three data lakes: an AI/ML metadata repository, a synthetic retail banking system, and a biomedical corpus. Table 2 outlines the high-level abstractions used to preserve the native structures of these domains. Furthermore, Table 3 details the final benchmark coverage and human annotation outcomes across these three lakes.

AI/ML Data Lake The AI/ML data lake focuses on entity disambiguation and relation chaining over weakly linked metadata. We construct the corpus using a snapshot of the Hugging Face Hub API (Hugging Face, 2025). The resulting lake consists of model cards and dataset cards, which are inherently semi-structured. We instantiate 5 distinct question families in this domain, yielding 3,578 human-validated questions (see Table 3 for a high-level summary and Table 7 in App A for the detailed breakdown).

These tasks map to our core reasoning taxonomy by requiring systems to deduce specific hyperparameters from isolated cards (**single-source**), synthesize constraints by jointly reasoning over explicitly linked model and dataset cards (**cross-source**), chain relations across the metadata graph to link datasets and subsequent models (**multi-hop**), and formulate queries around specific anchor entities despite lexical drift (**anchor-based**).

Retail Bank Data Lake The Retail Bank data lake evaluates grounded decision-making under strict policy constraints. Because real operational banking data is highly restricted, we generate a high-fidelity synthetic lake by deriving generation seed examples and instantiating an open-banking-inspired relational schema (client, account, transaction, and policy entities). We then populate client profiles, transaction ledgers, and institutional policy documents. We further validate that transaction distributions are similar between the synthetic lake and original real banking data by matching graph density on client profiles. Tasks are formulated across 9 persona-driven workflows (e.g., a fraud analyst investigating a transaction, or an auditor reviewing compliance), yielding 2,964 validated questions (detailed breakdown provided in Table 8, App A).

These workflows map to core reasoning families: **single-source lookups** for retrieving isolated facts (e.g., account tiers or policy limits); **cross-source composition** to combine unstructured policy rules with structured transaction ledgers (e.g., checking if a wire transfer violates a tiered policy); **multi-hop reasoning** to trace complex fund flows across multiple tables before applying operational rules; and **anchor-based retrieval** to disambiguate compliance policies based on client attributes amidst overlapping regional guidelines.

Table 3: Benchmark coverage summary across all three lakes after human validation.

Domain	Question Families	Questions	Correct	AIS	Answerable
AI/ML	5 source combos	3,935	3,750	3,687	3,578
Retail Bank	9 personas	3,312	3,047	3,003	2,964
Drug	5 source combos	3,574	3,434	3,373	3,304
Total	–	**10,821**	**10,231**	**10,063**	**9,846**

Drug Data Lake The Drug data lake stresses multimodal evidence integration. We construct this corpus by fusing structured XML records from DrugBank 6.0 (Knox et al., 2024) with open-access scientific literature indexed in PubMed (Sayers et al., 2026). This heterogeneous mixture of dense text and structured properties instantiates 5 question families, producing 3,304 validated questions (see Table 9 in App A for statistical breakdown).

The reasoning families in this domain stress multimodal evidence integration: **single-source lookups** extract precise values from isolated DrugBank tables or PubMed abstracts; **cross-source composition** synthesizes answers by linking unstructured text passages (e.g., mechanism of action) to structured tables (e.g., biochemical properties); **multi-hop reasoning** chains relationships across modalities (e.g., identifying a target in a table, then finding passages detailing its metabolic pathway); and **anchor-based retrieval** requires identifying correct active compounds amidst structurally similar distractors and complex aliases.

Splits and Licensing To prevent contamination, we define (open) validation (20%) and test (80%) splits at the *entity* rather than question level (i.e., $\text{Entity}(\text{val}) \cap \text{Entity}(\text{test}) = \emptyset$). The LAKEQUEST benchmark and underlying corpora are hosted on Hugging Face at `anonymousurl`. The AI/ML and Retail Bank subsets are released under an MIT license. For the Drug subset, we release benchmark annotations and evidence pointers, while use of underlying DrugBank tables remains subject to DrugBank licensing terms and PubMed-linked content remains subject to NLM/PubMed access policies and publisher rights.

5 Experimental Setup and Baselines

To benchmark grounded reasoning over data lakes, we define a unified evaluation contract and test representative QA baselines ranging from direct prompting to agentic tool use.

Evaluation Metrics Because data lake QA serves both high-assurance settings (where hallucination is unacceptable) and exploratory analytics (where cost and speed are critical), we treat evaluation as a multi-dimensional quality-cost trade-off. We evaluate systems across four axes:

- **Accuracy (Correctness):** We report binary correctness using rubric-based LLM-as-a-judge for baseline runs (including direct-answer and reasoning-answer correctness where applicable), and human-judged correctness for annotated benchmark slices in App A.
- **Faithfulness (Attribution):** Whether the predicted answer $\hat{a}$ is supported by the predicted evidence set $\hat{E}$. We operationalize this using Attributable to Identified Sources (AIS) style judgments (Rashkin et al., 2023), aligned with the human attribution protocol used throughout this benchmark.
- **Retrieval Quality:** Domain/modality-aware recall of gold evidence E^* in $\hat{E}$ (e.g., card retrieval in AI/ML, policy and table recall in Retail Bank, and passage/table recall in Drug). For unstructured evidence, success is chunk-level span inclusion; for structured evidence, we require retrieval of the referenced source object and apply row/field-level checks when explicit row-level pointers are available. Because real-world data lakes inherently contain redundant information (e.g., same fact appearing in multiple tables/passages), this metric serves as a strict lower bound. A system may retrieve valid alternative evidence that correctly answers the question despite missing the specific gold artifact.
- **Token Usage:** End-to-end retrieval-plus-synthesis token footprint, reported as total model tokens (and normalized per-query token count) under the evaluation-time configuration.

