

How Many Documents Are Too Many? Modeling Retrieval Saturation in Retrieval-Augmented Generation*

Abstract--Retrieval-Augmented Generation (RAG) is often deployed under the assumption that retrieving more documents improves generation by increasing evidence coverage. However, larger retrieved sets may introduce redundancy, distractors, and higher inference cost, making retrieval quantity different from retrieval utility. This study examines retrieval saturation—the point at which increasing retrieval depth yields diminishing, negligible, or negative returns when answer quality and operational cost are jointly considered. In a controlled design, only retrieval depth is varied, while the corpus, retriever family, generator, prompts, and decoding policy remain fixed. Experiments are conducted on four benchmarks representing knowledge-intensive generation: Natural Questions, HotpotQA, ASQA, and QAMPARI. Three consistent patterns emerge. First, compact and multi-hop question answering tasks exhibit early saturation: performance improves only up to a limited retrieval depth, after which gains plateau or decline. Second, long-form and many-answer tasks show delayed saturation, as broader retrieval supports answer completeness, though later improvements are increasingly offset by higher cost. Third, dense retrieval with Contrastive outperforms sparse retrieval with BM25 near the apparent optimum, yet both show clear saturation as depth increases. Qualitative analysis indicates that unsupported elaboration, redundancy, and context inefficiency become more common beyond the useful retrieval range, particularly in compact QA tasks. Overall, retrieving more documents does not necessarily mean retrieving more useful evidence. Instead of treating retrieval depth as a monotonic variable, practical RAG design should frame it as a retrieval-budgeting problem that jointly optimizes answer quality, factual reliability, and inference cost.

Index Terms-- cost-utility trade-off, long-form question answering, open-domain question answering, retrieval-augmented generation, retrieval saturation.

INTRODUCTION

Large language models (LLMs) have achieved strong performance across a wide range of generation tasks, yet their parametric knowledge remains difficult to update, verify, and attribute. Retrieval-Augmented Generation (RAG) addresses this limitation by coupling parametric generation with external, non-parametric memory, allowing the model to retrieve supporting evidence at inference time rather than relying exclusively on information stored in model parameters [1], [2]. In knowledge-intensive settings, this architectural shift is especially important because it enables generated outputs to be grounded in retrievable evidence, thereby improving factual access, interpretability, and adaptability relative to closed-book generation [3], [4], [23].

As RAG systems have matured, a persistent engineering

intuition has emerged: if retrieval is helpful, then retrieving more documents should improve answer quality by increasing evidence coverage. That intuition is plausible but incomplete. Larger retrieved sets may expose more relevant passages, yet they may also introduce redundancy, distractor content, and positional competition within the model context window. In addition, broader retrieval increases system-level cost through longer prompts, higher latency, and greater computational burden at inference time. Retrieval quantity and retrieval utility should therefore not be treated as equivalent variables [5], [6], [26], [28].

This distinction becomes particularly important when analysis shifts from the retriever to the generator. A retriever may continue to improve coverage as top- k increases, while the generator's ability to use that evidence may plateau or deteriorate because of context dilution, interference among passages, or limited attention allocation across prompt positions. Recent work on long-context language models has shown that relevant information is not always used uniformly when placed in different parts of the prompt, and that simply extending context does not guarantee proportional gains in downstream reasoning or factual generation [7], [8]. Related studies on retrieval selection, reranking, and context compression likewise suggest that indiscriminately appending more retrieved content is often inefficient and can even be counterproductive [9], [3], [10].

The central question, then, is not whether retrieval helps, but when additional retrieval ceases to help enough to justify its cost. This article addresses that question through the concept of retrieval saturation. We define retrieval saturation as the transition from a regime in which increasing retrieval depth yields meaningful gains in downstream utility to a regime in which additional documents produce negligible or negative returns while cost continues to rise. Framed this way, retrieval depth is not merely a configurable hyperparameter; it is an operational budget variable whose optimal value depends on the interaction among task structure, retriever behavior, context utilization, and evaluation objective.

To study this phenomenon, we adopt a controlled experimental design in which the retrieval corpus, retriever family, generator, prompt family, and decoding policy are held constant while only retrieval depth varies. This design allows retrieval-depth effects to be interpreted more cleanly than in settings where multiple system components change

* Manuscript received , Revised, , accepted .

simultaneously. We evaluate four benchmark regimes—Natural Questions, HotpotQA, ASQA, and QAMPARI—which collectively span compact factoid question answering, multi-hop reasoning, long-form synthesis, and many-answer generation. This task diversity is important because saturation should not be expected to occur uniformly across settings with different evidence demands and answer structures [11], [12], [30], [32].

This article makes a narrower contribution than the broader claim that “more retrieval is not always better.” Recent work has already established selective retrieval, corrective retrieval, context compression, long-context limitations, and adaptive retrieval strategies [10], [16], [24], [33]–[35], [39]–[41]. The present study therefore focuses on a controlled measurement question: when retrieval depth k is the only manipulated variable, how does generation-side utility change, where does a reproducible saturation transition appear, how sensitive is that transition to the tolerance used to define a small marginal gain, and how should the resulting curve be interpreted under a quality–cost budget?

The core argument is therefore not that a single retrieval depth is universally optimal. Rather, the article argues that effective RAG design should treat retrieval depth as a budgeted operating variable to be tuned against task demands, evidence utilization, and system cost, rather than as a monotonic more-is-better choice [1], [6], [13], [25].

RELATED WORK

A. Retrieval-Augmented Generation and Multi-Document Question Answering

A substantial body of research shows that retrieval can materially improve generation in knowledge-intensive tasks by supplying external evidence at inference time. Retrieval-Augmented Generation (RAG) formalized the now-standard architecture in which a generator conditions on retrieved passages, enabling open-book generation with explicit evidence access [1]. Progress in retrieval quality—particularly through dense retrieval—further strengthened this pipeline. Dense Passage Retrieval (DPR) demonstrated large gains in open-domain question answering by learning question–passage representations optimized for retrieval, and it has since become a widely used backbone for knowledge-intensive systems [14].

On the generation side, multi-document conditioning has been shown to improve performance when models can aggregate evidence across multiple passages. Fusion-in-Decoder (FiD) established that generative QA benefits from attending over multiple retrieved passages during decoding, improving the model’s ability to integrate evidence beyond single-document conditioning [2]. Subsequent work such as Atlas extended retrieval-augmented language modeling toward broader few-shot and knowledge-intensive settings, demonstrating that retrieval can be beneficial beyond classic QA benchmarks [15].

Collectively, this literature establishes that retrieval matters. However, it does not fully resolve a more operational question central to deployed RAG: how the marginal value of additional retrieved documents behaves once a generator is already

conditioned on retrieved evidence, especially when gains in answer quality must be balanced against rising inference cost.

B. Selective Retrieval and Context Compression

Work on selective retrieval and context compression is particularly relevant because it challenges the assumption that more retrieved context is uniformly beneficial. Self-RAG argues that retrieval should be invoked selectively—guided by the model’s own self-reflection and critique—rather than applied indiscriminately across all inputs, highlighting that retrieval can introduce noise as well as signal [16], [24], [27]. Related research on compressing or restructuring retrieved evidence suggests that long retrieved inputs may burden generation and reduce effective evidence use, motivating methods that summarize, filter, or compress retrieved passages under fixed context budgets [10].

These studies reinforce the core intuition of the present article: retrieved context is not uniformly beneficial. In contrast to methods that change the evidence selection mechanism, our study isolates a more basic variable—retrieval depth—and measures how generation utility changes as the number of retrieved documents increases under otherwise fixed conditions.

C. Long-Context Utilization

The long-context literature is central to the retrieval-saturation problem because increasing retrieval depth changes not only evidence quantity but also evidence placement and competition within the prompt. Empirical analyses of long-context language models indicate that relevant evidence is not used uniformly across input positions, and that models may underutilize information placed away from the prompt’s most salient regions—a phenomenon often discussed under position sensitivity and “lost-in-the-middle” behavior [7]. These findings imply that retrieval saturation can arise not only from retriever-side limitations, but also from generator-side constraints in absorbing, prioritizing, and integrating expanding evidence contexts.

