Measuring Retrieval Robustness in LLMs, Training Small LLM Agents for End-to-End RAG, and More!

Vol.106 for May 26 - Jun 01, 2025

SUMIT
MAY 30, 2025

Stay Ahead of the Curve with the Latest Advancements and Discoveries in Information Retrieval.

This week's newsletter highlights the following research:

1. Distilling Agentic Behavior into Compact Language Models, from Kang et al.
2. Training Small LLM Agents for End-to-End Retrieval-Augmented Generation, from Viettel Group
3. Self-Evolving Search Agents for Complex Question Answering, from Alibaba
4. Measuring and Understanding Retrieval Robustness in LLMs, from Bloomberg
5. Unifying Online Ad Ranking with Single-Model Architecture, from Meituan
6. Optimizing Query Decomposition for Multi-Vector Retrieval via LLM-Based Prompt Engineering, from Liu et al.
7. A Unified Framework for Hard Negative Mining in Enterprise Knowledge Retrieval, from Oracle AI
8. A Theoretical Analysis of Locality and Entropy in Neural Ranking, from The University of Glasgow
9. A Multi-Dataset Analysis of Chunk Size Effects in Dense Retrieval Systems, from Fraunhofer IAIS
10. A Systematic Framework for Understanding LLM-Based Recommendation Approaches, from NTU

[1] Distilling LLM Agent into Small Models with Retrieval and Code Tools

This paper from Kang et al. introduces Agent Distillation, a framework for transferring the complete task-solving capabilities of LLM agents to much smaller models while preserving their ability to use external tools like code execution and information retrieval. Unlike traditional chain-of-thought (CoT) distillation which only transfers static reasoning patterns, this approach enables small models (0.5B-7B parameters) to learn interactive "reason-act-observe" behaviors from a 32B parameter teacher model. The authors address key challenges in agent distillation through two innovative techniques: first-thought prefix, which improves teacher trajectory quality by incorporating initial CoT reasoning steps into agent prompts, and self-consistent action generation, which enhances student model robustness by sampling multiple action sequences and selecting the most consistent outcomes. Evaluated across eight reasoning benchmarks covering factual knowledge and mathematical reasoning, the distilled small agents consistently outperform CoT-distilled models of equivalent size and remarkably achieve performance comparable to CoT-distilled models that are 2-4 times larger.

📚 https://arxiv.org/abs/2505.17612

🧑🏽‍💻 https://github.com/Nardien/agent-distillation

[2] Agent-UniRAG: A Trainable Open-Source LLM Agent Framework for Unified Retrieval-Augmented Generation Systems

This paper from Viettel Group presents Agent-UniRAG, a trainable framework that unifies RAG systems by leveraging the LLM agent concept to handle both single-hop and multi-hop queries within a single end-to-end system. Unlike previous approaches that either handle query types separately or use classifiers to route queries to different specialized models, Agent-UniRAG employs a unified agent architecture with four key components: a Planning Module that determines when to search for external knowledge or provide final answers using ReAct mechanisms, a Search Tool for knowledge base interaction, a Reflector Module that filters irrelevant information from retrieved documents, and a Working Memory Module that maintains transparency by storing the reasoning process. Experimental results across six benchmark datasets demonstrate that Agent-UniRAG achieves competitive performance with significantly larger models.