Table 4: Baseline results for the AI/ML data lake. Metrics are Retrieval Recall (R), Accuracy (A), Faithfulness (F), and Token Count (T). Token Count (T) is reported in thousands of generation tokens per query. Model and Dataset are single-source lookup and anchor questions, Model-Dataset is cross-source composition, and Model-hop and Dataset-hop are multi-hop reasoning questions.

	RAG				Zero-shot	
Question Type	R	A	F	T	A	T
Model	0.09	0.35	0.25	0.39	0.25	0.06
Dataset	0.15	0.12	0.14	0.34	0.09	0.05
Model-Dataset	0.12	0.38	0.32	0.38	0.17	0.05
Model-hop	0.13	0.23	0.16	0.47	0.20	0.07
Dataset-hop	0.23	0.35	0.38	0.49	0.24	0.07

Evaluated Baselines We establish baseline performance using models with different levels of agentic complexity. We use GPT-5 as the generator (kept fixed within each experiment suite) (OpenAI, 2025) and FAISS-backed `text-embedding-3-small` for retrieval (Douze et al., 2026). Comprehensive details, including system instructions, retrieval formatting templates, and agentic tool-use prompts for the evaluated systems, are provided in App B.

- **Zero-Shot Prompting:** A naive baseline that directly prompts the generator LLM without explicit retrieval, testing the limits of parametric memory and instruction following.
- **Retrieval-Augmented Generation (RAG) (Lewis et al., 2020):** Used as our baseline for the document-only AI/ML data lake, this is a standard retrieve-then-generate pipeline. We chunk unstructured passages and structured metadata cards into overlapping text segments before indexing; we then retrieve the top-k most relevant chunks and synthesize a final answer. We experiment with different values of k and finalize on $k = 5$, which consistently attains strong performance.
- **Agentic and Structured Pipelines (XMODE, RESP) (Nooralahzadeh et al., 2025; Jiang et al., 2025):** Standard RAG is structurally insufficient for Retail Bank and Drug because it cannot natively execute the relational row operations that navigate Retail Bank transaction records and Drug ledgers. Instead, we evaluate stronger multi-step baselines that go beyond single-shot text retrieval by interleaving reasoning and tool use. In our implementation, XMODE acts as a tool-augmented RAG system: it uses an iterative loop with passage retrieval from a FAISS index alongside direct query/SQL access to structured tables. RESP follows an iterative retrieve–summarize–plan style decomposition.

6 Results and Cross-Domain Analysis

We evaluate baselines across the three domains to understand where modern retrieve-and-synthesize systems fail. Across domains, retrieval quality alone is insufficient; grounded data-lake QA requires stronger cross-source composition and attribution mechanisms.

6.1 AI/ML Data Lake: Discovery is the primary bottleneck

The AI/ML lake is challenging due to scale—it is built from a Hugging Face model-card corpus with over 1M cards and normalized into a large model–dataset graph with on-demand card-text retrieval—and the inconsistencies between how different authors document artifacts. As shown in Table 4, the RAG baseline exhibits a notable gap between exact-match Recall (R) and Accuracy (A). For example, on isolated Model questions, RAG achieves 0.35 Accuracy despite a recall of 0.09. This gap validates that real data lakes contain factual redundancy, for example from models that have been duplicated with minor modifications; the retriever frequently finds valid alternative evidence to synthesize the correct answer even if it misses the specific gold artifact (E^*). This reinforces that R acts as a strict lower bound and suggests improved retrieval as a logical next step for metadata QA. In summary, the wide gap between modest recall and reasoning accuracy suggests that navigating weakly linked metadata via dense retrieval is insufficient for reliable entity disambiguation and relation chaining.

Table 5: Baseline results for the Retail Bank data lake. Metrics are Table Recall (R_T), Policy Recall (R_P), Accuracy (A), Faithfulness (F), and Token Count (T; k-tokens/query, as in Table 4). Persona workflows span all four families in the task taxonomy depending on whether the task is isolated lookup, policy-ledger composition, multi-step trace, or anchor-conditioned policy selection.

	XMODE					RESP					Zero-shot	
Persona	R_T	R_P	A	F	T	R_T	R_P	A	F	T	A	T
Accountant	0.79	0.94	0.60	0.45	1.15	0.95	1.00	0.73	0.48	1.20	0.25	0.19
Auditor	0.84	0.90	0.58	0.51	1.00	0.94	0.85	0.58	0.55	1.12	0.69	0.19
CS manager	0.70	0.87	0.70	0.67	1.05	0.86	1.00	0.88	0.80	1.03	0.69	0.13
Customer service (CS)	0.82	0.98	0.64	0.64	1.26	0.93	1.00	0.58	0.50	1.07	0.69	0.16
Data analyst	0.62	0.69	0.59	0.49	1.10	0.89	0.65	0.55	0.38	1.04	0.31	0.16
Financial planner	0.64	0.99	0.65	0.59	1.24	0.85	0.90	0.68	0.70	1.05	0.88	0.16
Fraud analyst	0.72	0.92	0.61	0.47	1.10	0.95	0.90	0.68	0.60	1.27	0.81	0.15
Insurance analyst	0.52	0.89	0.58	0.45	1.14	0.74	0.75	0.62	0.48	1.24	0.38	0.20
Salesperson	0.56	0.66	0.44	0.42	1.15	0.77	0.60	0.30	0.23	1.09	0.38	0.15

