Retriever Portfolios: A Principled Approach to Adaptive RAG

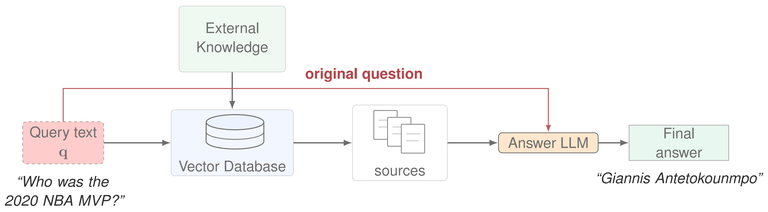
Miltiadis Stouras[1] Vincent Cohen-Addad[2] Silvio Lattanzi[2] Ola Svensson[1]

[1]EPFL [2]Google Research

ICML International Conference On Machine Learning

Paper Code

What is Retrieval-Augmented Generation (RAG)?



In standard deployments, the same knowledge index and retriever setting serve every query, making retrieval a fixed bottleneck when the question type changes.

Query needs are heterogeneous

Multi-hop

Needs a reasoning chain.

"How many NBA titles has the 2023–24 scoring leader won?"

2024 scoring leader → Luka Dončić → NBA titles

Best suited:
Graph / context-expansion retrieval

Diverse evidence

Needs coverage, not redundancy.

"What materials are sports balls made of?"

Best suited:
Diversity-aware retrieval

Specialized vocabulary

Needs domain semantics.

"In NSCLC, which EGFR mutation is associated with resistance to first-generation TKIs?"

Best suited:
Biomedical embedding retriever

No single retriever is uniformly best across these needs.

Can we adapt retrieval efficiently?

Different queries need different retrievers; the challenge is adapting cheaply.

Online adaptive search per query

q → try retriever settings → retrieve → generate → judge → repeat

Accurate, but too costly at inference time.

Retriever portfolio

Offline: choose k retrievers with complementary behaviors.

Online: route to a few of them, retrieve in parallel, and answer.

Question: Which k retrievers should we choose?

How do we measure the quality of a portfolio?

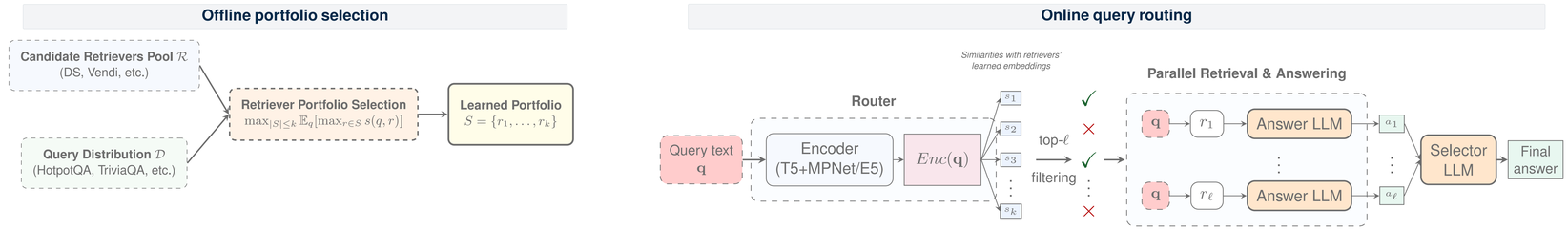
$$F(S) = \mathbb{E}_{q\sim\mathcal{D}}\left[\max_{r\in S} \text{score}(q, r)\right]$$

Promotes complementarity by definition: for each query, at least one member should be strong.

Submodular objective ⇒ greedy efficiently computes a near-optimal portfolio.

The problem is not about finding good retrievers; it's about finding complementary ones.