📚 https://arxiv.org/abs/2505.22571

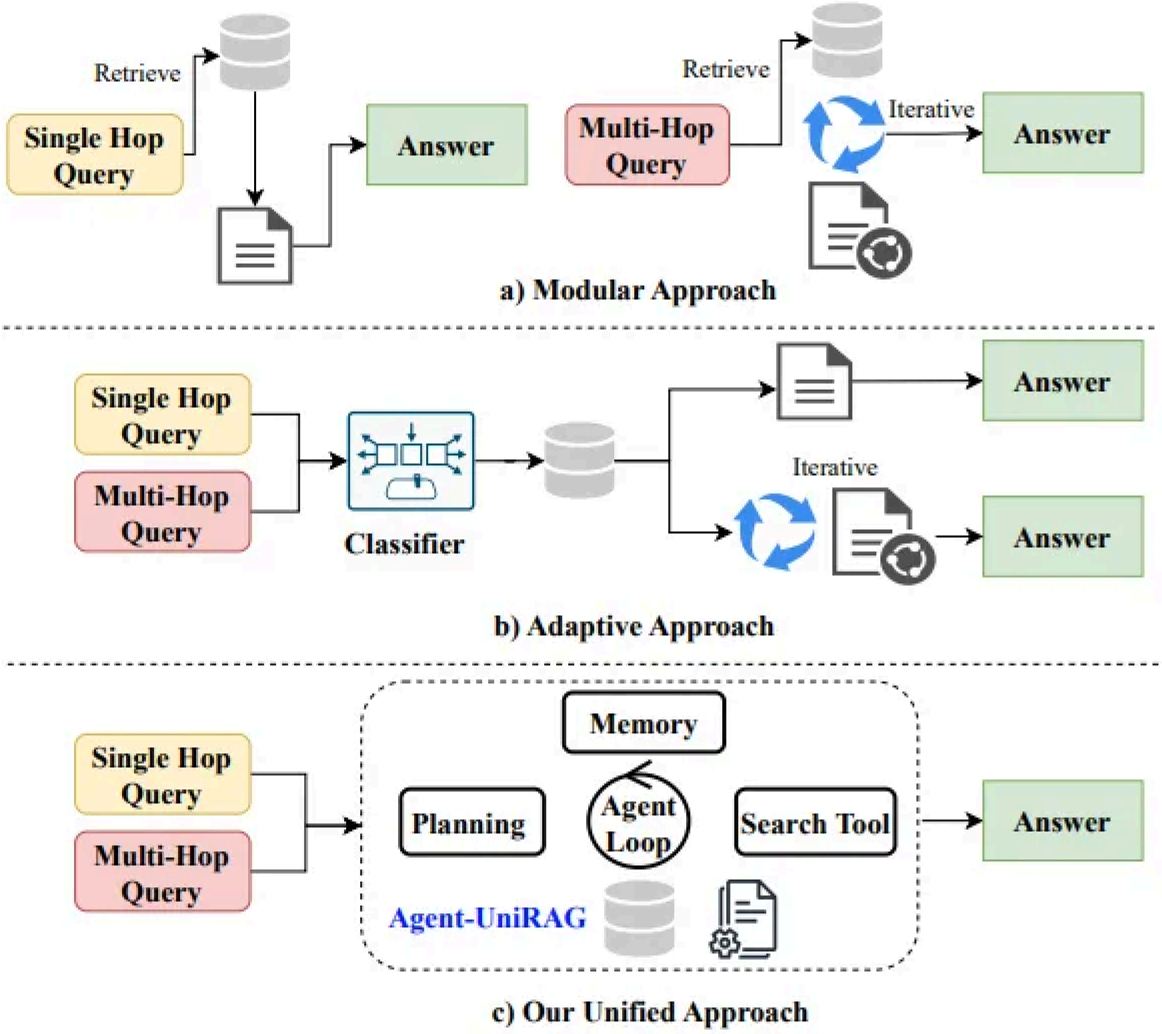


Figure 1: Conceptual analysis of previous works and Agent-UniRAG: (a) Modular approach handles query types separately; (b) Adaptive approach uses a classifier to determine query types before executing them separately; (c) Agent-UniRAG processes all query types within a unified system using the Agent LLM concept.

[3] EvolveSearch: An Iterative Self-Evolving Search Agent

This paper from Alibaba introduces EvolveSearch, an iterative self-evolution framework that combines supervised fine-tuning (SFT) and reinforcement learning (RL) to enhance LLMs' web search capabilities without requiring external human-annotated reasoning data. The approach alternates between two phases: an RL exploration phase where the model interacts with a web search environment using a hybrid reward mechanism (comprising format and answer rewards), and an SFT optimization phase where high-quality rollouts from RL are filtered using three criteria (high-reward selection, same query deduplication, and multi-calls selection) and used to train a better cold-start model for the next RL iteration.