6.2 Retail Bank Data Lake: A failure of grounding, not retrieval

Table 5 reports performance on the Retail Bank lake. This domain is distinctive because benchmark items are framed as operational decisions requiring both unstructured policy retrieval and structured table lookups. The XMODE baseline generally achieves high *Policy Recall* (0.85+ for seven of the nine personas, though struggling on the Data Analyst and Salesperson workflows), but *Table Recall* varies significantly across the board. More importantly, final answer accuracy is often much lower than policy recall. This exposes a synthesis gap: while agentic systems reliably retrieve abstract governing rules, they fundamentally struggle to ground those rules against specific operational ledgers to execute a faithful decision. Notably, the zero-shot baseline achieves high accuracy on certain workflows (e.g., 0.88 for the Financial planner). While our semantic filter successfully removes simple factoid lookups, frontier LLMs possess strong parametric priors for standard financial principles, allowing them to often deduce correct operational outcomes without context. However, because this zero-shot reasoning lacks explicit provenance (Faithfulness is strictly zero), it highlights why our benchmark requires grounded synthesis; in an enterprise setting, an accurate but ungrounded answer remains an unauditable liability.

6.3 Drug Data Lake: Joint tabular QA remains challenging

Table 6 reports performance on the Drug data lake, which stresses multimodal question answering across passages and tables. The strongest results appear on Passage questions, where XMODE achieves near-perfect recall (0.98), accuracy (0.97), and faithfulness (0.97). Passage-hop questions remain relatively strong: despite Passage Recall dropping to 0.72, answer accuracy is 0.86, suggesting systems can often synthesize correct answers from partially recovered text. This gap indicates that retrieving the correct table is often not sufficient; the remaining challenge lies in extracting the right fields and composing them into a correct answer. Difficulty increases further for Table-hop questions, where Table Recall drops to 0.48 and accuracy to 0.44, showing that multi-step reasoning over structured biomedical evidence remains substantially harder than single-table lookup.

The mixed Passage-Table family further reinforces this point. Although exact retrieval is very strong in both modalities—0.73 Passage Recall and 0.98 Table Recall—end-to-end accuracy reaches only 0.66. In other words, the main bottleneck is no longer finding the relevant evidence, but correctly integrating heterogeneous evidence across text and tables. Taken together, these results suggest that the Drug data lake is best characterized not as a retrieval-limited setting, but as a cross-modal synthesis setting in which joint reasoning over structured and unstructured biomedical evidence remains the main challenge.

6.4 Cross-Domain Takeaways

Across all three lakes, a consistent narrative emerges: *high retrieval quality alone does not guarantee correct reasoning.* We highlight two major takeaways for future system design:

Table 6: Baseline results for the Drug data lake. Metrics are Passage Recall (R_P), Table Recall (R_T), Accuracy (A), Faithfulness (F), and Token Count (T; k-tokens/query, as in Table 4). Passage/Table map to Single-source lookups and anchor questions, Passage-Table maps to Cross-source, and Passage-hop/Table-hop map to Multi-hop. RESP token counts are estimated full generation tokens per question (approximate).

	XMODE					RESP					Zero-shot	
Question Type	R_P	R_T	A	F	T	R_P	R_T	A	F	T	A	T
Passage	0.98	–	0.97	0.97	0.66	1.00	–	0.99	1.00	0.95	0.45	0.29
Table	–	0.81	0.60	0.76	0.59	–	0.75	0.65	0.70	0.73	0.25	0.23
Passage-Table	0.73	0.98	0.66	0.82	0.66	0.90	1.00	0.79	0.85	0.86	0.47	0.26
Passage-hop	0.72	–	0.86	0.90	0.84	0.65	–	0.86	0.90	1.30	0.57	0.28
Table-hop	–	0.48	0.44	0.62	0.67	–	0.60	0.51	0.63	0.87	0.32	0.26

The Divergence of Discovery vs. Synthesis: By isolating source discovery from synthesis using a controlled context-bundle ablation (App E), we find that failure modes are strictly domain-dependent. When provided with perfect gold evidence, AI/ML accuracy saturates at 98.7%, proving that the AI/ML lake is purely a *discovery* bottleneck. Conversely, Drug lake accuracy only reaches 62.5% even with perfect evidence, demonstrating that cross-modal reasoning over dense biomedical tables remains fundamentally broken in LLMs, independent of retrieval performance.

The Compounding Cost and Latency of Agentic Navigation: While standard single-shot RAG cannot natively navigate the relational topologies of the Retail Bank and Drug lakes, agentic loop-based methods (e.g., XMODE) introduce a compounding token and temporal overhead. Table 4 shows that the zero-shot baseline uses only ˜0.05–0.07 k-tokens per query (˜0.06 on average). In contrast, single-shot RAG requires retrieving and processing substantially larger context, increasing token usage to ˜0.34–0.49 k-tokens per query. Transitioning to iterative agentic navigation compounds this further. As reflected by the proxy token values in Tables 6 and 5, iterative navigation increases token consumption significantly—in the Retail Bank lake, to roughly 1.00–1.27 k-tokens per query (about 5–8× over zero-shot and up to ˜3× over standard RAG). Beyond token footprint, multi-step retrieval loops incur sequential latency bottlenecks, presenting a practical constraint for real-time operational environments like front-line customer service.

7 Limitations

Domain Scope and Synthetic Data: LAKEQUEST evaluates text and tables, excluding dynamic logs and raw audio/visual data (e.g., vision-model caveats rely on text metadata; see App D). Furthermore, the Retail Bank domain relies on synthetic data for privacy, which may alter real-world noise distributions.

Model Monoculture and Evaluator Bias: Using an LLM to synthesize candidate questions risks introducing distributional bias that favors the generator's reasoning patterns. We mitigated this via strict human validation. However, evaluating baselines and rubric-based correctness with the same LLM family (GPT-5) used for synthesis introduces a risk of self-preference bias (Zheng et al., 2023), which may overstate measured baseline performance. Future evaluations using independent frontier models will help separate general architectural bottlenecks from model-specific priors.