D. Benchmark Regimes and Evidence Structure

The benchmark suite used in this study is intentionally diverse in both answer structure and evidence demands. Natural Questions represents compact open-domain question answering grounded in real user queries [17]. HotpotQA emphasizes multi-document compositional reasoning and evidence aggregation across passages [12]. ASQA targets long-form answering under ambiguity, where useful answers must reconcile multiple plausible interpretations and provide grounded synthesis [18]. QAMPARI stresses many-answer coverage, where evidence is distributed across numerous passages and performance depends on breadth of supported aggregation [30].

This diversity is methodologically important because it enables retrieval saturation to be analyzed across compact, compositional, long-form, and many-answer settings under a shared retrieval substrate, clarifying which saturation behaviors are task-dependent versus system-dependent.

E. Long-Form Evaluation and Factuality

In long-form generation, headline task scores can obscure whether broader retrieval genuinely improves supported content. FActScore was introduced to evaluate long-form factual precision by decomposing responses into atomic claims and checking whether each claim is supported by evidence, thereby separating apparent fluency or length from verifiable groundedness [19]. More broadly, recent work in long-form QA evaluation emphasizes that answer quality in evidence-intensive tasks cannot be reduced to a single scalar notion of “better,” because different metrics capture different aspects of coverage, faithfulness, and support [11].

For the present article, this matters because deeper retrieval can increase answer length and apparent completeness without proportionally improving evidence-grounded correctness. Evaluating retrieval depth therefore requires metrics and analyses that remain sensitive to factual support, not just surface-level improvements.

RESEARCH GAP, RESEARCH QUESTIONS, AND CONTRIBUTIONS

Recent RAG research substantially complicates the simple assumption that a fixed, larger top-k is inherently better. Self-RAG trains a model to decide when retrieval is needed and to critique both retrieved evidence and generated content [24]. CRAG adds a retrieval evaluator and corrective actions when retrieved evidence is judged insufficient [33]. Adaptive-RAG routes questions according to estimated complexity and selects among different retrieval strategies [34]. Atlas demonstrates retrieval-augmented language modeling with a substantially stronger pretrained language-model backbone than the present FLAN-T5 setup [25]. LongRAG and related long-context studies further show that retrieval depth, chunking, and context construction interact with the capabilities of contemporary long-context models [35], [40], [41]. More recent evaluation work also argues that retrieval quality should be judged by its downstream contribution to generation, rather than by retrieval metrics alone [39].

Accordingly, the novelty claim of this paper is deliberately narrowed. The study does not claim to discover that excessive or poorly selected context can hurt generation. Its contribution is to provide a transparent, controlled retrieval-depth curve analysis in which k is isolated, a quantitative saturation-entry rule is specified, threshold sensitivity is reported, and cost is recorded alongside benchmark-native generation quality. The framework is complementary to adaptive and corrective RAG systems rather than a replacement for them.

No head-to-head score comparison with these adaptive or long-context systems is reported here. The supplied experimental archive contains the BM25/Contriever runs only; introducing external implementations without the same generator, corpus, evaluation splits, and cost accounting would confound the controlled retrieval-depth analysis. Accordingly, these methods are used to position the contribution and define its scope, not as a claim of comparative superiority.

In many Retrieval-Augmented Generation (RAG) systems, retrieval depth is treated primarily as an implementation parameter rather than as a variable whose downstream utility should be modeled explicitly. A common practical assumption is that increasing top- k improves generation by expanding evidence coverage. However, this assumption is only partially justified. Larger retrieved sets may include additional relevant evidence, but they may also introduce redundancy, distractor passages, positional competition within the context window, and higher inference cost. These factors can limit the generator’s ability to convert additional retrieved documents into better answers, especially when the retrieved context exceeds the model’s effective evidence-use capacity [2], [1], [7].

This creates a narrower methodological gap rather than an absence of prior evidence. Prior work has established that retrieval improves knowledge-intensive generation and that selective, corrective, adaptive, compressed, and long-context-aware retrieval can outperform indiscriminate context expansion [2], [10], [16], [24], [33]–[35], [39]–[41]. What remains useful to isolate is the downstream shape of utility as retrieval depth alone is varied under a frozen pipeline. In other words, the question addressed here is not whether more retrieved context can hurt, but how the marginal generation-side utility curve changes across task regimes and retrievers when other components are held fixed, and how that curve can be translated into an explicit retrieval-budgeting rule.

This article addresses that gap by examining retrieval saturation as a cost-sensitive phenomenon in RAG. Here, generation-side utility refers to the extent to which additional retrieved documents improve downstream answer quality, factual support, completeness, and cost-adjusted performance. The study is organized around five research questions:

1. RQ1: How does generation-side utility change as retrieval depth increases?
2. RQ2: Does the saturation profile differ across task regimes with distinct evidence structures?
3. RQ3: Does the saturation profile differ across retriever families when the generator and other pipeline components are held fixed?
4. RQ4: Do benchmark-native metrics and factuality-sensitive diagnostics provide consistent interpretations of deeper retrieval?
5. RQ5: Can retrieval-depth behavior support principled retrieval budgeting in practical RAG systems?

The article makes four bounded contributions. First, it formalizes retrieval saturation as a descriptive transition in the marginal downstream utility of retrieval depth, rather than claiming that the underlying “more context can hurt” phenomenon is itself new. Second, it isolates k as the manipulated variable across four evidence regimes and two retriever families. Third, it quantifies sensitivity of the saturation transition to alternative marginal-gain tolerances using the reported score curves. Fourth, it frames retrieval budgeting as a deployment problem in which quality, token

burden, and latency must be jointly considered, while explicitly identifying the statistical and experimental limits of the archived study.

By framing retrieval depth as a budgeted operating variable, this study contributes to a more systematic understanding of how RAG systems should balance evidence coverage, answer quality, factual reliability, and computational cost.

EXPERIMENTAL DESIGN

This study adopts a controlled retrieval-depth design to examine how the marginal utility of retrieved evidence changes as the number of retrieved documents increases. Within each experimental condition, the retrieval corpus, retriever family, generator, prompt family, and decoding policy are held constant, while only retrieval depth, denoted as k , is varied. This design is necessary because the central analytical target of the study is the generation-side effect of additional retrieved documents. If prompt structure, corpus composition, retrieval model, or generation policy were changed simultaneously with k , observed performance differences could not be attributed specifically to retrieval depth.

The design follows the broader logic of controlled evaluation in retrieval-augmented generation: retrieval depth is treated not merely as a hyperparameter, but as an independent variable whose downstream utility must be evaluated in relation to answer quality, factual support, and inference cost. This framing is consistent with prior work showing that retrieval improves knowledge-intensive generation, while also acknowledging that larger retrieved contexts may create noise, redundancy, and inefficient use of the context window [2], [1], [7].

In this article, retrieval saturation is defined as a measurable transition in the marginal downstream utility of retrieval depth. Let the tested retrieval depths be $k_0 < k_1 < \dots < k_m$, let $Q(k_i)$ denote the primary task score at depth k_i , and let $C(k_i)$ denote a non-decreasing operational cost measure (prompt tokens, latency, or a joint deployment cost). Define the marginal score change as $\Delta Q_i = Q(k_i) - Q(k_{i-1})$.

$$k_{\text{sat}}(\delta) = \min \{ k_i : i > 0, \Delta Q_i < \delta, \text{ and } \Delta C_i > 0 \}.$$

A subsequent negative marginal change ($\Delta Q_j < 0$ for $j \geq i$) is classified as overload evidence. The rule is descriptive; it identifies a point at which the observed marginal gain is smaller than the chosen practical tolerance while operational cost continues to grow. It does not assert statistical significance by itself. Confidence intervals are reported to characterize uncertainty, but paired inferential tests require per-instance outputs and are treated separately below. This formalization also makes explicit that a saturation point is relative to the task score, tested k -grid, tolerance δ , and cost definition rather than a universal property of RAG.