📚 https://arxiv.org/abs/2505.22501

---

[4] Evaluating the Retrieval Robustness of Large Language Models

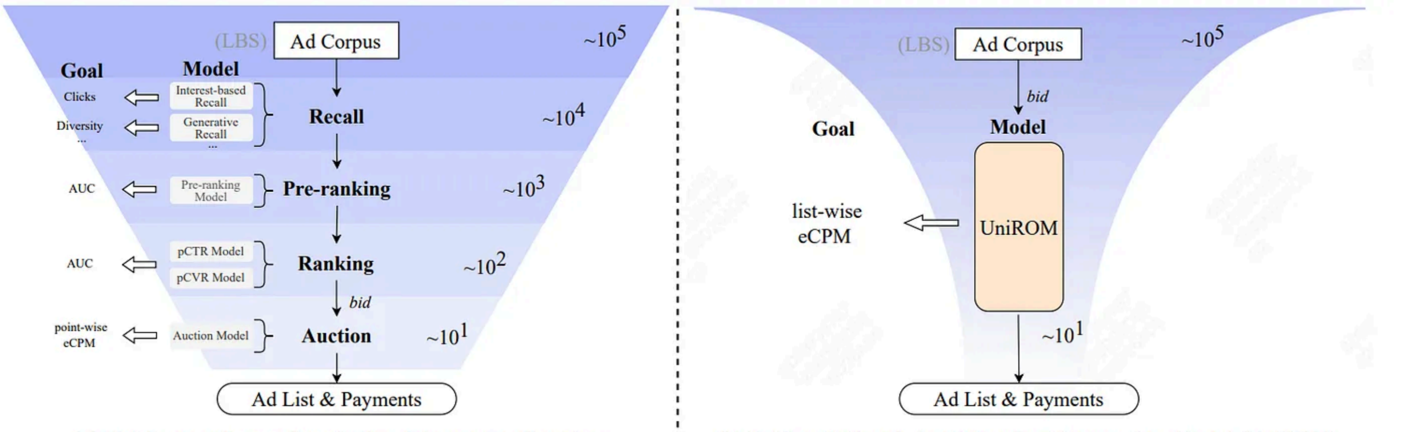
This paper from Bloomberg introduces novel metrics for evaluating the "retrieval robustness" of LLMs in realistic RAG setups, addressing three critical questions: whether RAG consistently outperforms non-RAG approaches, whether adding more retrieved documents improves performance, and whether document ordering affects results. The researchers developed three complementary metrics: No-Degradation Rate (NDR), Retrieval Size Robustness (RSR), and Retrieval Order Robustness (ROR). They compiled a benchmark of 1,500 open-domain questions from Natural Questions, HotpotQA, and ASQA datasets with Wikipedia documents retrieved using both BM25 and dense retrievers. Through comprehensive experiments across 11 LLMs from five families (including Llama, Mistral, Command, GPT-4o/o3-mini, and Claude) using three prompting strategies, they found that models generally demonstrate strong retrieval robustness, achieving over 80% scores on all metrics, with GPT-4o and o3-mini exceeding 90%. However, the research reveals that imperfect robustness leads to concerning sample-level trade-offs where performance improvements on some examples come at the cost of degradation on others, preventing models from fully capitalizing on RAG's benefits and creating unpredictable outcomes that pose risks for real-world deployment.

https://arxiv.org/abs/2505.21870

[5] One Model to Rank Them All: Unifying Online Advertising with End-to-End Learning

This paper from Meituan presents UniROM, an end-to-end generative architecture that unifies online advertising ranking into a single model, replacing the traditional multi-stage cascading architecture (MCA) used in industrial systems. The key innovation addresses two fundamental challenges in current advertising systems: performance inconsistencies arising from divergent optimization targets across different pipeline stages, and failure to account for advertisement externalities. UniROM introduces three main components: a Hybrid Feature Service (HFS) that decouples user and ad feature processing to reduce latency while maintaining expressiveness; RecFormer, which uses an innovative cluster-attention mechanism to efficiently model both user interests and contextual externalities across large candidate pools; and AucFormer, a non-autoregressive generator with permutation-aware evaluation that aligns with advertising platform objectives through reinforcement learning.

https://arxiv.org/abs/2505.19755



(a) Multi-stage Cascading Online Advertising Ranking

(b) Unifies Online Advertising Ranking as One Model (UniROM)

Figure 1: Illustrations of multi-stage cascading architecture and unified online advertising ranking.