Factual Redundancy: Data lakes inherently contain redundant facts. Because LAKEQUEST strictly tracks single-path gold evidence (E_i^*), systems might retrieve valid alternative provenance. Consequently, our retrieval metric serves as a conservative lower bound.

8 Conclusion

We introduced LAKEQUEST, a comprehensive benchmark and reproducible construction framework for grounded question answering over heterogeneous data lakes. Encompassing

AI/ML, retail banking, and biomedical domains, the benchmark provides a shared, human-validated evaluation environment that explicitly pairs natural language questions with precise, modality-aware evidence pointers. Through baseline evaluations, we demonstrated that high retrieval quality alone is insufficient; modern retrieve-and-synthesize architectures fail differently across domains, struggling with relation chaining, policy grounding, and multimodal composition. Ultimately, LAKEQUEST provides the necessary evaluation infrastructure to drive the next generation of reliable, agentic QA systems capable of rigorous discovery and faithful attribution over real-world data lakes.

9 Ethics Statement

LAKEQUEST includes retail banking and biomedical QA tasks where hallucinated or weakly grounded answers pose significant real-world risks. Consequently, this benchmark explicitly evaluates *Faithfulness* alongside accuracy, and human validation ensures that all queries are natively answerable from the retrieved context. However, strong performance on LAKEQUEST does not constitute an endorsement for the autonomous deployment of these systems in high-stakes, real-world decision pipelines. LAKEQUEST is intended as a rigorous evaluation resource, not a substitute for domain expertise or human oversight.

For the human validation, we recruited and paid data annotators using the Prolific platform (Prolific, 2026), ensuring that research ethics were upheld. We acknowledge that ethical considerations are extremely important for any research involving human subjects, and we have ensured that our research meets the Code of Ethics for COLM 2026. Our research plan was reviewed and approved by the ethics review board at our institution. In addition, we took steps to ensure research ethics including: reviewing our institution's protocols for ethical research; only showing suitable content to participants; collecting consent from annotators regarding how the data would be collected, used, and retained; and paying all recruits a fair wage above the minimum wage in the jurisdiction of the recruitment platform.

10 Reproducibility Statement

To ensure the full reproducibility of our five-stage dataset construction pipeline and evaluation framework, we release all artifacts associated with this work. The LAKEQUEST benchmark, including human validation annotations, explicit evidence pointers, and the underlying data lake corpora (subject to their respective licensing), is hosted on Github at `www.github.com/michael0402/LakeQuest`. Furthermore, our complete codebase—containing the automated generation pipeline, human-annotation UI, and all baseline implementations (RAG, XMODE, RESP) with exact prompts—is detailed in App B and will be open-sourced upon publication.

Acknowledgments

We are grateful to the Vector Institute for their support of this work. This research was also supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) through Discovery Grant.

References

Pete Bryan, Giorgio Severi, Joris de Gruyter, Daniel Jones, Blake Bullwinkel, Amanda Minnich, Shiven Chawla, Gary Lopez, Martin Pouliot, Adam Fourney, Whitney Maxwell, Katherine Pratt, Saphir Qi, Nina Chikanov, Roman Lutz, Raja Sekhar Rao Dheekonda, Bolor-Erdene Jagdagdorj, Eugenia Kim, Justin Song, Keegan Hines, Daniel Jones, Richard Lundeen, Sam Vaughan, Victoria Westerhoff, Yonatan Zunger, Chang Kawaguchi, Mark Russinovich, and Kumar Ram Shankar Siva. Taxonomy of failure mode in agentic ai systems. Technical report, Microsoft, 2025. URL `https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Taxonomy-of-Failure-Mode-in-Agentic-AI-Systems-Whitepaper.pdf`.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1026–1036, 2020. doi: 10.18653/v1/2020.findings-emnlp.91.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. Open question answering over tables and text. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Yuhao Deng, Chengliang Chai, Lei Cao, Qin Yuan, Siyuan Chen, Yanrui Yu, Zhaoze Sun, Junyi Wang, Jiajun Li, Ziqi Cao, Kaisen Jin, Chi Zhang, Yuqing Jiang, Yuanfang Zhang, Yuping Wang, Ye Yuan, Guoren Wang, and Nan Tang. LakeBench: A benchmark for discovering joinable and unionable tables in data lakes. *Proceedings of the VLDB Endowment*, 17(8):1925–1938, 2024. doi: 10.14778/3659437.3659448.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss Library. *IEEE Transactions on Big Data*, 12(2):346–361, 2026. doi: 10.1109/TBDATA.2025.3618474.

Mohamed Y. Eltabakh, Mayuresh Kunjir, Ahmed K. Elmagarmid, and Mohammad Shahmeer Ahmad. Cross modal data discovery over structured and unstructured data lakes. *Proceedings of the VLDB Endowment*, 16(11):3377–3390, 2023. doi: 10.14778/3611479.3611533.

Grace Fan, Jin Wang, Yuliang Li, Dan Zhang, and Renée J. Miller. Semantics-aware dataset discovery from data lakes with contextualized column-based representation learning. *Proceedings of the VLDB Endowment*, 16(7):1726–1739, 2023. doi: 10.14778/3587136.3587146.

Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. Aurum: A data discovery system. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pp. 1001–1012, 2018. doi: 10.1109/ICDE.2018.00094.

Hugging Face. Hugging Face Hub API, 2025. URL https://huggingface.co/docs/huggingface_hub/v0.5.1/en/package_reference/hf_api. Accessed 2025-12-01.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, 2021. doi: 10.18653/v1/2021.eacl-main.74.