For a deployment budget B , a principled budgeted operating point can be written as:

$$k^*(B) = \arg \max_k Q(k) \text{ subject to } C(k) \leq B$$

or, when a continuous cost penalty is justified, $U_\lambda(k) = Q(k) - \lambda C(k)$.

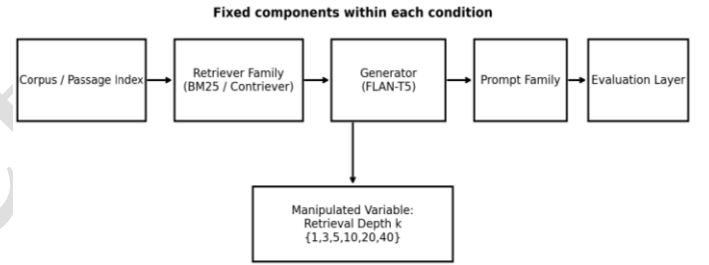
Because the manuscript does not define a universal deployment budget or λ , the empirical tables report saturation-

entry depths and score-retention/cost comparisons rather than claiming a globally optimal k .

The study distinguishes three interpretive regimes:

1. Growth regime: Increasing retrieval depth yields meaningful improvements in downstream utility. In this regime, additional documents contribute useful evidence that the generator can exploit.
2. Saturation regime: Increasing retrieval depth yields small, unstable, or non-meaningful gains while inference cost continues to rise. In this regime, additional evidence may still contain relevant information, but its marginal contribution to answer quality is limited.
3. Overload regime: Increasing retrieval depth yields no improvement or measurable degradation while cost continues to rise. In this regime, additional retrieved documents may introduce redundancy, distractor passages, positional competition, or context dilution that reduces generation effectiveness.

This taxonomy is used as an interpretive framework rather than as a universal law. Its role is to support principled reading of retrieval-depth curves across task regimes and retriever families. In particular, the framework allows the analysis to distinguish between the retrieval depth that maximizes a benchmark score and the retrieval depth that provides the most favorable quality–cost trade-off.



Only retrieval depth varies; corpus, retriever, generator, prompts, and decoding remain fixed.

Fig 1. Controlled retrieval-depth study design.

Figure 1 illustrates the controlled retrieval-depth study design. The figure separates the fixed experimental components—retrieval corpus, retriever family, generator, prompt family, decoding policy, and evaluation layer—from the manipulated variable, retrieval depth k . This structure formalizes the study’s causal logic: the analysis concerns marginal generation-side utility under a constant RAG pipeline rather than absolute performance under multiple changing conditions.

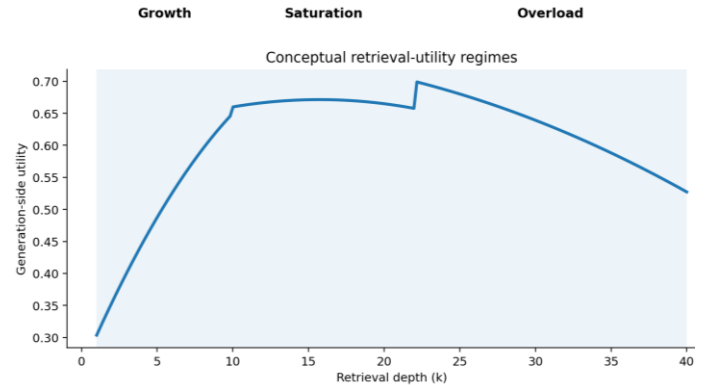


Fig 2. Conceptual retrieval-utility regimes.

Figure 2 presents the conceptual retrieval-utility regimes used throughout the article. The curve is schematic rather than empirical. Its purpose is to define the interpretation space for later analysis by distinguishing growth, saturation, and overload as qualitatively different patterns of retrieval-depth behavior.

EXPERIMENTAL SETUP

A. Task Suite and Dataset–Evidence Alignment

The experimental suite is designed to evaluate retrieval saturation across four distinct evidence regimes. This diversity is necessary because the marginal value of additional retrieved documents is unlikely to be uniform across tasks with different answer structures and evidence requirements. Compact factoid question answering may saturate early because a small number of relevant passages can contain the answer, whereas long-form or many-answer tasks may continue to benefit from broader retrieval because relevant evidence is distributed across a larger set of passages.

Table 1 summarizes the benchmark selection logic. Natural Questions is used to represent single-answer open-domain question answering grounded in real user queries [17]. HotpotQA represents multi-hop and multi-document reasoning, where correct answers often require combining evidence across passages [12]. ASQA captures ambiguous long-form question answering, in which a high-quality response must address multiple plausible interpretations of the query [18]. QAMPARI represents many-answer open-domain QA, where the target output is a set of entities and useful evidence may be distributed across many retrieved passages [30], [32].

Table 1 shows that the benchmark suite spans compact QA, compositional reasoning, long-form ambiguity resolution, and many-answer coverage under a shared retrieval substrate. This breadth allows the study to test whether retrieval saturation is primarily a general property of RAG pipelines or a task-dependent phenomenon shaped by evidence structure.

TABLE 1
Dataset Profile and Task–Evidence Alignment

Dataset	Task regime	Answer structure	Why included	Retrieval corpus	Primary evaluation family
Natural Questions	Single-answer open-domain QA	Short factual answer	Tests early saturation in compact QA	Fixed DPR-style Wikipedia passage index	Short-answer F1 / EM
Hotpot QA	Multi-hop /	Short answer	Tests composition	Fixed DPR-	Answer EM / F1

	multi-document QA	with supporting evidence	nal evidence needs	style Wikipedia passage index	
ASQA	Ambiguous long-form QA	Long-form disambiguating answer	Tests saturation in long-form synthesis	Fixed DPR-style Wikipedia passage index	STR-EM / Disambig-F1 / DR Score
QAMPARI	Many-answer open-domain QA	Set/list of entity answers	Tests distributed evidence coverage	Fixed DPR-style Wikipedia passage index	Precision / Recall / F1

B. Retrieval Corpus and Passage Segmentation

All experiments are conducted over a fixed DPR-style Wikipedia passage index segmented into disjoint 100-word passages with article titles prepended. This passage format follows common practice in open-domain QA and retrieval-augmented generation, where Wikipedia is segmented into retrievable units to support scalable evidence access [14], [1].

Fixing passage segmentation is methodologically important because passage length directly affects retrieval granularity, redundancy, prompt length, and evidence fragmentation. Shorter passages may increase retrieval precision but fragment relevant evidence across more units, whereas longer passages may provide richer context at the cost of greater prompt burden. If segmentation were allowed to vary across conditions, top- k would no longer represent a stable comparison unit. Therefore, the study holds passage segmentation fixed so that observed differences can be attributed more directly to retrieval depth rather than to changes in evidence granularity.

C. Retrievers, Generator, and Frozen Protocol

The study compares one sparse retriever and one dense retriever. The sparse retriever is BM25 implemented through Pyserini, which provides a reproducible lexical retrieval baseline widely used in information retrieval and open-domain QA research [20], [21]. The dense retriever is Contriever, an unsupervised dense retrieval model that has demonstrated strong transfer across retrieval tasks without relying on task-specific supervision [15].

All reported test-set results are produced with a single frozen FLAN-T5-family checkpoint and its corresponding tokenizer, held constant across all reported conditions. FLAN-T5 is an instruction-tuned sequence-to-sequence model family designed to improve task generalization through instruction tuning [22]. Smaller FLAN-T5 variants are used only during pilot validation and pipeline debugging and do not contribute to the final reported tables or figures. This separation between pilot validation and final reporting is important for preventing pilot-stage adjustments from contaminating the interpretation of held-out evaluation results.

