Test-time Corpus Feedback: From Retrieval to RAG

Mandeep Rathee
L3S Research Center
rathee@l3s.de

Venktesh V
Stockholm University
venktesh.viswanathan@dsv.su.se

Sean MacAvaney
University of Glasgow
sean.macavaney@glasgow.ac.uk

Avishek Anand
TU Delft
avishek.anand@tudelft.nl

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a standard framework for knowledge-intensive NLP tasks, combining large language models (LLMs) with document retrieval from external corpora. Despite its widespread use, most RAG pipelines continue to treat retrieval and reasoning as isolated components—retrieving documents once and then generating answers without further interaction. This static design often limits performance on complex tasks that require iterative evidence gathering or high-precision retrieval. Recent work in both the information retrieval (IR) and NLP communities has begun to close this gap by introducing adaptive retrieval and ranking methods that incorporate feedback. In this survey, we present a structured overview of advanced retrieval and ranking mechanisms that integrate such feedback. We categorize feedback signals based on their source and role in improving the query, retrieved context, or document pool. By consolidating these developments, we aim to bridge IR and NLP perspectives and highlight retrieval as a dynamic, learnable component of end-to-end RAG systems.

1 Introduction

Large language models (LLMs) augmented with retrieval have become a dominant paradigm for knowledge-intensive NLP tasks. In a typical *retrieval-augmented generation* (RAG) setup, an LLM retrieves documents from an external corpus and conditions generation on the retrieved evidence (Lewis et al., 2020b; Izacard and Grave, 2021). This setup mitigates a key weakness of LLMs—hallucination—by grounding generation in externally sourced knowledge. RAG systems now power open-domain QA (Karpukhin et al., 2020), fact verification (V et al., 2024; Schlichtkrull et al., 2023), knowledge-grounded dialogue, and explanatory QA. RAG has seen widespread adoption in commercial systems: including deployment of RAG for document QA, customer support, and knowledge management. This adoption underscores the practical importance of improving retrieval quality in production settings.

Despite their widespread use, many RAG systems rely on static, off-the-shelf retrieval modules, e.g., BM25 (Robertson et al., 1995) or dense dual encoders (Karpukhin et al., 2020), that are minimally adapted to the downstream task or domain. While re-rankers (Nogueira et al., 2020; Pradeep et al., 2023b) can improve ranking precision, the underlying retrieval often remains brittle in scenarios that demand complex reasoning: multi-hop QA, claim verification, procedural queries, or dialogue-based question answering. These tasks frequently require iterative lookups, query decomposition, or high-precision evidence, capabilities that static retrieval pipelines lack.

In contrast to the prevailing view of retrieval as a fixed first step, a growing body of work in the IR community treats retrieval as a *feedback-driven, adaptive process*, where signals from the output stage are used to guide when to retrieve, how to reformulate queries, and which evidence to include.

Definition

In this survey, we define **feedback** in RAGs as any signal, derived from the *corpus* at different levels – retrieval, ranking, or generation components. This feedback is used to improve the query, the context used for generation, or the set of retrieved documents.

Feedback vs. Traditional Relevance Feedback. Our notion of feedback differs fundamentally from classical IR relevance feedback (Rocchio, 1971). Traditional relevance feedback is an *interactive, user-driven* process: users explicitly mark documents as relevant or non-relevant, and the system updates query representations (e.g., via Rocchio

Findings of the Association for Computational Linguistics: EACL 2026, pages 5637–5656

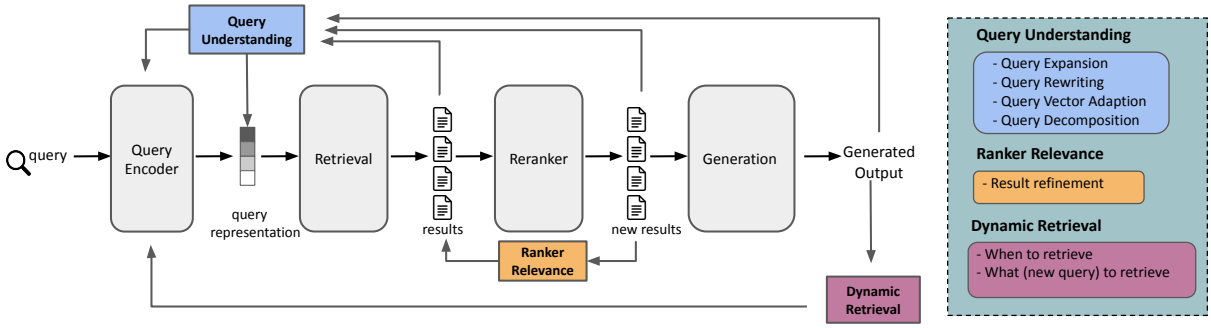

Figure 1: Illustration of feedback signals across the RAG pipeline. Feedback can modify the query (e.g., rewriting), the retrieved pool (e.g., ranker-based expansion), or the generation loop (e.g., retrieval triggers based on uncertainty).

algorithm) to improve subsequent retrievals. In contrast, feedback in modern RAG systems is *automatic, multi-stage, and model-driven*: it leverages signals from retrieval confidence, generation quality, semantic consistency, or intermediate reasoning steps, without explicit user judgments.

We note that such feedback may be applied in one or multiple rounds and can originate from internal model signals (e.g., uncertainty or confidence), external modules (e.g., rankers or verifiers), or user behavior (e.g., clicks or clarifications). Our notion of corpus feedback or simply feedback arises at three key stages:

1. **Query-level feedback** , where the input query is rewritten, expanded, or decomposed using model introspection or relevance signals (refer to Section 3).

2. **Retrieval-level feedback**, where rankers or corpus structure are used to revise or expand the document pool across rounds (refer to Section 4).

3. **Generation-time feedback**, where confidence, or verifier critiques trigger new retrievals or corrections (refer to Section 5).

Figure 1 shows an overview of these feedback stages. This survey synthesizes recent work that operationalizes these feedback signals across RAG pipelines. We organize methods by where and how feedback is applied, not by architecture or dataset, emphasizing how feedback improves retrieval adaptively rather than statically. There exist other surveys like, Retrieval methods survey (Hambarde and Proença, 2023), RAG survey (Gao et al., 2024), Reasoning RAG survey (Li et al., 2025c), Retrieval and Structuring augmented generation survey (Jiang et al., 2025b), and the Agentic RAG survey (Singh et al., 2025) covering retrieval techniques, prompting, reasoning, or agentic methods. These surveys provide broad overviews of the field, but they do not focus specifically on **test time corpus feedback**. Our survey is uniquely positioned to fill this gap between the advancements made by the IR research to improve search quality and RAG techniques developed in NLP research, for example, when to retrieve and what to retrieve. We provide a structured taxonomy of feedback mechanisms organized by *signal type* (query-level, retrieval-level, generation-time), bridging IR and NLP perspectives.

Our scope is deliberately focused on retrieval-centric innovations in RAG. We do not cover standalone prompting or answer-generation strategies unless they directly influence the retrieval component. Our goal is to help NLP researchers treat retrieval as a dynamic, learnable component—just as vital as the generator, especially in tasks that require reasoning over incomplete, multi-part, or contextual knowledge. We also review the experimental landscape for retrieval-centric RAG in Section 6: common benchmarks, evaluation metrics, and emerging standards for assessing retriever quality in knowledge-intensive tasks. By consolidating these developments, this survey attempts to bridge the gap between information retrieval and NLP communities, highlighting how feedback can drive the next generation of retrieval-aware, reasoning-capable RAG systems.