Zhouyu Jiang, Mengshu Sun, Lei Liang, and Zhiqiang Zhang. Retrieve, summarize, plan: Advancing multi-hop question answering with an iterative approach. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 1677–1686, 2025. doi: 10.1145/3701716.3716889.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Lucy Chin, Seth A Strawbridge, Marysol Garcia-Patino, Ray Kruger, Aadhavya Sivakumaran, Selena Sanford, Rahil Doshi, Nitya Khetarpal, Omolola Fatokun, Daphnee Doucet, Ashley Zubkowski, Dorsa Yahya Rayat, Hayley Jackson, Karxena Harford, Afia Anjum, Mahi Zakir, Fei Wang, Siyang Tian, Brian Lee, Jaanus Liigand, Harrison Peters, Ruo Qi Rachel Wang, Tue Nguyen, Denise So, Matthew Sharp, Rodolfo da Silva, Cyrella Gabriel, Joshua Scantlebury, Marissa Jasinski, David Ackerman, Timothy Jewison, Tanvir Sajed, Vasuk Gautam, and David S Wishart. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 2024. doi: 10.1093/nar/gkad976.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35*, 2022.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina

Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276.

Kin Kwan Leung, Mouloud Belbahri, Yi Sui, Alex Labach, Xueying Zhang, Stephen Anthony Rose, and Jesse C. Cresswell. Classifying and addressing the diversity of errors in retrieval-augmented generation systems. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2026.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33*, 2020.

Farhad Nooralahzadeh, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Multi-modal data exploration via language agents. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 795–813, 2025.

OpenAI. GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025. Version dated August 13, 2025.

Prolific. Prolific Recruitment Platform, 2026. URL https://www.prolific.com/. Accessed 2026-03-31.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016. doi: 10.18653/v1/D16-1264.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486.

Eric W Sayers, Evan E Bolton, Anna M Fine, Christopher Kelly, Sunghwan Kim, Melissa Landrum, Stacy Lathrop, Adriana Malheiro, Terence D Murphy, Lon Phan, Shashikant Pujar, Barton W Trawick, Valerie A Schneider, and Kim D Pruitt. Database resources of the National Center for Biotechnology Information in 2026. *Nucleic Acids Research*, 54(D1): D20–D27, 2026. doi: 10.1093/nar/gkaf1060.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems 36*, 2023.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. MultiModalQA: Complex question answering over text, tables and images. In *9th International Conference on Learning Representations*, 2021.

Nan Tang, Chenyu Yang, Zhengxuan Zhang, Yuyu Luo, Ju Fan, Lei Cao, Sam Madden, and Alon Y. Halevy. Symphony: Towards trustworthy question answering and verification using rag over multimodal data lakes. *IEEE Data Engineering Bulletin*, 48(4):135–146, 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. doi: 10.18653/v1/D18-1259.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36*, 2023.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287, 2021. doi: 10.18653/v1/2021.acl-long.254.

A Detailed Dataset Statistics

This section provides a granular breakdown of the human validation results across all three data lakes. As described in Sec 3.3, all candidate question-answer pairs generated by the synthesis pipeline underwent rigorous human review. The tables below report the total number of synthesized candidates (**Questions**) alongside the number of instances that successfully passed three key human-annotation criteria:

- **Correct:** The reference answer is factually accurate and logically sound.
- **AIS (Attributable to Identified Sources):** The reference answer is strictly entailed by the provided evidence pointers, with no hallucinations or reliance on external knowledge.
- **Answerable:** The question provides sufficient context to be answered exclusively using the retrieved data lake artifacts.

Only instances that passed all criteria were included in the final benchmark.

A.1 AI/ML Data Lake Statistics

Table 7 presents the validation breakdown for the AI/ML repository. In this domain, questions are structurally categorized based on the underlying metadata entities (Model Cards vs. Dataset Cards) and whether the reasoning requires multi-hop entity resolution across the AI/ML graph.

Table 7: Detailed human-annotation results for the AI/ML data lake, categorized by metadata entity types and graph topology.

Question Type	Questions	Correct	AIS	Answerable
Model	897	876	834	815
Dataset	975	925	929	896
Model-Dataset	623	592	570	562
Model-hop	741	701	706	683
Dataset-hop	699	656	648	622
Total	**3935**	**3750**	**3687**	**3578**

A.2 Retail Bank Data Lake Statistics

Table 8 outlines the validation results for the synthetic Retail Bank data lake. Unlike the other domains, Retail Bank QA instances are categorized by operational personas, testing how well models can ground their reasoning in strict policy constraints under different institutional workflows.

Table 8: Detailed human-annotation results for the Retail Bank data lake, categorized by persona-driven operational workflows.

Persona	Questions	Correct	AIS	Answerable
Accountant	549	478	461	454
Auditor	370	321	309	305
Customer service (CS)	328	306	311	303
CS manager	303	293	294	293
Data analyst	375	362	361	360
Financial Planner	228	216	215	215
Fraud analyst	416	391	381	378
Insurance analyst	429	378	376	366
Salesperson	314	302	295	290
Total	**3312**	**3047**	**3003**	**2964**

A.3 Drug Data Lake Statistics

Table 9 details the validation outcomes for the Drug data lake. The task families here stress cross-modal synthesis, dividing questions based on whether the required evidence resides in unstructured scientific text (Passages), structured biochemical records (Tables), or a combination of both.

Table 9: Detailed human-annotation results for the Drug data lake, categorized by modality-specific reasoning requirements.

Question Type	Questions	Correct	AIS	Answerable
Passage	797	768	762	754
Table	776	739	720	701
Passage-Table	447	429	421	410
Passage-hop	706	681	668	646
Table-hop	848	817	802	793
Total	**3,574**	**3434**	**3373**	**3304**

B Prompt Templates

This section details the exact prompt templates used for dataset synthesis, LLM-as-a-judge filtering, and baseline evaluation.