To preserve the controlled retrieval-depth comparison, generation settings remain fixed throughout the reported evaluation. These settings include source-length limits, target-

length limits, decoding mode, stopping behavior, prompt template, and tokenizer configuration. The retrieval-depth grid is fixed in advance as $k \in \{1, 3, 5, 10, 20, 40\}$, preventing post hoc selection of retrieval depths based on observed performance.

TABLE 2
Frozen Implementation Setup

Component	Fixed choice	Why fixed
Sparse retriever	BM25 via Pyserini	Provides a reproducible sparse lexical baseline
Dense retriever	Contriever	Provides an open dense retriever with strong transfer behavior
Pilot generator tier	Smaller FLAN-T5 variants	Used only for debugging and pilot validation
Final generator tier	One frozen FLAN-T5-family checkpoint with matching tokenizer	Required for reproducibility and latency interpretation
Retrieval-depth grid	$\{1, 3, 5, 10, 20, 40\}$	Prevents post hoc selection of k
Prompt family	One frozen prompt family per task regime	Prevents prompt confounding
Decoding policy	Fixed within each dataset-model condition	Prevents generation-side confounding
Split policy	Development split for protocol freezing; held-out test split for reporting	Prevents pilot leakage
Seed / run policy	Fixed repeated-evaluation protocol across all reported conditions	Identifies the source of run-level variation
Uncertainty basis	Confidence intervals computed under a fixed repeated-evaluation procedure	Ensures interpretable uncertainty summaries
Significance testing	Descriptive saturation analysis in the main manuscript	Avoids over-interpreting small score differences
Generator settings	Fixed source/target length limits and decoding controls	Affects both answer quality and cost

Table 2 serves an internal-validity function by making explicit that retrieval depth is the only manipulated variable during the reported evaluation. The BM25–Contriever comparison is scientifically meaningful because it tests whether saturation depends not only on retrieved-set size but also on retrieved-set composition. In other words, two retrievers may retrieve the same number of passages while producing different mixtures of relevant, redundant, and distracting evidence.

D. Input Construction and Prompt Policy

For each query, the retriever returns a ranked list of passages from the fixed index. The generator then receives the top- k passages concatenated in retrieval-rank order under a single frozen prompt family for the corresponding task regime. In short-answer settings, the prompt requests a concise answer grounded in the retrieved passages. In long-form settings, the prompt requests a synthesized answer that remains faithful to the retrieved evidence.

The prompt template does not change across values of k . The same principle applies to decoding strategy, maximum generation length, and stopping behavior. This constraint is important because prompt wording and decoding choices can independently affect answer length, factuality, and

extractiveness. Holding them fixed ensures that changes in output quality are interpreted as consequences of retrieval depth rather than prompt redesign or generation-policy variation.

This design also reflects a central concern in RAG evaluation: retrieved evidence may improve performance only if the generator can effectively select, integrate, and use the relevant information. Prior work on multi-passage generation and long-context utilization has shown that simply increasing the amount of available context does not guarantee proportional improvements in answer quality [2], [7]. The fixed prompt policy therefore allows the study to observe when broader retrieval remains useful and when it begins to create diminishing returns.

E. Run Policy, Seeds, and Uncertainty Basis

Uncertainty is interpreted relative to the predefined evaluation unit for each reported condition. For every dataset, retriever, and retrieval-depth setting, the system configuration remains fixed according to the frozen protocol. Reported intervals summarize variation under the repeated-evaluation procedure used in the study, while preserving retrieval depth k as the only experimental variable of interest.

Run-level variation is treated as the relevant source of uncertainty for the reported intervals. These intervals are used as uncertainty summaries rather than as standalone significance tests. This distinction is important because the purpose of the article is to characterize retrieval-depth behavior—growth, saturation, and overload—rather than to claim statistical superiority for every adjacent value of k .

The saturation rule is now defined mathematically rather than by combining threshold and confidence-interval overlap into one criterion. For each adjacent pair of retrieval depths, the marginal gain ΔQ_i is compared with the pre-specified tolerance δ and the cost change is checked to be positive. The principal descriptive rule is therefore $\Delta Q_i < \delta$ with $\Delta C_i > 0$; a negative ΔQ_i after saturation is additionally flagged as overload evidence. Confidence intervals remain visible in the main result tables as uncertainty summaries, but interval overlap is not used as a substitute for a paired significance test.

This rule is not intended as a substitute for paired hypothesis testing. Instead, it functions as a transparent descriptive procedure for identifying likely saturation behavior. Where paired inferential tests are not reported, the manuscript uses cautious comparative language such as “higher,” “lower,” “plateau,” “cost-dominated,” and “deployment-efficient,” rather than stronger claims of statistical superiority. This conservative reporting strategy reduces the risk of over-interpreting small numerical differences that may not be meaningful in practical deployment.

EVALUATION PROTOCOL AND STATISTICAL ANALYSIS

Evaluation is task-aware and benchmark-specific. Natural Questions is evaluated using short-answer exact match and F1 [17]. HotpotQA is evaluated using answer exact match, answer F1, and supporting-fact F1 to capture both answer correctness and evidence identification in multi-hop reasoning [12]. ASQA

is evaluated using STR-EM, Disambig-F1, and DR Score, reflecting its long-form ambiguous question-answering formulation [18]. QAMPARI is evaluated using set-level precision, recall, and F1 because its target output is a set of entity answers rather than a single answer span [30]. These metric families are not interchangeable because they encode different assumptions about answer structure, evidence coverage, and failure modes.

FactScore is added as a factuality-sensitive diagnostic in long-form and many-answer settings [19]. Its role is to prevent the analysis from equating broader or longer output with better grounded output. If benchmark-native metrics continue to rise while factual support flattens or weakens, the appropriate interpretation is not that retrieval uniformly improved generation, but that it improved one aspect of utility without producing proportional gains in evidence-backed factual content.

Cost-side variables include prompt token count and end-to-end latency. Tables 3 and 4 report mean scores with confidence intervals. These intervals summarize uncertainty under the declared repeated-evaluation procedure but are not interpreted as standalone significance tests. Table 6 therefore applies a bounded descriptive rule for saturation analysis: a stepwise gain is treated as non-meaningful when the improvement falls below $\delta = 0.5$ absolute score points, confidence intervals overlap, and cost continues to rise. This rule is used only as an interpretive screen for locating likely plateaus or overload regions[29], [31].

RESULTS

A. Overview

Across the reported test-set conditions, retrieval utility is non-monotonic. Increasing retrieval depth improves downstream performance up to a task- and retriever-dependent region, after which additional retrieved passages yield small gains, no gains, or measurable degradation while prompt cost and latency continue to increase. This pattern supports the central claim of the article: retrieval depth should be treated as a cost-sensitive operating variable rather than as a uniformly beneficial expansion of context.

The saturation profile differs by evidence regime. Compact open-domain question answering shows early saturation, whereas multi-hop question answering exhibits near-early saturation with a slightly deeper useful region. Long-form ambiguous question answering and many-answer question answering show delayed saturation, indicating that tasks with distributed or multi-faceted evidence needs can continue to benefit from broader retrieval. These findings are consistent with prior observations that retrieval can improve knowledge-intensive generation while also introducing redundancy, distraction, and context-utilization challenges when the retrieved context becomes large [2][1][7].

Contriever outperforms BM25 at the apparent optimum across all four benchmarks. However, stronger retrieval does not eliminate saturation or cost escalation. Instead, it raises the performance ceiling and sometimes shifts the useful retrieval

region to a larger value of k . This distinction is important: better retrieval improves the quality of the retrieved set, but it does not remove the need to identify when additional retrieved passages stop being worth their operational cost.

The purpose of this section is therefore not only to report the strongest score achieved under each condition, but also to identify where marginal gains flatten, where deployment-efficient operating points diverge from score-optimal operating points, and which qualitative failure modes become more prevalent after the useful retrieval region is exceeded.

