Measuring and Closing the Retrieval Gap in Financial Question Answering

Amine Kobeissi Philippe Langlais

Université de Montréal RALI

Canadian AI 2026
25-29 May 2026, Vancouver, BC

Introduction

Challenges in Financial QA:

- Retrieval quality is rarely measured carefully and most evaluations mask where pipelines fail.
- Retrieve the correct document but fail to locate the right page or evidence passage within it.

Why is this difficult?

Long Documents
Hundred-page filings with repetitive sections and dense tables

Temporal Ambiguity
Similar numerical data across successive annual reports

Complex Queries
Multi-step reasoning across specific passages

Our approach:

- Systematic comparison of retrieval strategies on FinanceBench
- Oracle framework to quantify upper bounds at document, page, and chunk level
- Domain fine-tuned page scorer as a targeted within-document intervention

Key findings:

- Page-level retrieval systematically lags document recall across all baselines
- Domain and task adaptation show promise to close the retrieval gap
- Performance varies substantially by question type with no single strategy performing best across all settings

Problem Formulation & Oracle Framework

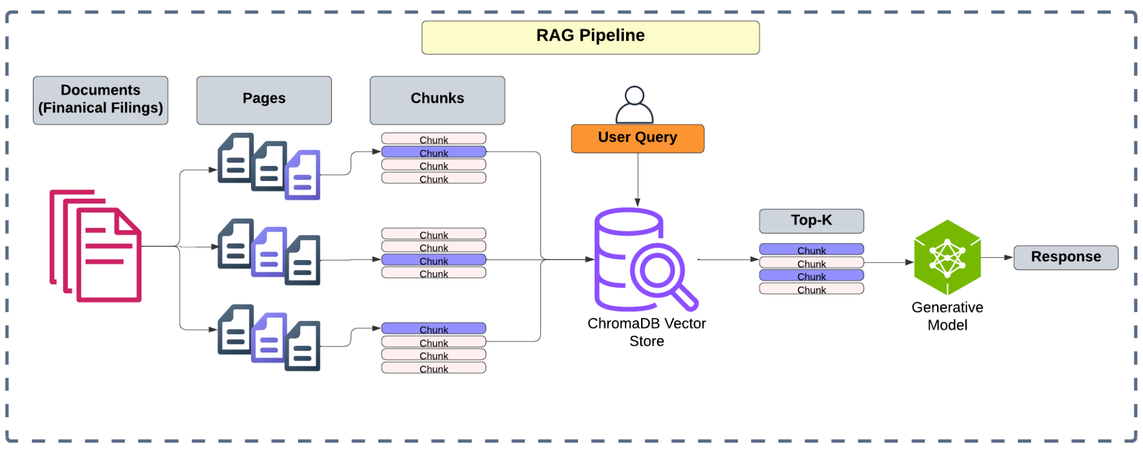

Figure 1. Evidence pages and chunks (blue) are the retrieval targets. Evaluation at document, page, and chunk granularity.

Three oracle conditions isolate error sources by progressively restricting the candidate space:

Standard
Full corpus $\mathcal{C}$

Oracle-Doc
Restricted to gold document
$\{c \in \mathcal{C} \mid d(c) = d^\star\}$

Oracle-Page
Restricted to gold pages
$\{c \in \mathcal{C} \mid d(c) = d^\star, p(c) \in P^\star\}$

Dataset: FinanceBench (Islam et al., 2023)

- **Document types:** 10-K (74.7%), 10-Q (10%), Earnings Calls (9.3%), 8-K (6%)
- **Question types:** Metrics-generated (50), Domain-relevant (50), Novel-generated (50)

Field	Example
Company	Amcor
Document	AMCOR_2020_10K
Question type	Metrics-generated
Question	What is net AR in FY2020 (in USD millions)?
Answer	$1,616.00
Gold page	49
Reference text	Amcor plc and Subsidiaries Consolidated Balance Sheet (in millions). As of June 30, 2020. Current assets: Cash and cash equivalents $742.6 $601.6. Trade receivables, net **1,615.9** 1,864.3. Inventories $1,437.2 $1,561.4.

Retrieval Baselines

- **Dense:** BGE-M3 bi-encoder with cosine similarity
- **Sparse:** BM25 (finance-oriented tokenization), SPLADE
- **Hybrid:** Reciprocal Rank Fusion of BM25 and BGE-M3
- **Parent-Child:** fine-grained chunks mapped to larger parent spans
- **HyDE / Multi-HyDE:** LLM-generated hypothetical passages to bridge lexical mismatch
- **Reranking:** BGE-Reranker-v2-m3 cross-encoder over 20 first-stage candidates

Proposed Method: Domain Fine-Tuned Page Scorer

Motivation: standard chunk retrieval fails to close the retrieval gap.