B.1 Candidate Synthesis Prompts

To generate candidate QA pairs, we provided the language model with a sampled evidence subgraph and the following instructions. The system prompt was shared across synthesis settings, while the task prompt was instantiated with minor wording changes depending on whether the evidence subgraph contained a model–dataset pair, multiple related model cards, multiple related dataset cards, or an anchor-plus-positive-plus-distractor setup.

System Instructions:
You are a data scientist working with models and data on huggingface.

Input Context:
{evidence_subgraph}

Task:
I'm going to show you one or more model cards and/or dataset cards. What natural question and answer might you ask that implicitly requires consulting information from the shown evidence subgraph? You should pretend that you cannot see the cards when asking the question (except in the anchored setting, where the anchor card is visible but the supporting and distractor cards are not). If you cannot think of one, then return an empty string. This should be a question that naturally requires consulting information from the provided evidence and should not be a chain of multiple unrelated questions that merely happen to require multiple cards. After the question, provide: (1) a direct answer (short, exact answer), and (2) an explained answer (slightly longer answer with brief justification from the cards). Always answer in the following format:

```
<question>
<direct_answer>
<explained_answer>
```

B.2 Baseline Evaluation Prompts: Agentic Tool Use (XMODE)

For our agentic baselines, we framed the data lake navigation as an iterative tool-use task.

Agent Instructions:
You are an AI agent for data lake question answering. Given a user question, iteratively reason over the question, retrieve the top-k relevant documents from the indexed data lake for the current question or sub-question, and summarize each document with respect to both the original question and the current sub-question. After each retrieval round, decide whether the accumulated evidence is sufficient to answer the original question; if not, propose the next sub-question and continue. We use up to 3 iterations, retrieve $k = 5$ documents per step, and use either similarity search or MMR retrieval. After the final step, generate the answer only from the accumulated evidence.

C Human Annotation Guidelines and Interface

To ensure high data quality, annotators were provided with a custom web interface that displayed the generated question, the synthesized answer, and the exact supporting evidence chunks. Annotators were tasked with independently verifying two criteria: *Answerability* (can the question be answered exclusively using the cited evidence?) and *Attribution* (is the generated answer completely faithful to the cited text/table?). We emphasize that this task evaluates strict textual entailment and reading comprehension based entirely on the provided evidence, rather than requiring annotators to verify external scientific or financial validity.

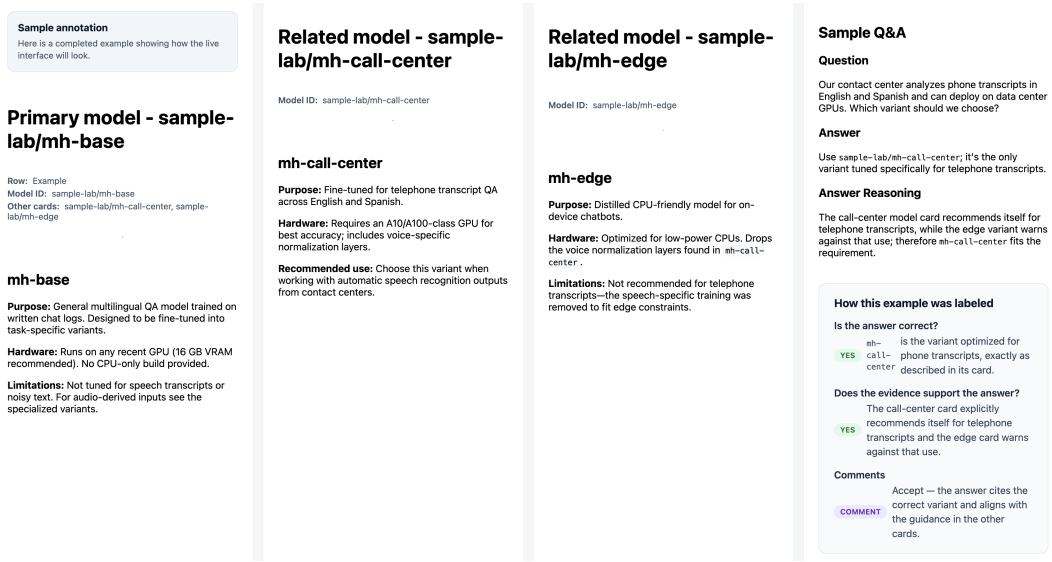

Figure 3: Screenshot of the custom validation UI. Annotators are presented with the highlighted evidence constraints and must explicitly accept or reject the instance based on the provided rubrics.

- **Is the answer correct?** Mark "Yes" only when the answer matches facts stated in the provided evidence.
- **Does the evidence support the answer?** Mark "Yes" only when specific sentences/rows in the provided evidence explicitly support the answer.
- **Any additional comments/thoughts.** Add a short note explaining the decision, especially for "No" labels, and reference the relevant evidence span.
- **Use only provided evidence.** Do not use outside knowledge; if the answer contradicts the provided evidence, mark it incorrect and unsupported.

D Extended Case Studies and Error Analysis

To qualitatively illustrate the quantitative failure modes discussed in Sec 6, we provide representative trace-level examples from the AI/ML and Drug data lakes. The cases below show three recurring error patterns: schema-confusable retrieval, relation chaining failure across closely related model cards, and multimodal metadata composition failure in vision–language datasets.

Table 10: A representative schema-grounding failure in the AI/ML data lake: retrieval finds the right capability class (function calling) but misses the target card's exact token schema.