2 Preliminaries

2.1 Retrieval System

The core objective of a retrieval system is to identify and rank a subset of documents $(d_1, d_2, ..., d_k)$ from a large corpus $\mathcal{C}$ based on their estimated relevance to a query q. Classical retrieval approaches, such as BM25 (Robertson et al., 1995), rely on exact term matching and produce sparse relevance scores. In contrast, dense retrieval methods employ

Cat.1	Approach	Approach Description
Query Level	**Lexical Pseudo Relevance Feedback**	
	Lexical PRF (Jaleel et al., 2004)	Expand queries using top-k document terms
	Rocchio (Rocchio, 1971)	Adjust vector using relevant feedback
	KL Expansion (Zhai and Lafferty, 2001)	Optimize query based on feedback documents
	Adaptive Relevance Feedback (Lv and Zhai, 2009)	Adaptive weights per query and feedback set
	Relevance Modeling (Metzler and Croft, 2005)	Interpolate query with new expansion terms
	LCE (Metzler and Croft, 2007)	Discover latent concepts for expansion
	LCA (Xu and Croft, 1996)	Use co-occurrence statistics for expansion
	Semantic Pseudo Relevance Feedback	
	EQE (Zamani and Croft, 2016)	Words with similar embeddings are used in query expansion
	RLM/RPE (Zamani and Croft, 2017)	Train a models to output words relevance
	ANCE PRF (Yu et al., 2021)	Expand using contrastive dense embeddings
	Colbert PRF (Wang et al., 2023b)	Contextual embedding expansion with late interaction
	Generative Relevance Feedback	
	GRF (Mackie et al., 2023)	Generate contexts with LLMs for queries
	GAR (Mao et al., 2021)	Expand using answer and passage metadata
	QueryExpansion (Jagerman et al., 2023)	Prompt-based query rewriting techniques
	LameR (Shen et al., 2024)	Append generated answers to original query
	InteR (Feng et al., 2024)	Alternate between generation and retrieval
	Iter-Retgen (Shao et al., 2023)	Interplay between GAR (or GRF) and RAG to improve answer generation
	BlendFilter (Wang et al., 2024)	Use both LLM-generated query and original query for retrieval
	RRR (Arora et al., 2023)	Interplay between GAR (or GRF) and RAG to improve retrieval
	MILL (Jia et al., 2024)	Use both PRF and GRF for query expansion
	ReAL (Chen et al., 2025a)	Learn original and expanding query terms weights
	Word2Passage (Choi et al., 2025)	Use granular word-level importance for query expansion
	DeepRetrieval (Jiang et al., 2025a)	RL training to optimize the rewritten query
Retrieval Level	GraphAR (MacAvaney et al., 2022)	Adaptive retrieval using a corpus graph
	LADR (Kulkarni et al., 2023)	Use lexical results for dense retrieval
	QUAM (Rathee et al., 2025b)	Adaptive retrieval using doc-doc similarities as feedback
	LexBoost (Kulkarni et al., 2024)	Improve lexical retrieval using semantic corpus graph
	SUNAR (V et al., 2025)	Use answer uncertainty as feedback
	ORE (Rathee et al., 2025c)	Dynamic documents selection for ranking
	SlideGAR (Rathee et al., 2025a)	Use LLM-based listwise ranker's feedback for adaptive retrieval
	ReFIT (Gangi Reddy et al., 2025)	Update query vector using Ranker feedback
	TOUR (Sung et al., 2023)	Update query representation using ranker feedback
Generation-Time	**Rule-Based Retrieval**	
	SKR (Wang et al., 2023c)	Ask LLM if information needed
	IRCoT (Trivedi et al., 2023)	Retrieve if CoT has not provided the final answer
	Adaptive RAG (Jeong et al., 2024)	Classifier's feedback for retrieval
	Retrieval-on-Demand via Feedback Signals	
	FLARE (Jiang et al., 2023)	Token probability as feedback
	DRAD (Su et al., 2024a)	Check hallucination in answer and trigger retrieval to mitigate
	DRAGIN (Su et al., 2024b)	Token probability as feedback
	Rowen (Ding et al., 2024)	Answer consistency as feedback
	SeaKR (Yao et al., 2025)	Internal states of the LLM as feedback
	CRAG (Yan et al., 2024)	Use retrieval evaluator to judge if context is relevant
	CoV-RAG (He et al., 2024)	Chain-of-Verification using a trained model
	SIM-RAG (Yang et al., 2025b)	External critic model to judge if context is sufficient
	Prompt-Based Methods	
	Self-Ask (Press et al., 2023)	Decompose the complex query into sub-queries
	DeComP (Khot et al., 2023)	Decompose complex query into sub-queries
	ReAct (Yao et al., 2022)	Use each reasoning step to trigger retrieval
	Searchain (Xu et al., 2024)	Generate chain-of-questions and trigger retrieval if needed
	MCTS-RAG (Hu et al., 2025)	Dynamically integrates reasoning and retrieval in MCTS
	SMR (Lee et al., 2025)	Mitigates overthinking in retrieval by guiding LLMs through discrete actions
	Learned or Agentic Methods	
	Self-RAG (Asai et al., 2024)	Train LLM to predict reflection tokens that trigger retrieval and judge context
	DynamicRAG (Sun et al., 2025b)	Trains LLM for dynamically selecting documents
	Search-R1 (Jin et al., 2025)	Train LLM to decompose query and generate tokens that trigger retrieval
	Search-O1 (Li et al., 2025a)	Decide autonomously when to retrieve by detecting the presence of uncertain words
	R1-Searcher (Song et al., 2025)	Reward for triggering search tokens
	ReZero (Dao and Le, 2025)	Introduces an RL framework that rewards the act of retrying search queries
	DeepResearcher (Zheng et al., 2025)	Use F1 score-based reward for answer accuracy
	WebThinker (Li et al., 2025b)	Adapt model to use commercial search engines during training
	ZeroSearch (Sun et al., 2025a)	Approximate the real search engine behavior during training

Table 1: Summary of feedback-based retrieval and RAG methods.

neural encoders to project queries and documents into a shared embedding space, enabling semantic similarity matching (Karpukhin et al., 2020; Shao et al., 2025). There have been proposed hybrid retrieval techniques (Cormack et al., 2009; Bruch et al., 2023) as well, which use both lexical and semantic query-document similarities.

Since first-stage retrievers often produce noisy candidates, modern pipelines incorporate a second-stage *re-ranking* step using more expressive models. This includes LLM-based rankers (Pradeep et al., 2023b; Ma et al., 2024; Sun et al., 2023) and reasoning-augmented models such as Rank-1 (Weller et al., 2025), and Rank-R1 (Zhuang et al., 2025), which refine the initial rankings by modeling deeper interactions between the query and candidate documents.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020b) is a hybrid paradigm that enhances the generative capabilities of large language models (LLMs) by incorporating non-parametric external knowledge during inference. This design mitigates well-documented limitations of standalone LLMs, including hallucinations, confident but incorrect outputs, and inability to reflect up-to-date or domain-specific information due to static pretraining (Hurst et al., 2024; Grattafiori et al., 2024; Yang et al., 2025a).

RAG introduces an explicit retrieval step: for a query q, a retriever selects a set of top-k documents $\{d_1, \ldots, d_k\}$ from an external corpus. A generator G then conditions on both q and the retrieved context to produce the output $a = G(q, d_1, \ldots, d_k)$ where G is typically an encoder-decoder or decoder-only LLM fine-tuned to integrate retrieved evidence into its generation process.

2.3 Challenges in RAG

A central challenge in RAG is that generation quality is tightly coupled with retrieval quality. This manifests in several failure modes:

Hallucination: When relevant documents which potentially contain the correct answers are not retrieved, the generator may produce plausible but factually incorrect content, confidently filling gaps with parametric knowledge (Cuconasu et al., 2024).