[6] POQD: Performance-Oriented Query Decomposer for Multi-vector retrieval

This paper from Liu et al. introduces POQD (Performance-Oriented Query Decomposer), a framework that optimizes query decomposition for Multi-Vector Retrieval (MVR) systems to enhance RAG performance. Traditional MVR approaches like ColBERT decompose queries into individual tokens, but POQD demonstrates that decomposing queries into slightly coarser-grained units like phrases yields superior results for RAG-based question answering tasks. The framework employs two LLMs: a Query Decomposer that generates sub-queries based on optimized prompts, and a Prompt Optimizer that iteratively refines these prompts to maximize downstream performance. To address the non-differentiable nature of the optimization problem and computational challenges of evaluating candidate prompts, POQD uses an alternating training algorithm that optimizes prompts while training downstream models for only a few epochs rather than full convergence. The authors provide theoretical analysis showing this approach effectively minimizes loss with appropriate hyperparameter configurations.

📚 https://arxiv.org/abs/2505.19189

🧑‍💻 https://github.com/PKU-SDS-lab/POQD-ICML25

---

[7] Hard Negative Mining for Domain-Specific Retrieval in Enterprise Systems

This paper from Oracle AI presents a hard negative mining framework specifically designed to enhance domain-specific retrieval in enterprise systems, where overlapping terminologies and semantic mismatches often degrade performance in applications like knowledge management and RAG. The authors propose a scalable approach that integrates multiple embedding models (six diverse bi-encoders), applies PCA-based dimensionality reduction for computational efficiency, and introduces two semantic conditions to dynamically select high-quality hard negatives that are semantically similar to queries but contextually distinct from true positive documents. Evaluated on both proprietary enterprise data and public domain-specific datasets, their method demonstrates substantial improvements compared to state-of-the-art baselines including BM25, In-batch negatives, STAR, and ADORE+STAR. The framework addresses key limitations of existing negative sampling techniques by ensuring selected hard negatives closely resemble query semantics while remaining contextually irrelevant.

📚 https://arxiv.org/abs/2505.18366

---

[8] Disentangling Locality and Entropy in Ranking Distillation

This paper from the University of Glasgow conducts a comprehensive theoretical and empirical investigation of two fundamental components in modern neural ranking systems: hard negative sampling (which selects challenging examples for training) and knowledge distillation (which transfers ranking knowledge from teacher models to students). The authors develop a generalization bound for ranking distillation that separates the effects of locality (geometric properties of the data) and entropy (uncertainty in teacher rankings), demonstrating that these factors have orthogonal influences on model performance. Locality affects the bias term while teacher entropy affects optimization dynamics. Through extensive experiments, they reveal that complex multi-stage hard negative mining pipelines (like those used in SentenceTransformers with up to 12 models) provide minimal benefits over simpler sampling strategies under distillation settings, contradicting conventional wisdom.

📚 https://arxiv.org/abs/2505.21058

🧑🏽‍💻 https://github.com/Parry-Parry/locality-and-entropy

---

[9] Rethinking Chunk Size For Long-Document Retrieval: A Multi-Dataset Analysis

This paper from Fraunhofer IAIS presents a systematic evaluation of fixed-size chunking strategies in RAG systems across six diverse question-answering datasets. The researchers tested chunk sizes ranging from 64 to 1024 tokens using two embedding models (Stella and Snowflake) and found that optimal chunk size depends heavily on dataset characteristics and answer types. Smaller chunks (64-128 tokens) work best for datasets with concise, fact-based answers like SQuAD, while larger chunks (512-1024 tokens) are necessary for datasets requiring broader contextual understanding such as NarrativeQA and TechQA. The study reveals that embedding models exhibit distinct chunking sensitivities. Stella benefits from larger chunks due to its ability to leverage global context from its extensive training on long texts, whereas Snowflake performs better with smaller chunks, excelling at fine-grained entity matching.