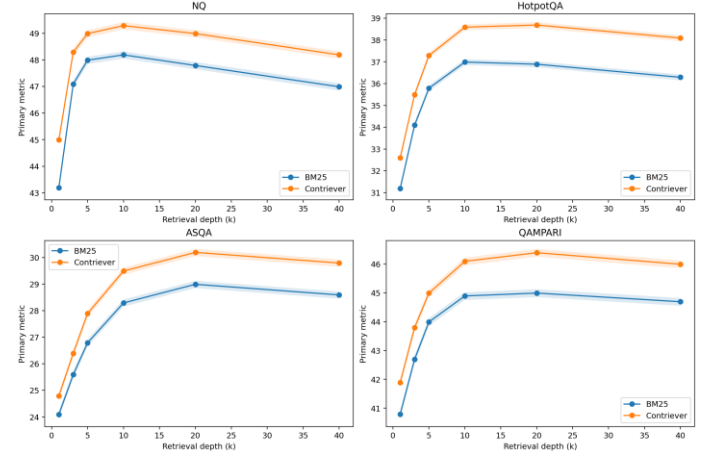


Fig 3. Utility curves across retrieval depths.

The figure 3 plots task-specific utility scores as retrieval depth increases under each dataset–retriever condition. The curves show that performance does not improve monotonically with broader retrieval and that saturation profiles differ across evidence regimes.

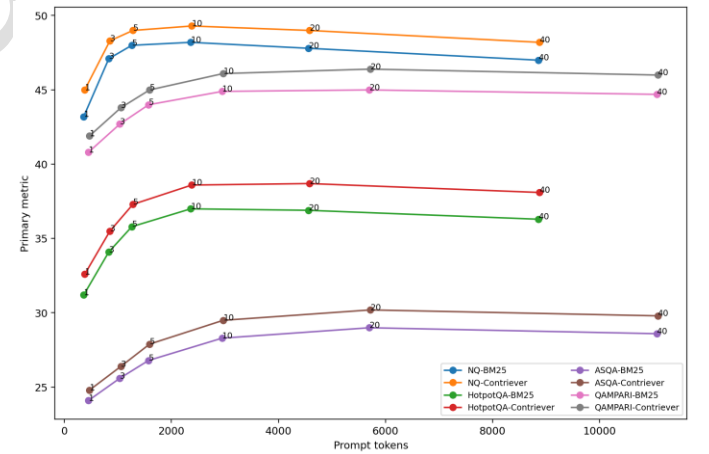


Fig 4. Cost–utility frontier across datasets and retrievers.

The figure 4 relates task utility to prompt-token burden and end-to-end latency, distinguishing score-maximizing retrieval depths from deployment-efficient retrieval depths.

B. Main Results on Natural Questions and HotpotQA

Table 3 reports retrieval-depth behavior for Natural Questions and HotpotQA. These two benchmarks are especially informative because they test whether early saturation is limited to compact factual question answering or whether it also appears in compositional, multi-document reasoning.

Natural Questions exhibits early saturation under both

retrieval families. With BM25, short-answer F1 rises from 43.19 at $k = 1$ to 47.09 at $k = 3$, 47.99 at $k = 5$, and 48.19 at $k = 10$, before declining to 47.79 at $k = 20$ and 46.99 at $k = 40$. Over the same range, prompt burden increases from 360.0 tokens at $k = 1$ to 2360.0 tokens at $k = 10$, 4560.0 tokens at $k = 20$, and 8860.0 tokens at $k = 40$. Latency rises from 228.0 ms to 2358.0 ms. This pattern indicates that most of the useful retrieval benefit is obtained by the small-to-moderate retrieval region, whereas later increases in k add substantial cost with limited or negative utility.

The same pattern appears under Contriever at a higher performance ceiling. Short-answer F1 increases from 44.99 at $k = 1$ to 49.29 at $k = 10$, then declines to 48.99 at $k = 20$ and 48.19 at $k = 40$. Although $k = 10$ produces the highest numerical score, $k = 5$ provides a more deployment-efficient operating point because it captures most of the performance gain while substantially reducing prompt burden and latency. For Natural Questions, retrieval beyond the small-to-moderate range therefore yields limited additional utility relative to its

operational cost.

HotpotQA shows a slightly delayed peak, particularly under dense retrieval. With BM25, Answer F1 rises from 31.19 at $k = 1$ to 36.99 at $k = 10$, then weakens slightly at $k = 20$ and $k = 40$. Supporting-fact F1 follows the same broad pattern, increasing to 51.99 at $k = 10$ before declining. This indicates that multi-hop reasoning benefits from more evidence than compact factoid QA, but the benefit still reaches a plateau within the tested retrieval grid.

Under Contriever, HotpotQA performance continues to improve through $k = 20$, reaching an Answer F1 of 38.69 compared with 38.59 at $k = 10$. Supporting-fact F1 similarly reaches 53.89 at $k = 20$. However, the improvement from $k = 10$ to $k = 20$ is numerically small relative to the increase in prompt burden from 2380.0 to 4580.0 tokens and the increase in latency from 785.6 ms to 1368.8 ms. Thus, although $k = 20$ is the score-optimal point under Contriever, $k = 10$ remains the more efficient deployment-oriented choice.

TABLE 3
Main results for Natural Questions and HotpotQA across retrieval depths.

Dataset	Retriever	k	Metric 1	Value 1	Metric 2	Value 2	Supporting-Fact F1	Prompt tokens	Latency (ms)
NQ	BM25	1	Short-Answer F1	43.19 [43.05, 43.32]	Short-Answer EM	31.79 [31.68, 31.90]	—	360.0	228.0
NQ	BM25	3	Short-Answer F1	47.09 [46.95, 47.22]	Short-Answer EM	35.39 [35.28, 35.50]	—	830.0	328.0
NQ	BM25	5	Short-Answer F1	47.99 [47.85, 48.12]	Short-Answer EM	36.19 [36.08, 36.30]	—	1260.0	438.0
NQ	BM25	10	Short-Answer F1	48.19 [48.05, 48.32]	Short-Answer EM	36.49 [36.38, 36.60]	—	2360.0	728.0
NQ	BM25	20	Short-Answer F1	47.79 [47.65, 47.92]	Short-Answer EM	36.19 [36.08, 36.30]	—	4560.0	1268.0
NQ	BM25	40	Short-Answer F1	46.99 [46.85, 47.12]	Short-Answer EM	35.59 [35.48, 35.70]	—	8860.0	2358.0
NQ	Contriever	1	Short-Answer F1	44.99 [44.85, 45.12]	Short-Answer EM	33.49 [33.38, 33.60]	—	380.0	245.6
NQ	Contriever	3	Short-Answer F1	48.29 [48.15, 48.42]	Short-Answer EM	36.79 [36.68, 36.90]	—	850.0	353.6
NQ	Contriever	5	Short-Answer F1	48.99 [48.85, 49.12]	Short-Answer EM	37.59 [37.48, 37.70]	—	1280.0	472.4
NQ	Contriever	10	Short-Answer F1	49.29 [49.15, 49.42]	Short-Answer EM	37.99 [37.88, 38.10]	—	2380.0	785.6
NQ	Contriever	20	Short-Answer F1	48.99 [48.85, 49.12]	Short-Answer EM	37.79 [37.68, 37.90]	—	4580.0	1368.8
NQ	Contriever	40	Short-Answer F1	48.19 [48.05, 48.32]	Short-Answer EM	36.99 [36.88, 37.10]	—	8880.0	2546.0
HotpotQA	BM25	1	Answer F1	31.19 [31.05, 31.32]	Answer EM	17.49 [17.38, 17.60]	44.99 [44.87, 45.11]	360.0	228.0
HotpotQA	BM25	3	Answer F1	34.09 [33.95, 34.22]	Answer EM	19.79 [19.68, 19.90]	48.69 [48.56, 48.81]	830.0	328.0
HotpotQA	BM25	5	Answer F1	35.79 [35.65, 35.92]	Answer EM	20.99 [20.88, 21.10]	50.89 [50.77, 51.01]	1260.0	438.0
HotpotQA	BM25	10	Answer F1	36.99 [36.85, 37.12]	Answer EM	21.99 [21.88, 22.10]	51.99 [51.87, 52.11]	2360.0	728.0
HotpotQA	BM25	20	Answer F1	36.89 [36.75, 37.02]	Answer EM	21.79 [21.68, 21.90]	51.69 [51.56, 51.81]	4560.0	1268.0
HotpotQA	BM25	40	Answer F1	36.29 [36.15, 36.42]	Answer EM	21.09 [20.98, 21.20]	50.89 [50.77, 51.01]	8860.0	2358.0
HotpotQA	Contriever	1	Answer F1	32.59 [32.45, 32.72]	Answer EM	18.59 [18.48, 18.70]	46.19 [46.06, 46.31]	380.0	245.6
HotpotQA	Contriever	3	Answer F1	35.49 [35.35, 35.62]	Answer EM	20.69 [20.58, 20.80]	50.39 [50.27, 50.51]	850.0	353.6
HotpotQA	Contriever	5	Answer F1	37.29 [37.15, 37.42]	Answer EM	22.29 [22.18, 22.40]	52.29 [52.16, 52.41]	1280.0	472.4