Information Omission: Missing key evidence leads to incomplete answers, where critical facts are absent despite being available in the corpus (Cuconasu et al., 2025).

Distraction by Noise: Including irrelevant or contradictory documents causes the generator to lose focus on pertinent information, degrading answer quality even when relevant documents are present (Liu et al., 2024).

Attribution Failure: The model may correctly retrieve relevant documents, but fail to properly ground its generation in them, instead relying on memorized patterns in parametric memory.

Consequently, improving the top-k retrieval results is crucial. This can be viewed both as a *selection* problem (how to retrieve the most relevant documents) and a *filtering* problem (how to suppress distracting or noisy context). To this end, several methods have been proposed that incorporate various forms of *feedback*, ranging from simple lexical overlap to more sophisticated agentic or reasoning-based signals, to guide and refine the retrieval process.

In this survey, we systematically categorize these feedback mechanisms and analyze their effectiveness across different components of the RAG pipeline. We focus on how feedback is acquired, represented, and integrated into retrieval, with the aim of providing a comprehensive taxonomy and highlighting open research challenges.

3 Query-level feedback

We first focus on the first aspect which feedback in RAG systems impact – *the query*. A fundamental factor influencing the performance of RAG systems is indeed the formulation of the input query. Poorly phrased, underspecified, or ambiguous queries can lead to irrelevant retrieval, ultimately degrading the quality of the generated output.

To address this, a variety of *feedback-driven* query reformulation methods have been proposed. Feedback may be applied in one or multiple rounds to iteratively enhance retrieval effectiveness and overall answer quality. In this section, we focus on the feedback improving query representations and categorize them into two broad families based on the source and representation of feedback: (i) pseudo-relevance feedback from retrieved documents, and (ii) generative relevance feedback from large language models.

3.1 Pseudo-Relevance Feedback (PRF)

Pseudo-relevance feedback (PRF) techniques modify queries based on the content of top-k retrieved documents, assumed to be relevant. These methods operate either in the lexical space or in dense embedding spaces.

Lexical PRF. Classical PRF methods such as RM3 (Jaleel et al., 2004), Rocchio (Rocchio, 1971), and KL-divergence-based models (Zhai and Lafferty, 2001) extract high-frequency terms from pseudo-relevant documents to expand the original query. These approaches rely on exact term matching and term frequency statistics. Enhancements like Latent Concept Expansion (LCE) (Metzler and Croft, 2007) and Local Context Analysis (LCA) (Xu and Croft, 1996) leverage co-occurrence patterns or latent topic structures but still operate in the discrete term space. While effective for certain domains, these methods are limited by *the vocabulary mismatch problem*: relevant documents may not share terms with the query, especially in low-resource or noisy scenarios.

Semantic PRF. To address lexical mismatch (Zamani and Croft, 2016, 2017) use word embeddings to expand queries with semantically related terms. More recent techniques adopt dense retrieval settings – (Wang et al., 2023b) performs feedback-based expansion in contextualized token embedding space, while ANCE-PRF (Yu et al., 2021) averages document embeddings to interpolate with the query vector. These methods enable richer semantic matching but *remain sensitive to the ambiguity* or sparsity of the original query (Jagerman et al., 2023).

Key Insights

Expands queries using information from top-ranked documents, improving recall but still limited by vocabulary mismatch, query ambiguity, and noise in initial retrievals.

3.2 Generative Relevance Feedback (GRF)

Generative relevance feedback (GRF) methods employ large language models (LLMs) to generate query expansions, reformulations, or conceptually enriched representations. Unlike PRF, where feedback is extracted from retrieved documents, GRF generates feedback via prompting, generation, or learning signals.

LLM-only Feedback. Several methods prompt pre-trained LLMs to produce reformulated or expanded queries. QueryExpansion (QE) (Jagerman et al., 2023) employs different prompting styles, including Chain-of-Thought (CoT) prompting (Wei et al., 2022), to elicit stepwise explanations and derive new query terms. While these methods can function without initial retrieval, many still use retrieved documents as input, which can introduce noise. Hybrid systems such as MILL (Jia et al., 2024), GRF+PRF (Mackie et al., 2023), and Blend-Filter (Wang et al., 2024) combine lexical PRF and GRF by verifying consistency between generated expansions and retrieved evidence. Word-level filtering methods like Word2Passage (Choi et al., 2025) and ReAL (Chen et al., 2025a) further refine queries using token importance estimates.

Feedback from Generated Answers. Beyond generating expansions, some methods use LLM-generated answers as implicit feedback. Generation-Augmented Retrieval (GAR) (Mao et al., 2021) generates answer-like contexts (titles, passages, summaries) using a model like BART (Lewis et al., 2020a), which are then concatenated to the query. However, this introduces risks of hallucination and irrelevant additions. To refine this idea, RRR (Arora et al., 2023) iteratively updates the query based on retrieval performance, using a feedback loop constrained by a document budget. LameR (Shen et al., 2024) first generates multiple answers, augments them with the query, and performs a second retrieval pass—effectively building a feedback loop from generation to retrieval. InteR (Feng et al., 2024) and Iter-RetGen (Shao et al., 2023) perform tighter integration between RAG and GAR by alternating between generation and retrieval for iterative refinement.

Optimization-based Feedback. Recent work aims to move beyond prompting heuristics by directly optimizing queries for retrieval objectives. DeepRetrieval (Jiang et al., 2025a) introduces a reinforcement learning framework where the query generation process is trained end-to-end to maximize retrieval metrics (e.g., recall, nDCG), using document-level reward signals. This eliminates reliance on manual prompting or ground truth supervision.

We refer readers to comprehensive surveys such as (Song and Zheng, 2024) for broader coverage of query rewriting and optimization techniques be-

yond the RAG context.

Key Insights

GRF methods use LLMs to generate or optimize query reformulations, offering richer semantics and adaptability, but are prone to hallucination (irrelevant but plausible-sounding terms) and require strategies to control noise.

4 Retrieval-level feedback

Retrieval in RAG pipelines is often bottlenecked by the bounded recall of the first-stage retriever. Once the top-k documents are selected, re-ranking can improve their ordering, but cannot recover relevant documents missed in the initial retrieval. This limitation motivates *adaptive retrieval* methods that incorporate feedback, often from neural rankers or structural knowledge of the corpus, to refine or expand the retrieved document set across one or more rounds. In this section, we examine two prominent classes of adaptive retrieval strategies, *neighborhood-based corpus expansion* and *query vector adaptation*.

Neighborhood-based Corpus Expansion relies on the *clustering hypothesis* that posits that co-relevant documents tend to be similar to one another. GraphAR (MacAvaney et al., 2022) formalizes this intuition by constructing a corpus graph using lexical similarity between documents. After reranking an initial retrieved set, the method expands the document pool by including neighbors of top-ranked documents in the graph. Variants such as LADR (Kulkarni et al., 2023) and LexBoost (Kulkarni et al., 2024) improve efficiency by using dense bi-encoders and incorporating query-document and document-document edges. QUAM (Rathee et al., 2025b) generalizes these approaches by introducing query affinity modeling, taking into account the degree of similarity between neighbors and their relevance. The ORE framework (Rathee et al., 2025c) further refines this strategy by prioritizing expanded documents based on their expected utility toward the ranker's final relevance. SUNAR (V et al., 2025) incorporates uncertainty over multiple LLM-generated answers to adjust retrieval weights, offering a feedback loop grounded in generation uncertainty, though it may amplify hallucinated answers. SlideGAR (Rathee et al., 2025a) uses LLM-based listwise rankers (Pradeep et al., 2023b,a) to iteratively expand and refine the document pool over a document graph, closing the loop between ranking, selection, and feedback-driven retrieval.