Method



Experimental setup

Retriever pool

Algorithms
- DiscountedSimilarity (DS)
- Vendi
- GraphDense

+

Dense backbones
- MPNet: sentence-transformers/all-mpnet-base-v2
- E5: intfloat/e5-large-v2

→ **360 total configurations**

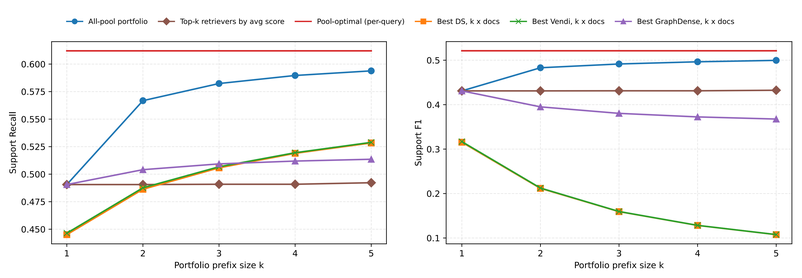
Benchmarks
- HotpotQA
- 2WikiMultiHopQA
- MuSiQue
- TriviaQA

Models / baselines

Models
- Gemma-3-27B-It
- Llama-3.1-70B-Instruct

Baselines
- Nearest-neighbor retrieval
- Best single-family retriever
- Best single retriever fetching more documents
- Vendi-RAG

Portfolios improve retrieval coverage

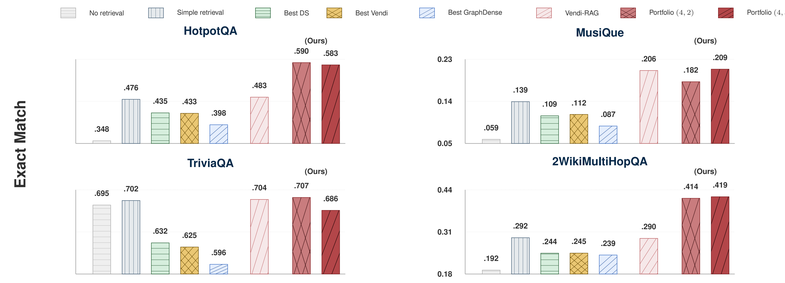


Takeaway. Portfolios outperform all single-retriever baselines, even when those baselines retrieve more documents.

Key point. The gains come from complementarity: top-k retrievers by average score are often redundant, while a learned portfolio of 5 members nearly matches pool-optimal adaptation.

Retrieval gains translate to QA accuracy

Answer Exact Match is calculated for Llama-3.1-70B-Instruct according to the labeled answers of each dataset. All baselines use the same answer model.

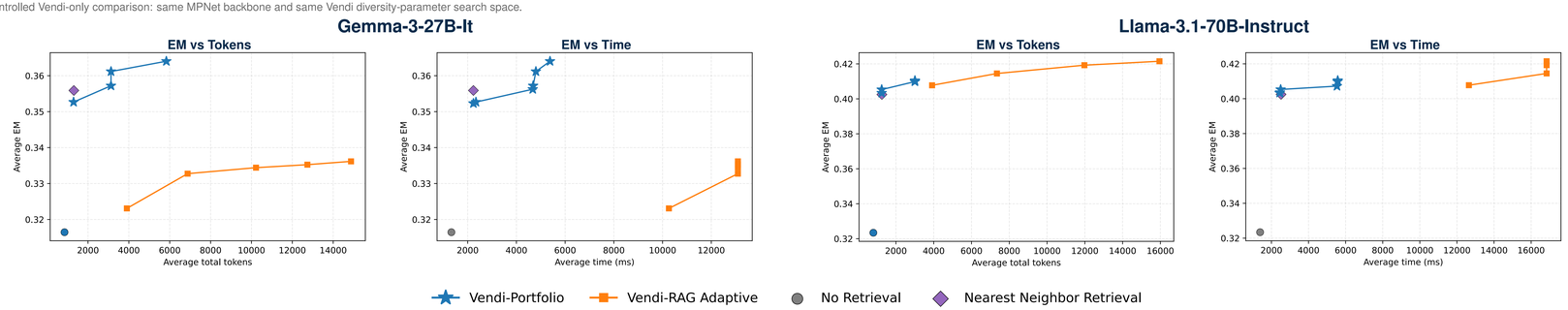


Takeaway. Routed portfolios give the highest Exact Match in all four datasets.

Key point. Complementary retrieval improves final answers, outperforming static baselines and online Vendi-RAG.

Offline adaptation is more efficient

Controlled Vendi-only comparison: same MPNet backbone and same Vendi diversity-parameter search space.



Takeaway. Portfolios avoid expensive per-query search while retaining strong Exact Match.