Domain	**AI/ML (Schema-Confusable Retrieval)**
Question	What prompt format does `0-hero/Matter-0.1-7B` use, and which special tokens does it provide for function calling?
Gold Evidence	The `Matter-0.1-7B` card states that the model uses ChatML prompt format; its Function Calling section lists `<|begin_func|>`, `<|end_func|>`, `<|begin_func_response|>`, and `<|end_func_response|>`.
Gold Answer	ChatML prompt format; function-call tokens `<|begin_func|>` and `<|end_func|>`; function-response tokens `<|begin_func_response|>` and `<|end_func_response|>`.
RAG Predicted Evidence	The retriever surfaced schema-similar tool-use models such as `meetkai/functionary-medium-v3.1` and `meetkai/functionary-small-v3.1`, plus unrelated cards, instead of grounding on the target Matter card.
RAG Predicted Answer	"Prompt format: ChatML-style. Function-calling special tokens: `<|tool_call|>` and `<|tool_response|>`."
Failure Mode Analysis	The retriever found semantically similar function-calling cards, but not the target-specific token schema. The generator then copied the more common `tool_call/tool_response` convention, yielding a partially correct yet still wrong answer.

Table 11: A relation-chaining failure: the system retrieves locally similar model cards but does not gather the exact sibling set required for accurate cross-card comparison.

Domain	**AI/ML (Relation Chaining)**
Question	Among fsicoli's fine-tuned Whisper models (`whisper-small-pt-1000h`, `whisper-medium-pt-1000h`, and `whisper-large-v3-pt-1000h`), which achieves the lowest WER on the shared evaluation dataset, and what are the WERs for each model?
Gold Evidence	The three model cards report WERs of 0.1490 (`whisper-small-pt-1000h`), 0.1147 (`whisper-medium-pt-1000h`), and 0.1113 (`whisper-large-v3-pt-1000h`), all on the same Portuguese speech dataset family.
Gold Answer	`whisper-large-v3-pt-1000h` has the lowest WER (0.1113), followed by `whisper-medium-pt-1000h` (0.1147) and `whisper-small-pt-1000h` (0.1490).
RAG Predicted Evidence	The retriever pulled neighboring Whisper cards such as `fsicoli/whisper-large-v3-pt-cv16` and several unrelated Whisper models, but failed to assemble the three required `pt-1000h` cards into a single comparison set.
RAG Predicted Answer	"I don't see WERs for fsicoli's `pt-1000h` models in the provided context. The only fsicoli metric here is for `whisper-large-v3-pt-cv16` with WER $\approx$ 0.1076."
Failure Mode Analysis	This error is not a pure hallucination; it is a chaining failure. The system retrieved a nearby family member with a very similar identifier, but missed the exact trio needed for cross-card aggregation, so it could not complete the comparison.

Table 12: A multimodal composition failure: the answer depends on jointly grounding license metadata and image-rights caveats across two vision–language dataset cards.

Domain	**AI/ML (Multimodal Composition)**
Question	Do `FreedomIntelligence/ALLaVA-4V` and `FreedomIntelligence/ALLaVA-4V-Chinese` allow commercial use?
Gold Evidence	The `ALLaVA-4V` dataset card lists license `cc-by-nc-4.0`; `ALLaVA-4V-Chinese` lists `apache-2.0`. The `ALLaVA-4V` card also states that the authors "do not own right to any image contained within the `images.zip` file," requiring users to respect original image copyrights.
Gold Answer	No. `ALLaVA-4V` is non-commercial, while `ALLaVA-4V-Chinese` is Apache-2.0; moreover, commercial reuse still requires checking the rights of the underlying third-party images.
RAG Predicted Evidence	The retriever returned image-generation model cards such as `LarryAIDraw/*` and `svjack/Stable-Diffusion-FineTuned-zh-*`, rather than the two ALLaVA dataset cards whose licenses had to be composed.
RAG Predicted Answer	"I can't confirm that from the provided context. If the license is Apache-2.0, MIT, BSD, or CC-BY, commercial use is generally allowed; if it is CC-BY-NC, it is not allowed."
Failure Mode Analysis	The query requires composing metadata from two vision–language dataset cards and incorporating an image-rights caveat. Retrieval drifted toward visually related generative model cards, so the generator defaulted to generic licensing heuristics instead of grounding on the relevant multimodal datasets.

Table 13: Representative Drug failure case from the evidence-provided ablation: the model receives the correct table evidence but still fails to produce the correct answer.

Domain	**Drug (Evidence-Provided Generation Failure)**
Question	Which cited reference mentions "Lorazepam", and does that reference include an ISBN?
Gold Evidence	In `drugbank-drugs_textbooks`, the row with `ref-id=T385` has citation "Ghiasi N. and Marwaha R. (2018). Lorazepam" and `isbn = NaN` (`parent_key=DB00186`).
Gold Answer	`T385`; it does not include an ISBN (missing / NaN).
LLM Generated Answer	"Not determinable from the provided snippets; no citation mentioning 'Lorazepam' is shown, so the ISBN status is unknown."
Failure Mode Analysis	In the context-bundle ablation, the model was given the correct evidence chunk (plus four distractor tables), so retrieval was not the bottleneck. The key row was present in context, but the model failed to attend to and extract it, defaulting to an abstention-style response. This is a generation-under-sufficient-evidence failure rather than a retrieval miss.

E Controlled Context-Bundle Ablation

We ran a fixed context-bundle ablation on the AI/ML, Retail Bank, and Drug data lakes to reduce retrieval variance and better isolate answer synthesis. In each case, the model was given a fixed evidence bundle containing the gold evidence plus relevant distractors, and we evaluated both a direct answer and a reasoning answer using the same LLM-as-a-judge pipeline as in the main experiments.

Table 14: LLM correctness results for the fixed context-bundle ablation across data lakes.