Query Vector Adaptation updates the query representation based on feedback from ranked documents. ReFIT (Gangi Reddy et al., 2025) and TOUR (Sung et al., 2023) both adjust the query vector in dense retrieval space using intermediate relevance scores from neural rankers. These adapted queries are used to perform second-stage retrieval, improving coverage of relevant documents.

Key Insights

Relevance feedback improves recall via efficient corpus expansion or query adaptation, but risks adding noise when similarity links or feedback are unreliable.

5 Generation-time feedback

RAG systems face two fundamental challenges: determining *when to retrieve* external knowledge and *how to retrieve* relevant content effectively (Su et al., 2024b). Classical RAG pipelines follow a fixed sequence of retrieval, ranking, and generation, limiting their adaptability. Recent work introduces *adaptive RAG*, where retrieval strategies are dynamically adjusted based on query characteristics, model feedback, or task complexity. We categorize these approaches into three main classes.

5.1 Rule-Based and Discriminative Approaches

These methods use fixed heuristics or external models to guide retrieval. In-Context RALM (Ram et al., 2023) retrieves documents at regular token intervals, while IRCoT (Trivedi et al., 2023) interleaves retrieval within chain-of-thought frameworks (Wei et al., 2022). However, these strategies often perform retrieval regardless of necessity, leading to latency overheads and noisy context. To overcome over-retrieval limitations, retrieval-on-demand approaches trigger retrieval only when needed, based on external feedback or LLM assessment.

Key Insights

Rule-based methods help in answer generation but result in over-retrieval, latency costs, and noisy context that can lead to wrong answers.

5.2 Retrieval-on-Demand via Feedback Signals

These methods trigger retrieval based on feedback signals from answer uncertainty, model internal states, or context quality. SKR (Wang et al., 2023c) asks the LLM itself if additional information is needed to answer the query. If yes, retrieval is triggered; otherwise, the answer is generated from the LLM's internal knowledge. However, SKR's judgment is solely based on the LLM without external validation, and LLMs tend to be overconfident in their assessments without context (Xiong et al., 2024), potentially missing critical retrieval opportunities. To address the overconfidence issue, FLARE (Jiang et al., 2023) uses token probability as an objective signal, retrieving documents only if the probability falls below a predefined threshold. It uses the last generated sentence (excluding uncertain tokens) as a query for retrieval. While this reduces overconfidence bias, FLARE suffers from two limitations: first, not all uncertain tokens are equally important for triggering retrieval; second, using only the last sentence for query formulation may miss broader contextual information needs. Building on FLARE's limitations, DRAD (Su et al., 2024a) introduces an external hallucination detection module to identify hallucinated entities in generated answers, triggering retrieval when hallucinations are detected. The last generated sentence (without hallucinated entities) is used as the retrieval query. However, DRAD's query formulation still relies on heuristic strategies, limiting its ability to capture the model's true information needs. DRAGIN (Su et al., 2024b) addresses this by reformulating queries using keywords extracted from the model's internal attention weights and reasoning, providing a more principled approach to query generation. Like FLARE, it uses token probabilities to trigger retrieval but excludes uncertain tokens from the query itself. SeaKR (Yao et al., 2025) takes a different approach by computing self-aware uncertainty directly from LLM internal states, triggering retrieval when uncertainty exceeds a threshold. More adaptive approaches include CtRLA (Huanshuo et al., 2025), which probes LLM latent states to inform retrieval timing.

Alternative feedback signals have also been explored. Rowen (Ding et al., 2024) uses answer consistency across languages and across different LLMs as a retrieval trigger. If total consistency falls below a threshold, retrieval is triggered. However, consistency-based approaches have a fundamental flaw: models can be consistently wrong, leading to high consistency scores for incorrect answers (V et al., 2025).

A critical limitation of the above methods is that they treat all queries as equally complex and do not assess whether the retrieved context is actually relevant or sufficient. Adaptive RAG (Jeong et al., 2024) addresses query complexity by using a routing mechanism to predict whether retrieval is needed and determining the number of retrieval rounds based on complexity. However, it assumes retrieved context is relevant without verification, potentially propagating noisy or irrelevant information.

To address context quality, CRAG (Yan et al., 2024) introduces relevance assessment, using a fine-tuned model to classify retrieved documents as correct, incorrect, or ambiguous. If the context is incorrect, a rewritten query is issued to a web search engine. While CRAG focuses on relevance, it does not assess sufficiency—relevant documents may still lack complete information. SIM-RAG (Yang et al., 2025b) addresses this gap by training a lightweight critic model to evaluate context sufficiency. If information is insufficient, a new query is formulated using both the original query and already retrieved context. CoV-RAG (He et al., 2024) takes a comprehensive approach by identifying multiple error types (reference correctness, answer correctness, and truthfulness) and scoring them using a trained verifier, triggering retrieval or refinement based on these scores.

Key Insights

External feedback signals help reduce retrieval rounds but may still retrieve noisy context, and complexity assessment remains challenging.

5.3 Self-Triggered Retrieval via Reasoning

These approaches enable LLMs to autonomously decide when and how to retrieve through query decomposition or planning, termed *Reasoning RAG* or *Agentic RAG*. They fall into two categories: prompt-based methods with few-shot examples, and learned methods where models decide autonomously.

Prompt-Based Methods. DeComP (Khot et al., 2023) divides tasks into sub-tasks but only trig-

gers retrieval without reasoning steps. ReAct (Yao et al., 2022) interleaves reasoning traces with actions, while Self-Ask (Press et al., 2023) decomposes queries into sub-questions. However, these approaches lack error correction mechanisms, leading to cascading failures. Searchain (Xu et al., 2024) constructs global reasoning chains to mitigate this. SMR (Lee et al., 2025) addresses redundant and misguided reasoning through Refine, Rerank, and Stop actions. MCTS-RAG (Hu et al., 2025) uses Monte Carlo Tree Search but is computationally expensive. Search-O1 (Li et al., 2025a) lets Large Reasoning Models decide autonomously when to retrieve.

Key Insights

Prompt-based query decomposition and interleaving reasoning with retrieval improve coverage for complex questions by deciding when and what to retrieve, but are prone to cascading errors and redundant steps.

Learned or Agentic Methods. These models are trained to use search/retrieval as tools during answer generation, with rewards for correct tool calls and context usage. Self-RAG (Asai et al., 2024) trains reflection tokens for retrieval decisions and document relevance estimation. Search-R1 (Jin et al., 2025) and R1-Searcher (Song et al., 2025) use reinforcement learning to optimize retrieval timing and query formulation. DynamicRAG (Sun et al., 2025b) trains LLM for dynamically selecting and ranking documents that are sufficient to answer the query. ReZero (Dao and Le, 2025) rewards retry actions after failed searches. DeepResearcher (Zheng et al., 2025) and WebThinker (Li et al., 2025b) interact with commercial search engines during training, leading to noisy context and high API costs. ZeroSearch (Sun et al., 2025a) approximates search engine behavior using LLM parametric memory to reduce costs and noise. A detailed analysis of the Agentic method is provided in Appendix B due to space limitations.

Verifier-Based Feedback. Re^{2}Search++ (Xiong et al., 2025) uses fine-tuned critics to verify intermediate answers and improve query quality.

Key Insights

Learned methods boost autonomy and integration but introduce retrieval latency, noise, and dependence on external web search, making evaluation difficult.

6 Evaluation Datasets and Benchmarks

Evaluating RAG systems requires diverse datasets that test retrieval accuracy, ranking quality, and answer generation across varying complexity levels.