📚 https://arxiv.org/abs/2505.21700

🧑🏽‍💻 https://github.com/fraunhofer-iais/chunking-strategies

---

[10] Augment or Not? A Comparative Study of Pure and Augmented Large Language Model Recommenders

This paper from NTU presents a systematic taxonomy for LLM recommender systems, categorizing existing approaches into two distinct branches: Pure LLM Recommenders, which rely solely on LLM capabilities, and Augmented LLM Recommenders, which integrate additional non-LLM techniques to enhance performance. The Pure category encompasses methods like naive embedding utilization (BERT4Rec), pretrained language model fine-tuning (P5, POD), instruction tuning (TALLRec, GenRec), architectural adaptations (LITE-LLM4REC), and reflect-and-rethink approaches that improve through output refinement. The Augmented category includes semantic identifier augmentation using techniques like RQ-VAE-based clustering (TIGER), collaborative modality augmentation that projects traditional collaborative filtering embeddings into language space (CoLLM, LLaRA), prompts augmentation, and retrieve-and-rerank methods. Through experimental evaluations, the authors demonstrate that augmented approaches generally outperform pure LLM methods, with semantic identifiers and collaborative signals providing significant improvements.

📚 https://arxiv.org/abs/2505.23053

🧑‍💻 https://github.com/MiuLab/LMRec-Survey [not public as of 05/30]