Data Lake	Baseline	Avg. Correctness	Direct Correctness	Reasoning Correctness
AI/ML	Context Bundle	0.9875	0.9750	1.0000
Retail Bank	Context Bundle	0.7500	0.7500	0.7500
Drug	Context Bundle	0.6250	0.6000	0.6500

The results show a sharp cross-domain difference. On AI/ML, performance is nearly saturated once gold evidence is present; Retail Bank is materially higher but not saturated, while Drug remains lowest under the same setup, indicating that synthesis over heterogeneous evidence remains a substantial challenge; reasoning is equal to or stronger than direct answers across lakes.

F LLM-as-a-Judge Consistency Analysis

To assess the stability of our rubric-based LLM-as-a-judge evaluation, we repeated the same AI/ML data lake evaluation three times under an identical setup and report per-question-family correctness in Table 15. The results are highly consistent across runs, with only minor variation across the three repeated tests, suggesting that the judge-based correctness metric is sufficiently stable for the comparative analyses reported in the main paper.

Table 15: Repeated-run LLM correctness on the AI/ML data lake.

Question Type	LLM Correctness – Test 1	LLM Correctness – Test 2	LLM Correctness – Test 3
Model	0.250	0.260	0.245
Dataset	0.140	0.145	0.140
Model-Dataset	0.320	0.340	0.320
Model-hop	0.155	0.165	0.160
Dataset-hop	0.380	0.385	0.375

G Retail Bank Data Lake QA Generation Example

Figure 4 illustrates the end-to-end question-answer synthesis pipeline for the Retail Bank data lake. The process begins by sampling a dense evidence subgraph consisting of a client profile, a transaction ledger, and a governing institutional policy. The generator model is instructed to synthesize a realistic operational query (e.g., a fraud analyst checking a transaction limit) that strictly requires joining facts across all three sampled artifacts. Following generation, the pipeline programmatically filters deictic or ill-formed queries, and an LLM-as-a-judge verifies that the question cannot be answered via parametric memory alone. Finally, the instance undergoes human validation to ensure strict attribution to the highlighted evidence spans.

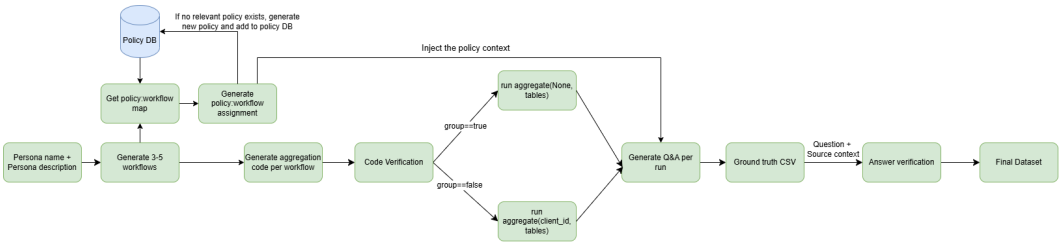

Figure 4: Example of QA generation for the Retail Bank data lake.

H Open Source Model Results Ablation

Table 16: Baseline results for the AI/ML data lake using gpt-oss-20b

	RAG				Zero-shot	
Question Type	R	A	F	T	A	T
Model	0.09	0.29	0.21	0.36	0.20	0.05
Dataset	0.15	0.09	0.11	0.31	0.07	0.05
Model-Dataset	0.12	0.30	0.25	0.35	0.13	0.05
Model-hop	0.13	0.17	0.12	0.43	0.14	0.06
Dataset-hop	0.23	0.28	0.30	0.44	0.18	0.06

Table 17: Baseline results for the Retail Bank data lake using gpt-oss-20b.

	XMODE					RESP					Zero-shot	
Persona	R_T	R_P	A	F	T	R_T	R_P	A	F	T	A	T
Accountant	0.79	0.94	0.50	0.37	1.02	0.95	1.00	0.61	0.40	1.08	0.18	0.15
Auditor	0.84	0.90	0.49	0.42	0.92	0.94	0.85	0.50	0.47	1.00	0.56	0.16
Customer service (CS)	0.82	0.98	0.55	0.53	1.12	0.93	1.00	0.49	0.39	0.96	0.56	0.14
CS manager	0.70	0.87	0.61	0.56	0.95	0.86	1.00	0.77	0.69	0.93	0.57	0.11
Data analyst	0.62	0.69	0.50	0.40	1.00	0.89	0.65	0.46	0.31	0.95	0.24	0.13
Financial planner	0.64	0.99	0.56	0.50	1.10	0.85	0.90	0.58	0.60	0.96	0.72	0.14
Fraud analyst	0.72	0.92	0.52	0.39	0.99	0.95	0.90	0.59	0.50	1.14	0.67	0.13
Insurance analyst	0.52	0.89	0.49	0.37	1.02	0.74	0.75	0.54	0.39	1.10	0.29	0.17
Salesperson	0.56	0.66	0.35	0.33	1.01	0.77	0.60	0.23	0.18	0.97	0.29	0.13

Table 18: Baseline results for the Drug data lake using gpt-oss-20b.

	XMODE					RESP					Zero-shot	
Question Type	R_P	R_T	A	F	T	R_P	R_T	A	F	T	A	T
Passage	0.98	–	0.90	0.91	0.58	1.00	–	0.93	0.94	0.84	0.34	0.24
Table	–	0.81	0.50	0.65	0.52	–	0.75	0.55	0.60	0.64	0.18	0.19
Passage-Table	0.73	0.98	0.56	0.70	0.58	0.90	1.00	0.68	0.75	0.75	0.36	0.22
Passage-hop	0.72	–	0.74	0.79	0.74	0.65	–	0.75	0.80	1.12	0.43	0.24
Table-hop	–	0.48	0.35	0.50	0.58	–	0.60	0.41	0.52	0.75	0.23	0.21