6.1 Information Retrieval Benchmarks

Information retrieval has a long history of evaluation campaigns, including those from TREC, CLEF, NTCIR, and FIRE. The queries used in these collections are often developed to strike a balance between having too many relevant documents in the target collection (which can be too easy to retrieve and too difficult to properly annotate (Voorhees et al., 2022)) and too few relevant documents. Sometimes challenging topics are also developed deliberately (Voorhees, 2005). The topic development process often involves manual reformulation of queries to ensure good coverage of relevance assessments. Mirroring the annotation process itself, it can be beneficial for automated retrieval systems to also perform various forms of query understanding (expansion or rewriting) to help ensure high recall. This connects with methods in Section 3 and therefore, these approaches show performance gains. Comprehensive benchmarks like BEIR (Thakur et al., 2021) enable zero-shot generalization assessment across multiple domains. TREC Deep Learning tracks (Craswell et al., 2020, 2021, 2022, 2023) provide additional evaluation sets for modern retrieval systems.

6.2 Question Answering Benchmarks

QA datasets are categorized by reasoning complexity. **Single-hop benchmarks** like Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), SQuAD (Rajpurkar et al., 2016), and PopQA (Mallen et al., 2023) test direct information retrieval and utilization, where the query-level feedback (reformulated queries covering all aspects of the original query/question) is beneficial. **Multi-hop benchmarks** including 2WikiMultiHopQA (Ho et al., 2020), HotpotQA (Yang et al., 2018), and MuSiQue (Trivedi et al., 2022) require compositional reasoning over multiple documents. In multi-hop QA, the retrieval with the

original query/question does not perform well due to lexical mismatch (since the intermediate answer also needs to retrieve the golden documents). In such cases, the generation-time feedback, such as reasoning-aware query decomposition and retrieval on demand, is helpful. During generation, when LLMs hallucinate, the generation time feedback (retrieval on demand) helps in boosting LLM generation confidence. This connects with Section 5.2.

6.3 Fact Verification Benchmarks

These datasets evaluate claim verification against retrieved evidence. FEVER (Thorne et al., 2018) provides simple claims requiring Wikipedia evidence retrieval and NLI classification. HoVeR (Jiang et al., 2020) extends this to multi-hop reasoning scenarios. QuanTemp (V et al., 2024) focuses on numerical claim verification requiring claim decomposition and iterative retrieval. This relates to query-level feedback and generation-time feedback. AveriTeC (Schlichtkrull et al., 2023) provides real-world claims with web-based evidence, testing RAG systems against noisy and contradictory sources.

6.4 Complex Reasoning Benchmarks

BRIGHT (SU et al., 2025) introduces reasoning-aware relevance criteria, defining document relevance by the presence of logical constructs (deductive steps, analogies, constraints) rather than topical alignment, challenging lexical retrievers. However, reasoning-augmented retrieval methods with query rewriting using CoT (Chain-of-Thought) have shown performance gains (Weller et al., 2025). This connects with methods in Section 3. Agentic reasoning tasks like Report Generation (Jiang et al., 2024; Huot et al., 2024), Deep Research (Wu et al., 2025), GPQA (Rein et al., 2023), and MATH500 (Cobbe et al., 2021) involve web search access but traditionally lack corpus availability for retrieval evaluation (Li et al., 2025a; Wei et al., 2025). BrowseComp-Plus (Chen et al., 2025b) addresses this by providing curated corpora for Deep Research tasks. We refer to Appendix C for a comparative overview of key benchmarks used in RAG and IR systems. We provide details on the evaluation metrics and framework used in RAG, as well as feedback-specific metrics in Appendix D.

7 Challenges and Future Directions

Despite recent advances, test-time corpus-level feedback in RAG systems faces several key limitations related to computational cost, feedback quality, decision-making, and evaluation.

Computational Cost of Adaptive Retrieval. Many feedback-driven approaches involve costly operations such as multiple retrieval rounds, reranking with large models, or traversing corpus graphs (Rathee et al., 2025b; Kulkarni et al., 2023; Hu et al., 2025). These methods often apply uniformly across queries, regardless of complexity. Efficient strategies, e.g., lightweight rankers, selective triggering, or confidence-aware stopping (Jiang et al., 2023; Yao et al., 2025), are crucial to make such systems viable at scale. Recent work shows that smaller models can achieve competitive performance with high-quality context (V et al., 2025), highlighting the importance of retrieval efficiency.

Noisy and Unstructured Corpus Feedback. Retrieved documents often contain redundant or irrelevant content, and most systems lack mechanisms to assess document utility beyond relevance ranking. Few methods exploit inter-document structure such as semantic similarity or topical diversity (MacAvaney et al., 2022; Rathee et al., 2025b). Structured representations (e.g., retrieval graphs, clusters) could improve feedback signals by enabling more targeted document selection and filtering.

Lack of Feedback-Aware Decision Policies. Many RAG systems perform fixed sequences of retrieval and reformulation without explicit decision criteria for feedback sufficiency or action (reranking, rewriting, reretrieval) to take (Su et al., 2024b; Li et al., 2025a). Learning retrieval control policies based on document-level or generation signals is a promising but underexplored direction.

Inadequate Evaluation of Feedback Behavior. Existing benchmarks emphasize answer correctness or static retrieval recall, but rarely measure feedback effectiveness across rounds. Datasets often lack annotations for document utility, retrieval iteration, or evidence sufficiency. Metrics that credit systems for improving retrieval through feedback, e.g., via answer change, document set refinement, or reduced over-retrieval—are needed to advance the field (Zheng et al., 2025).

We believe that tackling these challenges is essential to make corpus-level feedback a robust and efficient component of real-world RAG pipelines, closing the loop between retrieval and reasoning in complex language tasks.

Limitations

This survey focuses exclusively on **test-time feedback mechanisms that involve interaction with the corpus** in Retrieval-Augmented Generation (RAG) systems. We refer to this as *corpus-level feedback*, signals derived from retrieved documents, re-rankers, document-document relationships, or other corpus-grounded structures. Several related forms of feedback fall outside our scope. First, we do not cover feedback mechanisms that operate independently of the corpus, such as LLM self-refinement, planning, reasoning without retrieval or end users. For example, techniques that rewrite queries based solely on model introspection (e.g., self-refine (Madaan et al., 2023)) without consulting retrieved content are not considered corpus feedback and are excluded.

Second, our focus is restricted to retrieval-centric adaptation. We do not survey approaches that modify the generation module unless they directly inform or adapt retrieval via corpus-level signals. Third, we do not cover training-time feedback or methods that rely on offline supervised signals to fine-tune retrievers. Our interest is in test-time feedback mechanisms that dynamically update the query, document pool, or ranking without modifying model parameters.

Finally, we do not cover feedback in multimodal RAG. Also, we do not cover other feedback, like user behavior, multimodal information, and knowledge graph structures.

References

Amin Abolghasemi, Leif Azzopardi, Seyyed Hadi Hashemi, Maarten de Rijke, and Suzan Verberne. 2025. Evaluation of attribution bias in generator-aware retrieval-augmented large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21105–21124, Vienna, Austria. Association for Computational Linguistics.

Daman Arora, Anush Kini, Sayak Ray Chowdhury, Nagarajan Natarajan, Gaurav Sinha, and Amit Sharma. 2023. Gar-meets-rag paradigm for zero-shot information retrieval. *CoRR*, abs/2310.20158.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Sebastian Bruch, Siyu Gai, and Amir Ingber. 2023. An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems*, 42(1):1–35.

Xinran Chen, Ben He, Xuanang Chen, and Le Sun. 2025a. Not all terms matter: Recall-oriented adaptive learning for PLM-aided query expansion in open-domain question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22139–22151, Vienna, Austria. Association for Computational Linguistics.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. 2025b. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent. In *First Workshop on Multi-Turn Interactions in Large Language Models*.