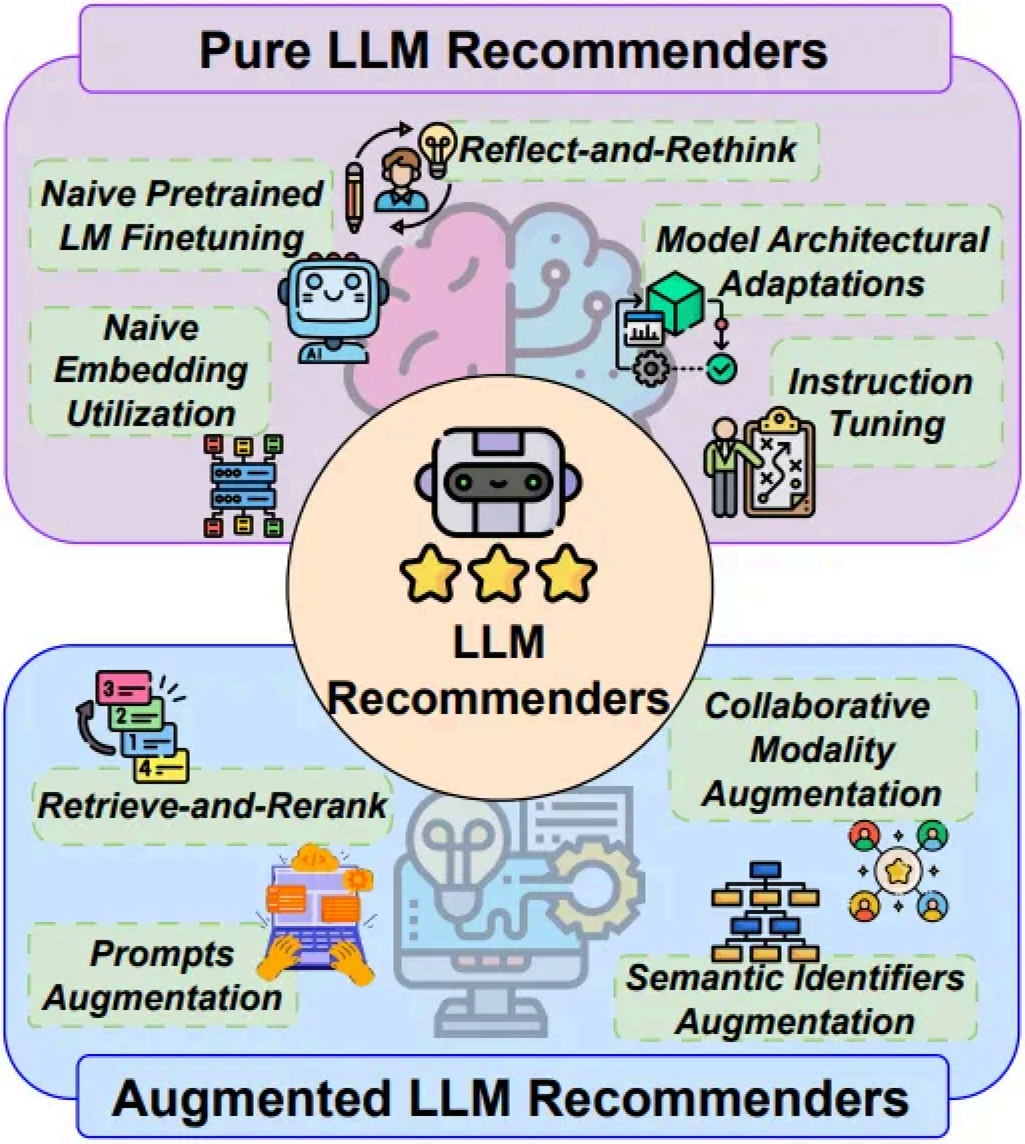

Pure LLM Recommenders
Reflect-and-Rethink
Naive Pretrained LM Finetuning
Model Architectural Adaptations
Naive Embedding Utilization
Instruction Tuning
LLM Recommenders
Retrieve-and-Rerank
Collaborative Modality Augmentation
Prompts Augmentation
Semantic Identifiers Augmentation
Augmented LLM Recommenders

Figure 1: **An illustration of the taxonomy.** LLM Recommenders can be categorized into Pure (up) and Augmented (down) LLM Recommenders, depending on whether they utilize non-LLM techniques to help the final decision making of LLMs.

Extras: Tools

🛠️ RankLLM: A Python Package for Reranking with LLMs

RankLLM is an open-source Python package for document reranking that supports pointwise, pairwise, and listwise approaches using LLMs. It provides a modular and configurable framework compatible with both proprietary and open-source models, enabling integration with various retrieval and inference tools. RankLLM includes optional components for document retrieval via Pyserini, training with Hugging Face, evaluation using standard IR metrics, and analysis of model outputs.

📝 https://arxiv.org/abs/2505.19284

🧑‍💻 https://github.com/castorini/rank_llm/

Extras: Benchmarks

⏱️ VIBE: Vector Index Benchmark for Embeddings

VIBE is an open-source benchmarking suite for evaluating approximate nearest neighbor (ANN) algorithms on modern embedding datasets. VIBE provides a pipeline for generating benchmark datasets using current embedding models and includes both in-distribution and out-of-distribution scenarios. It supports evaluation across 21 ANN implementations and includes performance metrics like recall and query throughput.

📝 https://arxiv.org/abs/2505.17810

🧑‍💻 https://github.com/vector-index-bench/vibe

Extras: Datasets

💾 Yambda-5B -- A Large-Scale Multi-modal Dataset for Ranking And Retrieval

Yambda-5B is a large-scale dataset of music streaming interactions released by Yandex, comprising 4.79 billion user-item events from 1 million users and over 9 million tracks. It includes five types of user feedback: listens, likes, dislikes, unlikes, and undislikes, annotated with timestamps and an is_organic flag that distinguishes organic from recommendation-driven interactions. The dataset also provides audio embeddings derived from a convolutional neural network and is distributed in multiple sizes and formats to support a range of modeling tasks, particularly in ranking and retrieval.

📝 https://arxiv.org/abs/2505.22238

🧑‍💻 https://huggingface.co/datasets/yandex/yambda

💾 Towards Better Instruction Following Retrieval Models

InF-IR is a large-scale training corpus designed for improving instruction-following capabilities in information retrieval models. The dataset contains over 38,000 positive samples structured as <instruction, query, passage> triplets. For each positive triplet, two hard negative examples are generated by modifying either the instruction or the query, with their semantic plausibility and instructional incorrectness validated by a reasoning model.

📝 https://arxiv.org/abs/2505.21439

🧑‍💻 https://huggingface.co/datasets/InF-IR/InF-IR

I hope this weekly roundup of top papers has provided you with valuable insights and a glimpse into the exciting advancements taking place in the field. Remember to look deeper into the papers that pique your interest.

I also blog about Machine Learning, Deep Learning, MLOps, and Software Engineering domains. I explore diverse topics, such as Natural Language Processing, Large Language Models, Recommendation Systems, etc., and conduct in-depth analyses, drawing insights from the latest research papers.

5 Likes · 1 Restack

← Previous

Discussion about this post

Comments Restacks

Write a comment...