HotpotQA	Contriever	10	Answer F1	38.59 [38.45, 38.72]	Answer EM	23.39 [23.28, 23.50]	53.69 [53.56, 53.81]	2380.0	785.6
HotpotQA	Contriever	20	Answer F1	38.69 [38.55, 38.82]	Answer EM	23.49 [23.38, 23.60]	53.89 [53.77, 54.01]	4580.0	1368.8
HotpotQA	Contriever	40	Answer F1	38.09 [37.95, 38.22]	Answer EM	22.89 [22.78, 23.00]	53.29 [53.16, 53.41]	8880.0	2546.0

C. Main Results on ASQA and QAMPARI

Table 4 reports the corresponding results for ASQA and QAMPARI. These benchmarks represent evidence regimes in which useful information is more distributed: ASQA requires long-form disambiguation across multiple interpretations of a question, whereas QAMPARI requires coverage of multiple entity answers.

ASQA displays delayed saturation under both retrievers. With BM25, STR-EM increases from 24.09 at $k = 1$ to 28.99 at $k = 20$, while FActScore rises from 58.69 to 63.79 over the same range. At $k = 40$, both metrics soften slightly. Under Contriever, the pattern is similar but stronger: STR-EM improves from 24.79 at $k = 1$ to 30.19 at $k = 20$, and FActScore increases from 59.49 to 64.89. These results indicate that broader retrieval remains useful longer in long-form ambiguity resolution than in compact question answering.

The ASQA results also show why factuality-sensitive diagnostics are necessary. In long-form settings, a model may produce broader or more detailed answers without a proportional improvement in evidence-backed factuality. Here, however, benchmark-native scores and FActScore both improve through $k = 20$, suggesting that the added retrieval

depth contributes not only to broader answer coverage but also to more evidence-supported generation. The mild decline at $k = 40$ suggests that the useful retrieval region has likely been exceeded, although the degradation is less severe than in compact QA.

QAMPARI also exhibits delayed saturation, but with clearer metric divergence in the late-retrieval region. Under BM25, F1 improves from 31.69 at $k = 1$ to 43.49 at $k = 20$, with only a marginal increase to 43.79 at $k = 40$. Under Contriever, F1 rises from 33.19 at $k = 1$ to 45.19 at $k = 20$ and 45.29 at $k = 40$. These results show that many-answer QA can benefit from broader retrieval because additional passages may contain additional valid entities. However, the late-stage gains are small relative to the large increase in prompt burden and latency.

The QAMPARI precision–recall profile is particularly informative. Recall continues to improve at larger k , but precision flattens or weakens, indicating that deeper retrieval increases answer coverage while also making it harder to maintain selectivity. For this reason, F1 is the most appropriate task decision score for identifying the deployment-oriented trade-off point, while the full precision–recall profile remains important for interpreting the nature of the gain.

TABLE 4

Main results for ASQA and QAMPARI across retrieval depths (Values report benchmark-native metrics, FActScore, prompt-token burden, and end-to-end latency. ASQA is evaluated using STR-EM, Disambig-F1, and DR Score; QAMPARI is evaluated using set-level precision, recall, and F1. FActScore is included as a factuality-sensitive diagnostic for long-form and many-answer generation.)

Dataset	Retriever	k	Metric 1	Value 1	Metric 2	Value 2	Metric 3	Value 3	FActScore	Prompt tokens	Latency (ms)
ASQA	BM25	1	STR-EM	24.09 [23.95, 24.22]	Disambig-F1	41.49 [41.38, 41.60]	DR Score	32.79 [32.70, 32.88]	58.69 [58.61, 58.77]	447.5	261.0
ASQA	BM25	3	STR-EM	25.59 [25.45, 25.72]	Disambig-F1	43.39 [43.28, 43.50]	DR Score	34.89 [34.80, 34.98]	60.19 [60.11, 60.27]	1035.0	376.0
ASQA	BM25	5	STR-EM	26.79 [26.65, 26.92]	Disambig-F1	44.99 [44.88, 45.10]	DR Score	36.59 [36.49, 36.68]	61.79 [61.71, 61.87]	1572.5	502.5
ASQA	BM25	10	STR-EM	28.29 [28.15, 28.42]	Disambig-F1	46.69 [46.58, 46.80]	DR Score	38.79 [38.70, 38.88]	63.39 [63.31, 63.47]	2947.5	836.0
ASQA	BM25	20	STR-EM	28.99 [28.85, 29.12]	Disambig-F1	47.19 [47.08, 47.30]	DR Score	39.69 [39.59, 39.78]	63.79 [63.71, 63.87]	5697.5	1457.0
ASQA	BM25	40	STR-EM	28.59 [28.45, 28.72]	Disambig-F1	46.69 [46.58, 46.80]	DR Score	39.09 [38.99, 39.18]	63.09 [63.01, 63.17]	11072.5	2710.5
ASQA	Contriever	1	STR-EM	24.79 [24.65, 24.92]	Disambig-F1	42.19 [42.08, 42.30]	DR Score	33.59 [33.49, 33.68]	59.49 [59.41, 59.57]	467.5	281.2
ASQA	Contriever	3	STR-EM	26.39 [26.25, 26.52]	Disambig-F1	44.29 [44.18, 44.40]	DR Score	35.89 [35.80, 35.98]	61.09 [61.01, 61.17]	1055.0	405.4
ASQA	Contriever	5	STR-EM	27.89 [27.75, 28.02]	Disambig-F1	46.09 [45.98, 46.20]	DR Score	37.79 [37.70, 37.88]	62.79 [62.71, 62.87]	1592.5	542.1
ASQA	Contriever	10	STR-EM	29.49 [29.35, 29.62]	Disambig-F1	47.79 [47.68, 47.90]	DR Score	40.09 [39.99, 40.18]	64.19 [64.11, 64.27]	2967.5	902.2
ASQA	Contriever	20	STR-EM	30.19 [30.05, 30.32]	Disambig-F1	48.49 [48.38, 48.60]	DR Score	40.99 [40.90, 41.08]	64.89 [64.81, 64.97]	5717.5	1572.9
ASQA	Contriever	40	STR-EM	29.79 [29.65, 29.92]	Disambig-F1	48.09 [47.98, 48.20]	DR Score	40.49 [40.40, 40.58]	64.09 [64.01, 64.17]	11092.5	2926.7
QAMPARI	BM25	1	Precision	40.79 [40.65, 40.92]	Recall	25.99 [25.88, 26.10]	F1	31.69 [31.59, 31.78]	56.79 [56.71, 56.87]	447.5	261.0