Jeonghwan Choi, Minjeong Ban, Minseok Kim, and Hwanjun Song. 2025. Word2Passage: Word-level importance re-weighting for query expansion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8276–8296, Vienna, Austria. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track. *Preprint*, arXiv:2102.07662.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2022. Overview of the trec 2021 deep learning track. In *Text REtrieval Conference (TREC)*. NIST, TREC.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2023. Overview of the trec 2022 deep learning track. In *Text REtrieval Conference (TREC)*. NIST, TREC.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the trec 2019 deep learning track. *Preprint*, arXiv:2003.07820.

Florin Cuconasu, Simone Filice, Guy Horowitz, Yoelle Maarek, and Fabrizio Silvestri. 2025. Do RAG systems suffer from positional bias? *CoRR*, abs/2505.15561.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 719–729, New York, NY, USA. Association for Computing Machinery.

Alan Dao and Thinh Le. 2025. Rezero: Enhancing LLM search ability by trying one-more-time. *CoRR*, abs/2504.11001.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *CoRR*, abs/2402.10612.

ExplodingGradients. 2024. Ragas: Supercharge your llm application evaluations. `https://github.com/explodinggradients/ragas`.

Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2024. Synergistic interplay between search and large language models for information retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9571–9583, Bangkok, Thailand. Association for Computational Linguistics.

Revanth Gangi Reddy, Pradeep Dasigi, Md Arafat Sultan, Arman Cohan, Avirup Sil, Heng Ji, and Hannaneh Hajishirzi. 2025. A large-scale study of reranker relevance feedback at inference. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3010–3014, New York, NY, USA. Association for Computing Machinery.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Kailash A. Hambarde and Hugo Proença. 2023. Information retrieval: Recent advances and beyond. *IEEE Access*, 11:76581–76604.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10371–10393, Miami, Florida, USA. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yunhai Hu, Yilun Zhao, Chen Zhao, and Arman Cohan. 2025. MCTS-RAG: enhancing retrieval-augmented generation with monte carlo tree search. *CoRR*, abs/2503.20757.

Liu Huanshuo, Hao Zhang, Zhijiang Guo, Jing Wang, Kuicai Dong, Xiangyang Li, Yi Quan Lee, Cong Zhang, and Yong Liu. 2025. CtrlA: Adaptive retrieval-augmented generation via inherent control. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12592–12618, Vienna, Austria. Association for Computational Linguistics.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2024. Agents' room: Narrative generation through multi-step collaboration. *arXiv preprint arXiv:2410.02603*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Gautier Izacard and Edouard Grave. 2021. 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*, pages 874–880, Online. Association for Computational Linguistics.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *CoRR*, abs/2305.03653.

Nasreen Abdul Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. Umass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC*

2004, Gaithersburg, Maryland, USA, November 16-19, 2004, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

Pengyue Jia, Yiding Liu, Xiangyu Zhao, Xiaopeng Li, Changying Hao, Shuaiqiang Wang, and Dawei Yin. 2024. MILL: Mutual verification with large language models for zero-shot query expansion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2498–2518, Mexico City, Mexico. Association for Computational Linguistics.

Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2025a. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *CoRR*, abs/2503.00223.

Pengcheng Jiang, Siru Ouyang, Yizhu Jiao, Ming Zhong, Runchu Tian, and Jiawei Han. 2025b. Retrieval and structuring augmented generation with large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6032–6042, New York, NY, USA. Association for Computing Machinery.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina J Semnani, and Monica S Lam. 2024. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. *arXiv preprint arXiv:2408.15232*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*, abs/2503.09516.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Jia-Huei Ju, Suzan Verberne, Maarten de Rijke, and Andrew Yates. 2025. Controlled retrieval-augmented context evaluation for long-form RAG. *CoRR*, abs/2506.20051.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Hrishikesh Kulkarni, Nazli Goharian, Ophir Frieder, and Sean MacAvaney. 2024. Lexboost: Improving lexical document retrieval with nearest neighbors. In *Proceedings of the ACM Symposium on Document Engineering 2024*, DocEng '24, New York, NY, USA. Association for Computing Machinery.

Hrishikesh Kulkarni, Sean MacAvaney, Nazli Goharian, and Ophir Frieder. 2023. Lexically-accelerated dense retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 152–162. ACM.

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. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Dohyeon Lee, Yeonseok Jeong, and Seung-won Hwang. 2025. From token to action: State machine reasoning to mitigate overthinking in information retrieval. *CoRR*, abs/2505.23059.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick 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. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025a. Search-o1: Agentic search-enhanced large reasoning models. *Preprint*, arXiv:2501.05366.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025b. Webthinker: Empowering large reasoning models with deep research capability. *CoRR*, abs/2504.21776.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, Chunkit Chan, Yankai Chen, Zhongfen Deng, Yinghui Li, Hai-Tao Zheng, Dongyuan Li, Renhe Jiang, Ming Zhang, Yangqiu Song, and Philip S. Yu. 2025c. Towards agentic rag with deep reasoning: A survey of rag-reasoning systems in llms. *Preprint*, arXiv:2507.09477.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yuanhua Lv and ChengXiang Zhai. 2009. Adaptive relevance feedback in information retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 255–264, New York, NY, USA. Association for Computing Machinery.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2421–2425, New York, NY, USA. Association for Computing Machinery.

Sean MacAvaney, Nicola Tonellotto, and Craig Macdonald. 2022. Adaptive re-ranking with a corpus graph. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 1491–1500. ACM.

Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. 2023. Generative and pseudo-relevant feedback for sparse, dense and learned sparse retrieval. *CoRR*, abs/2305.07477.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. 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)*, pages 4089–4100, Online. Association for Computational Linguistics.

Donald Metzler and W. Bruce Croft. 2005. A markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 472–479, New York, NY, USA. Association for Computing Machinery.

Donald Metzler and W. Bruce Croft. 2007. Latent concept expansion using markov random fields. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, page 311–318, New York, NY, USA. Association for Computing Machinery.

Rodrigo Frassetto Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 708–718. Association for Computational Linguistics.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *CoRR*, abs/2309.15088.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. Rankzephyr: Effective and robust

zero-shot listwise reranking is a breeze! *CoRR*, abs/2312.02724.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Mandeep Rathee, Sean MacAvaney, and Avishek Anand. 2025a. Guiding retrieval using llm-based listwise rankers. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part I*, volume 15572 of *Lecture Notes in Computer Science*, pages 230–246. Springer.

Mandeep Rathee, Sean MacAvaney, and Avishek Anand. 2025b. Quam: Adaptive retrieval through query affinity modelling. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 954–962, New York, NY, USA. Association for Computing Machinery.

Mandeep Rathee, Venktesh V, Sean MacAvaney, and Avishek Anand. 2025c. Breaking the lens of the telescope: Online relevance estimation over large retrieval sets. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 2287–2297, New York, NY, USA. Association for Computing Machinery.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022.

S.E. Robertson, S. Walker, and M.M. Hancock-Beaulieu. 1995. Large test collection experiments on an operational, interactive system: Okapi at trec. *Information Processing & Management*, 31(3):345–360. The Second Text Retrieval Conference (TREC-2).

JJ Rocchio. 1971. Relevance feedback in information retrieval. *The SMART Retrieval System-Experiments in Automatic Document Processing/Prentice Hall*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Harrisen Scells, Shengyao Zhuang, and Guido Zuccon. 2022. Reduce, reuse, recycle: Green information retrieval research. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2825–2837, New York, NY, USA. Association for Computing Machinery.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems*, volume 36, pages 65128–65167. Curran Associates, Inc.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, and Luke Zettlemoyer. 2025. Reasonir: Training retrievers for reasoning tasks. *CoRR*, abs/2504.20595.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. 2024. Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15933–15946, Bangkok, Thailand. Association for Computational Linguistics.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *Preprint*, arXiv:2501.09136.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *CoRR*, abs/2503.05592.

