Test-Time Strategies for More Efficient and Accurate Agentic RAG

Abhinav Sharma[1], Brian Zhang[1], Deepti Guntur[1], Zhiyang Zuo[1], Shreyas Chaudhari[1], Wenlong Zhao[1], Franck Dernoncourt[2], Puneet Mathur[2], Ryan A. Rossi[2], Nedim Lipka[2]

[1]*University of Massachusetts Amherst* [2]*Adobe Research*

Background & Motivation

Agentic RAG frameworks like **Search-R1** interleave reasoning and retrieval to handle complex multi-hop questions, reporting up to **41%** improvement over standard RAG.

Per-turn loop: the LLM emits query q_i, the retriever returns documents D_i, and D_i is appended to the reasoning trace until a final answer a is produced.

However, two recurring failures:

- **Information forgetting** — the model repeatedly retrieves previously seen documents, wasting turns and tokens.
- **Ineffective evidence integration** — retrieved passages are not effectively incorporated into the current reasoning state.

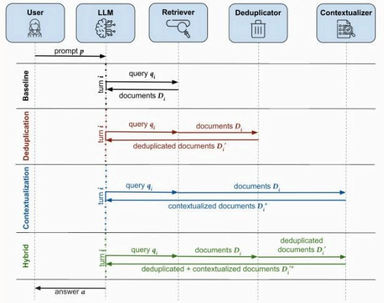
Question: Can lightweight, *training-free* changes at inference time fix both issues without modifying the agent or retriever?

Our Contributions

➤ **Contextualization** — an external LLM extracts task-relevant information from D_i after each retrieval, producing a concise summary D_i^* appended to a persistent memory cache.

➤ **De-duplication** — filters out documents already retrieved in earlier turns, replacing duplicates with the next highest-ranked unseen passage.

➤ **Hybrid** — sequentially applies de-duplication and contextualization, retaining novel evidence while compressing it for downstream reasoning.



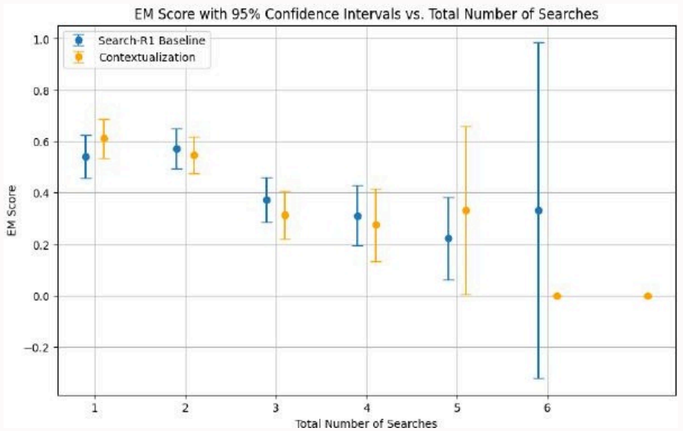
Information flow per turn for the four configurations.

TL;DR: Lightweight test-time modifications to Search-R1 yield a **+5.6% EM** gain and a **−10.5% retrieval-turn** reduction *without retraining* the agent or retriever.

Experiments & Results

Variant	Exact Match	LLM Match	Avg. # Searches
`Qwen2.5-3B base (PPO)`	0.292	0.356	1.410
`Qwen2.5-3B instruct (GRPO)`	0.310	0.396	2.054
`Qwen2.5-7B base (PPO)`	0.464	0.538	2.392
`+ Contextualization (Ours)`	**0.490**	**0.574**	**2.142**
`+ De-duplication (Ours)`	0.478	0.560	2.498
`+ Hybrid (Ours)`	0.480	0.568	2.154

500-question subset of HotpotQA & Natural Questions validation sets.



EM vs. retrieval count. Both methods show a downward trend — harder questions require more turns.

Key Findings

1. Contextualization is the only method that simultaneously improves accuracy *and* reduces retrievals: **+5.6% EM**, **+6.7% LLM**, **−10.5% turns**.

2. De-duplication alone *increases* retrieval count (2.498 vs. 2.392) without proportional accuracy gain — the model just issues new queries to fill the gap.

3. Repeated retrieval is *not* the main bottleneck. The dominant failure is ineffective *use* of evidence already retrieved.

4. LLM Match is consistently 16-18% higher than EM, mostly due to numerical answers and abbreviated names treated as wrong by EM.