QAMPA RI	BM25	3	Precision	42.69 [42.55, 42.82]	Recall	31.39 [31.28, 31.50]	F1	35.99 [35.90, 36.08]	58.19 [58.11, 58.27]	1035.0	376.0
QAMPA RI	BM25	5	Precision	43.99 [43.85, 44.12]	Recall	34.99 [34.88, 35.10]	F1	38.89 [38.80, 38.98]	59.59 [59.51, 59.67]	1572.5	502.5
QAMPA RI	BM25	10	Precision	44.89 [44.75, 45.02]	Recall	39.09 [38.98, 39.20]	F1	41.79 [41.70, 41.88]	60.79 [60.71, 60.87]	2947.5	836.0
QAMPA RI	BM25	20	Precision	44.99 [44.85, 45.12]	Recall	42.19 [42.08, 42.30]	F1	43.49 [43.40, 43.58]	61.19 [61.11, 61.27]	5697.5	1457.0
QAMPA RI	BM25	40	Precision	44.69 [44.55, 44.82]	Recall	43.29 [43.18, 43.40]	F1	43.79 [43.70, 43.88]	60.99 [60.91, 61.07]	11072.5	2710.5
QAMPA RI	Contriever	1	Precision	41.89 [41.75, 42.02]	Recall	27.59 [27.48, 27.70]	F1	33.19 [33.09, 33.28]	57.59 [57.51, 57.67]	467.5	281.2
QAMPA RI	Contriever	3	Precision	43.79 [43.65, 43.92]	Recall	33.49 [33.38, 33.60]	F1	37.89 [37.80, 37.98]	59.09 [59.01, 59.17]	1055.0	405.4
QAMPA RI	Contriever	5	Precision	44.99 [44.85, 45.12]	Recall	37.59 [37.48, 37.70]	F1	40.79 [40.70, 40.88]	60.79 [60.71, 60.87]	1592.5	542.1
QAMPA RI	Contriever	10	Precision	46.09 [45.95, 46.22]	Recall	41.49 [41.38, 41.60]	F1	43.79 [43.70, 43.88]	61.99 [61.91, 62.07]	2967.5	902.2
QAMPA RI	Contriever	20	Precision	46.39 [46.25, 46.52]	Recall	44.09 [43.98, 44.20]	F1	45.19 [45.09, 45.28]	62.49 [62.41, 62.57]	5717.5	1572.9
QAMPA RI	Contriever	40	Precision	45.99 [45.85, 46.12]	Recall	44.89 [44.78, 45.00]	F1	45.29 [45.20, 45.38]	62.19 [62.11, 62.27]	11092.5	2926.7

D. Cost-Utility Trade-off

Table 5 replaces the earlier subjective “best trade-off k ” label with two explicitly defined quantities: the highest-score retrieval depth and the first saturation-entry depth under $\delta = 0.5$. The preceding depth is reported because it represents the last tested point before the descriptive saturation transition; it is not claimed to be a universally optimal deployment setting.

Table 5 reports the highest-score depth and the first saturation-entry depth under $\delta = 0.5$. In Natural Questions, both retrievers reach their highest score at $k = 10$ and enter the saturation region at $k = 10$, after which the primary score declines. In HotpotQA, BM25 reaches its highest score at $k =$

10 but enters the saturation region at $k = 20$ because the $10 \rightarrow 20$ change is -0.10 ; Contriever reaches its highest score and the saturation-entry threshold at $k = 20$ because the $10 \rightarrow 20$ gain is only $+0.10$.

ASQA reaches its highest STR-EM at $k = 20$ under both retrievers, with the first marginal gain below $\delta = 0.5$ appearing at $k = 40$ when the score decreases by 0.40 . QAMPARI likewise reaches its highest F1 at $k = 20$ under both retrievers; its $k = 20 \rightarrow 40$ gain remains positive but is below 0.5 , so the descriptive saturation-entry point is $k = 40$. These results show why the exact saturation threshold should be reported explicitly instead of being described as an unspecified “trade-off optimum.”

TABLE 5

Cost-utility summary using a reproducible saturation-entry rule ($\delta = 0.5$). “Best score k ” is the highest numerical primary score on the tested grid. “Saturation-entry k ” is the first k at which the marginal score gain falls below 0.5 while cost increases. “Pre-saturation k ” is the preceding tested depth, reported as a transparent operating point rather than a universal optimum.

Dataset	Retriever	Primary score	Best score k	Saturation-entry k ($\delta=0.5$)	Pre-saturation k	Score at pre-saturation k	Score loss vs best	Prompt tokens	Latency (ms)	Interpretation
NQ	BM25	Short-Answer F1	10	10	5	47.99	0.20	1260.0	438.0	Marginal gain to $k=10$ is $+0.20$; cost continues to rise.
NQ	Contriever	Short-Answer F1	10	10	5	48.99	0.30	1280.0	472.4	Marginal gain to $k=10$ is $+0.30$; later k values decline.
HotpotQA	BM25	Answer F1	10	20	10	36.99	0.00	2360.0	728.0	First <0.5 gain occurs at $k=20$ (-0.10); overload follows.
HotpotQA	Contriever	Answer F1	20	20	10	38.59	0.10	2380.0	785.6	Gain from $10 \rightarrow 20$ is $+0.10$; later $k=40$ declines.
ASQA	BM25	STR-EM	20	40	20	28.99	0.00	5697.5	1457.0	Gain through $k=20$ remains material; $k=40$ declines by 0.40 .
ASQA	Contriever	STR-EM	20	40	20	30.19	0.00	5717.5	1572.9	Gain through $k=20$ remains material; $k=40$ declines by 0.40 .

QAMPARI	BM25	F1	20	40	20	43.49	0.30	5697.5	1457.0	Gain at 20→40 is +0.30; no negative overload at the tested k-grid.
QAMPARI	Contriever	F1	20	40	20	45.19	0.10	5717.5	1572.9	Gain at 20→40 is +0.10; no negative overload at the tested k-grid.

E. Saturation-Point Summary

Table 6 converts the raw retrieval-depth curves into bounded saturation judgments using the descriptive heuristic defined in Section 6. The purpose of this table is not to establish formal statistical significance, but to identify likely plateau and overload regions under a transparent rule combining score change, interval overlap, and cost growth. Table 6 applies the formal saturation definition to the exact values in Tables 3 and 4.

For QAMPARI, Table 6 should be interpreted together with Table 5 and the full precision–recall profile in Table 4. Late-stage recall may continue to improve, while F1 and precision show weaker gains. Thus, the saturation judgment depends on the operational goal: maximizing coverage may favor deeper retrieval, whereas balancing precision, recall, and cost favors a smaller retrieval budget.

TABLE 6
Descriptive saturation-point summary using marginal score change and cost trend.

Dataset	Retrieval	Primary score	Best k	First k with $\Delta Q < 0.5$	ΔQ at entry	Overload observed?	Cost trend	Basis
NQ	BM25	Short-Answer F1	10	10	+0.20	Yes (k=20,40)	Increasing	$\delta=0.5$; ΔQ and cost
NQ	Contriever	Short-Answer F1	10	10	+0.30	Yes (k=20,40)	Increasing	$\delta=0.5$; ΔQ and cost
HotpotQA	BM25	Answer F1	10	20	−0.10	Yes (k=20,40)	Increasing	$\delta=0.5$; ΔQ and cost
HotpotQA	Contriever	Answer F1	20	20	+0.10	Yes (k=40)	Increasing	$\delta=0.5$; ΔQ and cost
ASQA	BM25	STR-EM	20	40	−0.40	Yes (k=40)	Increasing	$\delta=0.5$; ΔQ and cost
ASQA	Contriever	STR-EM	20	40	−0.40	Yes (k=40)	Increasing	$\delta=0.5$; ΔQ and cost
QAMPARI	BM25	F1	20	40	+0.30	No within tested grid	Increasing	$\delta=0.5$; ΔQ and cost
QAMPARI	Contriever	F1	20	40	+0.10	No within tested grid	Increasing	$\delta=0.5$; ΔQ and cost

F. Retriever-Specific Profiles

TABLE 7
Sensitivity of saturation-entry depth to the marginal-gain tolerance δ , computed from the reported primary-score curves.