Mingyang Song and Mao Zheng. 2024. A survey of query optimization in large language models. *CoRR*, abs/2412.17558.

Hongjin SU, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Liu Haisu, Quan Shi, Zachary Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan Arik, Danqi Chen, and Tao Yu.

2025. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. In *International Conference on Representation Learning*, volume 2025, pages 48941–48991.

Weihang Su, Yichen Tang, Qingyao Ai, Changyue Wang, Zhijing Wu, and Yiqun Liu. 2024a. Mitigating entity-level hallucination in large language models. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, page 23–31, New York, NY, USA. Association for Computing Machinery.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024b. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12991–13013, Bangkok, Thailand. Association for Computational Linguistics.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. 2025a. Zerosearch: Incentivize the search capability of llms without searching. *CoRR*, abs/2505.04588.

Jiashuo Sun, Xianrui Zhong, Sizhe Zhou, and Jiawei Han. 2025b. DynamicRAG: Leveraging outputs of large language model as feedback for dynamic reranking in retrieval-augmented generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Mujeen Sung, Jungsoo Park, Jaewoo Kang, Danqi Chen, and Jinhyuk Lee. 2023. Optimizing test-time query representations for dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5731–5746, Toronto, Canada. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Venktesh V, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 650–660, New York, NY, USA. Association for Computing Machinery.

Venktesh V, Mandeep Rathee, and Avishek Anand. 2025. SUNAR: Semantic uncertainty based neighborhood aware retrieval for complex QA. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5818–5835, Albuquerque, New Mexico. Association for Computational Linguistics.

Ellen M. Voorhees. 2005. The TREC robust retrieval track. *SIGIR Forum*, 39(1):11–20.

Ellen M. Voorhees, Nick Craswell, and Jimmy Lin. 2022. Too many relevants: Whither cranfield test collections? In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2970–2980. ACM.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 22–32, New York, NY, USA. Association for Computing Machinery.

Haoyu Wang, Ruirui Li, Haoming Jiang, Jinjin Tian, Zhengyang Wang, Chen Luo, Xianfeng Tang, Monica Xiao Cheng, Tuo Zhao, and Jing Gao. 2024. BlendFilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1009–1025, Miami, Florida, USA. Association for Computational Linguistics.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

Xiao Wang, Craig MacDonald, Nicola Tonellotto, and Iadh Ounis. 2023b. Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Trans. Web*, 17(1).

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023c. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315, Singapore. Association for Computational Linguistics.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *CoRR*, abs/2504.12516.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn J. Lawrie, and Benjamin Van Durme. 2025. Rank1: Test-time compute for reranking in information retrieval. *CoRR*, abs/2502.18418.

Junde Wu, Jiayuan Zhu, and Yuyuan Liu. 2025. Agentic reasoning: Reasoning llms with tools for the deep research. *CoRR*, abs/2502.04644.

Kaige Xie, Philippe Laban, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. 2025. Do RAG systems cover what matters? evaluating and optimizing responses with sub-question coverage. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5836–5849, Albuquerque, New Mexico. Association for Computational Linguistics.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2025. Rag-gym: Systematic optimization of language agents for retrieval-augmented generation. *Preprint*, arXiv:2502.13957.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jinxi Xu and W. Bruce Croft. 1996. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, page 4–11, New York, NY, USA. Association for Computing Machinery.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In *The Web Conference 2024*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *CoRR*, abs/2401.15884.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Diji Yang, Linda Zeng, Jinmeng Rao, and Yi Zhang. 2025b. Knowing you don't know: Learning when to continue search in multi-round rag through self-practicing. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 1305–1315, New York, NY, USA. Association for Computing Machinery.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Liu Weichuan, Lei Hou, and Juanzi Li. 2025. SeaKR: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27022–27043, Vienna, Austria. Association for Computational Linguistics.

Xiaopeng Ye, Chen Xu, Chaoliang Zhang, Zhaocheng Du, Jun Xu, Gang Wang, and Zhenhua Dong. 2025. Q-PRM: Adaptive query rewriting for retrieval-augmented generation via step-level process supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15113–15128, Suzhou, China. Association for Computational Linguistics.

HongChien Yu, Chenyan Xiong, and Jamie Callan. 2021. Improving query representations for dense retrieval with pseudo relevance feedback. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3592–3596, New York, NY, USA. Association for Computing Machinery.

Hamed Zamani and W. Bruce Croft. 2016. Embedding-based query language models. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, page 147–156, New York, NY, USA. Association for Computing Machinery.

Hamed Zamani and W. Bruce Croft. 2017. Relevance-based word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 505–514, New York, NY, USA. Association for Computing Machinery.

Chengxiang Zhai and John Lafferty. 2001. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, page 403–410, New York, NY, USA. Association for Computing Machinery.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *CoRR*, abs/2504.03160.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *CoRR*, abs/2503.06034.

Boyang Zuo, Xiao Zhang, Feng Li, Pengjie Wang, Jian Xu, and Bo Zheng. 2025. VALUE: value-aware large language model for query rewriting via weighted trie in sponsored search. *CoRR*, abs/2504.05321.

A Literature Compilation

A.1 Search Strategy

We conducted a comprehensive search on Google Scholar. We first focused on highly relevant Natural language Processing (NLP) venues such as ACL, EMNLP, NAACL, COLM and journals like TACL to collect RAG related literature. We also extensively curated papers from IR venues like SIGIR, ECIR, CIKM, WSDM to cover information retrieval literature and recent advancements in RAG systems.

A.2 Compilation Strategy

After careful review of Abstract, Introduction, Conclusion and Limitations we only retained papers that employ feedback mechanisms for improving retrieval and other components of RAG system which also helped synthesize our definition of feedback described in Section 3

B Detailed discussion on Agentic Methods.

The Agentic models go beyond prompt instructions and use search/retrieval as a tool. These models are trained to trigger this tool during answer generation. The training process mainly focuses on giving rewards for correct tool calls and context usage. In addition, similar to RAG methods, the retrieved documents are used as context to generate intermediate answers or the final answer. The search tool might have access to a local database or a web search engine to retrieve up-to-date knowledge.

Self-RAG (Asai et al., 2024) trains to predict reflection tokens for deciding when to retrieve and for estimating the relevance of retrieved documents. In addition, it judges the retrieved documents based on the generated answers and their factuality. However, it can fail when its self-reflection misjudges retrieval needs or relevance, leading to missed information or reliance on irrelevant context.

Search-R1 (Jin et al., 2025) is an extension of the DeepSeek-R1 (Guo et al., 2025) model, where the retrieval is a component training process. It autonomously generates search queries and performs real-time retrieval during step-by-step reasoning processes through reinforcement learning, including GRPO and PPO. The retrieval is triggered by `<search>` and `</search>` tokens, and the retrieved context is enclosed in `<information>` and `</information>` tokens. Similarly, R1-Searcher (Song et al., 2025) also uses an RL framework and uses two-stage rewards. The first stage has a retrieval reward that helps the model to use the correct format to trigger the retrieval, and the second stage has an answer reward that encourages the model to learn to utilize external retrieval effectively. While both these methods encourage better integration of external knowledge, they still inherit the limitations of retrieval latency and potential noise from the search source.

ReZero (Retry-Zero) (Dao and Le, 2025) introduces an RL framework that rewards the act of retrying search queries following an unsuccessful initial attempt, and it encourages LLM to explore alternative queries rather than prematurely stopping. The training process provides positive signals/rewards (feedback) if the model executes a retry action after failed searches, teaching the philosophy of "try one more time". However, these local database-based searches might miss the up-to-date knowledge and could generate answers for queries that require such knowledge.