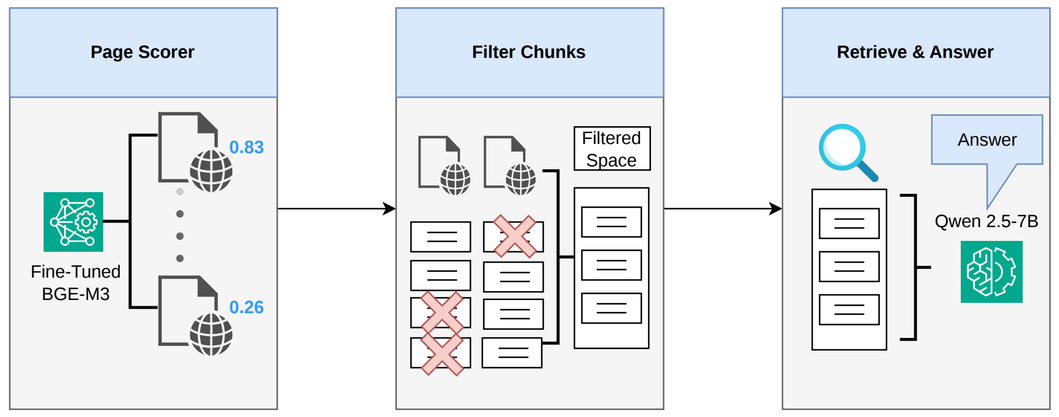

Figure 2. Page Scorer Methedology Pipeline

Results

Retrieval Results at $k = 5$

Method	DocRec ↓ Gold doc found in top-k	PageRec ↓ Fraction of gold pages covered	BLEU ↓ Lexical overlap with gold chunks	ROUGE-L ↓ LCS overlap with gold chunks
BM25	0.32	0.07	0.04	0.12
SPLADE	0.81	0.31	0.24	0.36
Dense (BGE-M3)	0.88	0.34	0.26	0.35
Hybrid (BM25 + BGE-M3)	0.61	0.23	0.13	0.27
Parent-Child	0.91	0.32	0.23	0.34
Dense + Multi-HyDE	0.85	0.42	0.27	0.39
Dense + Reranker	0.87	0.41	0.19	0.34
Dense + Multi-HyDE + Reranker	0.93	0.46	0.28	0.40
Page Scorer (Ours)	**0.95**	**0.55**	**0.33**	**0.46**
Oracle-Doc	*1.00*	*0.60*	*0.25*	*0.42*
Oracle-Page	*1.00*	*1.00*	*0.40*	*0.59*

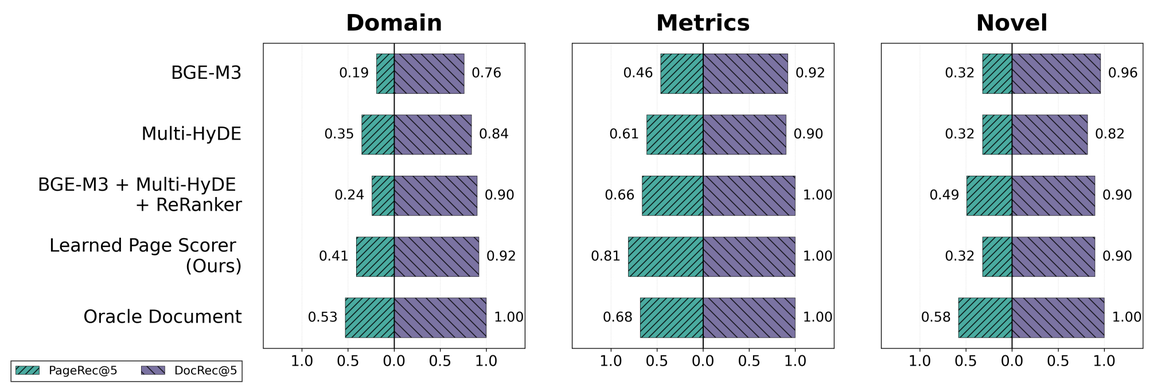

Figure 3. Document and page recall at $k = 5$ by question type (50 questions per type).

Generation Results (Qwen-2.5-7B-Instruct)

Method	ROUGE-L ↓ Lexical similarity to reference answer	Numeric Match ↓ Numerical correctness (±3% tolerance)
Dense (BGE-M3)	0.06	0.24
Dense + Multi-HyDE + Reranker	0.12	0.38
Page Scorer (Ours)	**0.15**	**0.50**
Oracle-Doc	*0.13*	*0.44*
Oracle-Page	*0.16*	*0.70*

Conclusion

- **Oracle framework** decomposes retrieval failures into document discovery, page localization, and chunk selection with empirical upper bounds at each level.
- **Systematic comparison** shows dense retrieval outperforms sparse and Multi-HyDE with reranking improves recall but does not close the within-document gap.
- **Page scorer** shows potential, surpassing the best RAG baseline method and Oracle-Doc on page recall.
- **Heterogeneous difficulty** with no strategy performing best across all question types.

Future work includes retraining the page scorer on external financial corpora, a systematic chunking study, and applying the frameworks to larger benchmarks.

amine.kobeissi@umontreal.ca