DeepResearcher (Zheng et al., 2025) and WebThinker (Li et al., 2025b) interact in real time with commercial search engines during training, which leads to noisy context from the web (since the quality of these documents is unpredictable) and a high number of API calls. To address these limitations, ZeroSearch (Sun et al., 2025a) argues that since LLM has acquired enough world knowledge during heavy pre-training, it does not need to use a search engine during training. Since the LLM itself can generate a good-quality document from its parametric memory that answers the query, as well as noisy documents. Hence, it can approximate the real search engine behavior during training and reduce the training costs and noise, but its effectiveness depends on the LLM's pre-trained knowledge.

C Dataset Overview

Table 2 provides a comparative overview of key benchmarks used in the evaluation of RAG and IR systems, categorised by their characteristics and the type of feedback mechanisms that help.

D Evaluation Metrics

Evaluating retrieval systems and Retrieval-Augmented Generation (RAG) pipelines is critical for ensuring the accuracy, relevance, and reliability of generated responses. Retrieval evaluation typically focuses on metrics such as recall@k, precision, mean reciprocal rank (MRR), and hit

Dataset	Type	Reasoning	Beneficial Feedback Types
Single-hop QA			
Natural Questions	QA	Single	Query + Retrieval
TriviaQA	QA	Single	Query + Retrieval
SQuAD	QA	Single	Retrieval
PopQA	QA	Single	Query + Retrieval
Multi-hop QA			
HotpotQA	QA	Multi	Query + Retrieval + Generation
2WikiMultiHopQA	QA	Multi	Query + Retrieval + Generation
MuSiQue	QA	Multi	All three (with iterative retrieval and verification)
Fact Verification			
FEVER	Verification	Single	Retrieval + Generation
HoVeR	Verification	Multi	Query + Retrieval + Generation
QuanTemp	Verification	Multi	Query + Retrieval + Generation
AveriTeC	Verification	Multi	All three (with critical noise filtering)
Information Retrieval			
BEIR	IR	Mixed	Query + Retrieval (re-ranking)
TREC-DL	IR	Single	Query + Retrieval (re-ranking)
Complex Reasoning			
BRIGHT	IR	Multi	Query + Retrieval + Generation
GPQA	QA	Multi	All three (with agentic reasoning)
BrowseComp-Plus	QA	Multi	All three (with multi-round agentic)

Table 2: Summary of benchmark datasets for evaluating feedback-driven RAG systems. Beneficial Feedback Types indicate which feedback mechanisms (Query-level, Retrieval-level, Generation-time) are most effective for each dataset based on its complexity and characteristics.

rate, which assess how effectively the system retrieves pertinent documents or passages given a query. In contrast, RAG evaluation is more holistic, combining retrieval quality with generation fidelity and coherence. Common approaches include measuring answer correctness using exact match (EM) or F1 score, assessing faithfulness to retrieved evidence to detect hallucinations, and evaluating relevance and fluency through human or automated scoring (e.g., BLEU, ROUGE, or BERTScore). Recent frameworks like RAGAS (Exploding-Gradients, 2024), ARES (Saad-Falcon et al., 2024), and CRUX (Ju et al., 2025) also emphasize end-to-end evaluation, where the interplay between retrieval accuracy and generation quality is analyzed to identify bottlenecks, such as irrelevant documents leading to incorrect answers, making comprehensive evaluation essential for diagnosing and improving RAG system performance.

However, the current RAG evaluation methods mainly focus on the retrieval and final answer performance. However, Reasoning RAG systems are highly dependent on intermediate reasoning steps and retrieval rounds. Therefore, it is also important to consider additional evaluation dimensions such as computational cost, efficiency, or number of retrieval rounds.

Feedback-Specific Evaluation Metrics. Emerging metrics specifically designed for feedback-driven RAG systems include:

- **Retrieval Efficiency:** Number of retrieval rounds required to obtain sufficient evidence, measured as average retrieval steps per query. Lower values indicate more efficient feedback mechanisms (Asai et al., 2024; Jiang et al., 2023). For instance, BrowseCompPlus (Chen et al., 2025b) shows tradeoff between search calls and effectiveness.

- **Query Reformulation Quality:** While query reformulation quality is indirectly measured by improvements in downstream retrieval performance and answer generation (Ma et al., 2023), several works also independently measure quality of generated reformulations. Common metrics involve BertScore (Zuo et al., 2025) to ascertain the semantic similarity between original and reformulated queries and measures like ROUGE which measure the lexical overlap (Ye et al., 2025). Effective feedback should increase relevance while preserving query intent (Wang et al., 2023a). Additionally, in RAG setup it is also necessary to ensure that reformulated queries lead to better answers apart from retrieval performance and hence (Ma et al., 2023) leverage high quality

feedback from answer generating LLM to update the query reformulator LLM using reinforcement learning. Alternatively, more recently, Process Reward models (Ye et al., 2025) have been employed to update the reformulator in a more adaptive and interpretable manner resulting in high-quality queries as measured by above metrics.

- **Evidence Coverage:** This metric tracks not only recall improves over rounds in complex tasks involving RAG but also if the content covers necessary information to answer complex queries. For instance, in CRUX (Ju et al., 2025), apart from evaluating relevance of retrieved documents, the content of each document is also examined to quantify coverage which is measured as the number of sub-questions that can be answered by the content in the document. (Xie et al., 2025) also propose a similar metric defining coverage of retrieved documents in terms of answerable sub-questions apart from relevance. Better feedback from generator helps retrieve documents that improve such coverage.

- **Feedback Signal Accuracy:** Precision of retrieval triggers (whether retrieval was needed when triggered) and relevance predictions from self-assessment modules. This measures the quality of meta-reasoning (Asai et al., 2024).

- **Cost-Effectiveness:** Trade-off between computational cost (LLM calls, retrieval operations) and final answer quality. Measured as performance per unit cost or Pareto efficiency (Chen et al., 2025b; Jeong et al., 2024). Some works such as (V et al., 2025) showed that smaller LLMs can produce good quality answers if provided relevant, high-quality context and recommends to allocate more compute on retrieval which results in low cost at generation level. Also, these GPU heavy models, used at ranking or generation stage, are costly in terms of carbon footprints. Recent work, GreenIR (Scells et al., 2022), studies costs of such ranking models.

- **Attribution Quality:** Whether generated answers correctly cite specific retrieved passages that support each claim, evaluated through citation precision and recall (Asai et al., 2024; Abolghasemi et al., 2025). More recently (Wallat et al., 2025), posit that citation correctness which measures whether attributed documents support the corresponding statements is not enough. They propose a new metric **citation faithfulness** which additionally measures whether the model is genuinely grounded on attributed documents than post-rationalized to fit pre-existing parametric knowledge.

Limitation of Feedback-Specific Evaluation Metrics Despite progress in RAG evaluation, several critical limitations remain, particularly for feedback-driven systems.

Ground Truth Limitations for Multi-Hop Tasks. Annotating "correct" retrieval paths for multi-hop reasoning is subjective and expensive. Multiple valid reasoning chains may exist, yet most benchmarks provide only one gold path. This makes it challenging to evaluate whether alternative retrieval strategies are genuinely inferior or simply different (Trivedi et al., 2022; Ho et al., 2020).

Feedback Quality Attribution. When RAG systems use multiple feedback sources (rankers, verifiers), it is unclear which component contributed to improvements or failures. Current metrics provide aggregate scores but lack fine-grained attribution to diagnose whether retrieval, ranking, or generation feedback was most impactful.