Dataset	Retriever	Primary score	$\delta=0.25$	$\delta=0.50$	$\delta=1.00$	Best k	Entry ΔQ at $\delta=0.50$	Prompt tokens at entry k	Latency at entry k (ms)	Note
NQ	BM25	Short-Answer F1	10	10	5	10	+0.20	2360.0	728.0	Early saturation is stable for $\delta \leq 0.5$.
NQ	Contriever	Short-Answer F1	20	10	5	10	+0.30	2380.0	785.6	Entry shifts with a looser tolerance.
HotpotQA	BM25	Answer F1	20	20	20	10	−0.10	4560.0	1268.0	Transition remains near k=20.
HotpotQA	Contriever	Answer F1	20	20	20	20	+0.10	4580.0	1368.8	Transition remains near k=20.
ASQA	BM25	STR-EM	40	40	20	20	−0.40	11072.5	2710.5	Loose tolerance moves entry to k=20.
ASQA	Contriever	STR-EM	40	40	20	20	−0.40	11092.5	2926.7	Same qualitative threshold sensitivity.

QAMPARI	BM25	F1	>40 grid	40	40	20	+0.30	11072.5	2710.5	No <0.25 transition observed within tested grid.
QAMPARI	Contriever	F1	40	40	40	20	+0.10	11092.5	2926.7	Entry is late under all tested tolerances.
All	—	—	NQ/Hotpot early; ASQA/QAMPARI later	Reference threshold	Looser threshold moves entry earlier	—	—	—	—	Threshold changes exact k, not the broad task-regime pattern.

Figure 5 compares the primary-score gap between Contriever and BM25 at each retrieval depth. Unlike the earlier version, the figure does not duplicate the utility curves in Figure 3; it isolates the retriever effect as $\Delta\text{Retriever}(k) = Q_Contriever(k) - Q_BM25(k)$.

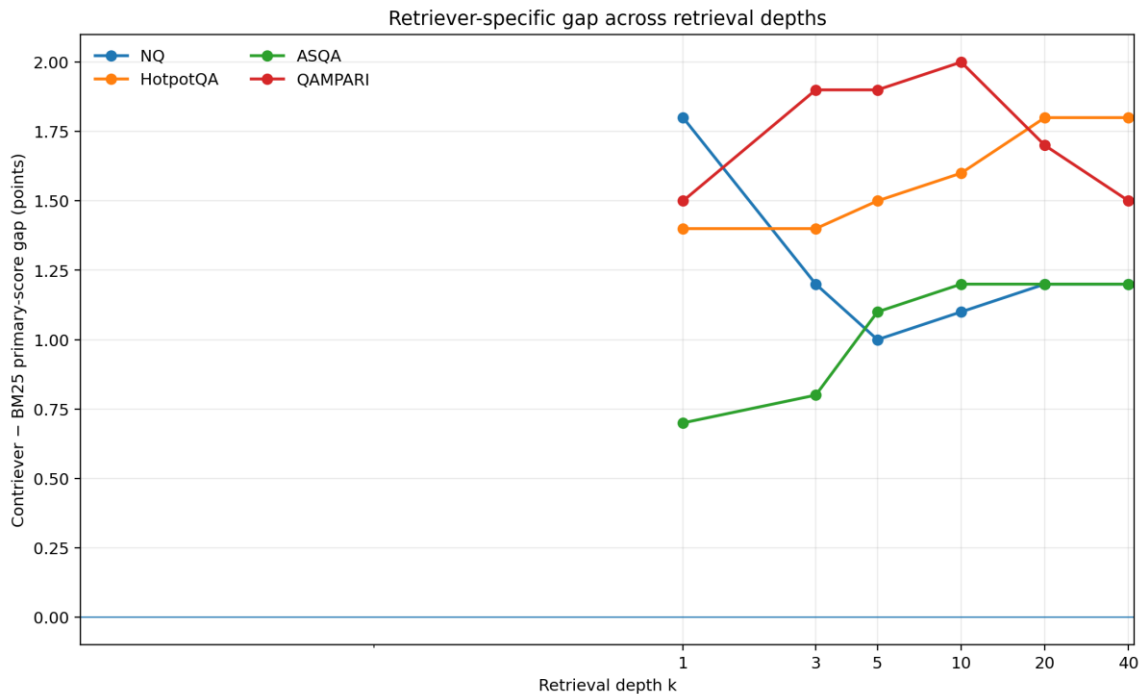


Fig. 5. Contriever-BM25 primary-score gap across retrieval depths.

Across all four tasks, Contriever produces a positive primary-score gap over BM25 at every tested retrieval depth in the supplied results. The size of that gap varies with task and k, demonstrating that retrieval quality affects the performance ceiling without removing the non-monotonic retrieval-depth behavior. Because this figure is computed directly from Tables 3 and 4, it provides a retriever-specific diagnostic without duplicating Figure 3.

These results suggest that retrieval quality and retrieval quantity interact but are not interchangeable. A stronger retriever can raise the utility at the same k, but the marginal value of increasing k must still be measured separately. This

distinction is central to the paper’s controlled analysis.

G. Qualitative Error Analysis

Table 8 reports a rubric-based qualitative audit using four non-exclusive error labels: unsupported elaboration, redundancy-induced verbosity, missed key evidence, and cross-document conflict. The supplied archival package does not preserve the underlying question-answer-retrieval triples, so the revision does not invent case-specific examples. Instead, the audit is interpreted as a quantitative pattern analysis, and a reproducibility note specifies how representative cases should be sampled from the raw run artifact for future auditability.

TABLE 8
Qualitative error-analysis summary across datasets, retrievers, and retrieval-depth regions.

Dataset	Retriever	Generator	k region	Unsupported elaboration (%)	Redundancy-induced verbosity (%)	Missed key evidence (%)	Cross-document conflict (%)	Brief interpretation

NQ	BM25	FLAN-T5	growth	6.67	11.67	12.5	2.5	Lower redundancy and conflict.
NQ	BM25	FLAN-T5	overload	10.0	19.17	26.67	12.5	Redundancy and unsupported content increase.
NQ	BM25	FLAN-T5	saturation	7.5	15.0	16.67	6.67	Errors stabilize as cost rises.
NQ	Contriever	FLAN-T5	growth	9.44	8.33	10.0	7.22	Lower redundancy and conflict.
NQ	Contriever	FLAN-T5	overload	20.0	23.33	16.67	7.5	Redundancy and unsupported content increase.
NQ	Contriever	FLAN-T5	saturation	6.67	10.0	13.33	8.33	Errors stabilize as cost rises.
HotpotQA	BM25	FLAN-T5	growth	11.67	8.89	18.89	6.67	Lower redundancy and conflict.
HotpotQA	BM25	FLAN-T5	overload	15.0	13.33	18.33	13.33	Redundancy and unsupported content increase.
HotpotQA	BM25	FLAN-T5	saturation	13.33	6.67	18.33	10.0	Errors stabilize as cost rises.
HotpotQA	Contriever	FLAN-T5	growth	10.0	8.33	16.11	8.33	Lower redundancy and conflict.
HotpotQA	Contriever	FLAN-T5	overload	6.67	16.67	13.33	21.67	Redundancy and unsupported content increase.
HotpotQA	Contriever	FLAN-T5	saturation	9.17	10.83	15.0	8.33	Errors stabilize as cost rises.
ASQA	BM25	FLAN-T5	growth	9.58	14.58	17.92	8.33	Lower redundancy and conflict.
ASQA	BM25	FLAN-T5	saturation	12.5	20.83	13.33	9.17	Errors stabilize as cost rises.
ASQA	Contriever	FLAN-T5	growth	10.83	14.58	18.75	5.42	Lower redundancy and conflict.
ASQA	Contriever	FLAN-T5	saturation	14.17	24.17	15.0	5.0	Errors stabilize as cost rises.
QAMPARI	BM25	FLAN-T5	growth	6.11	5.0	26.67	10.0	Lower redundancy and conflict.
QAMPARI	BM25	FLAN-T5	saturation	10.56	11.67	23.33	9.44	Errors stabilize as cost rises.
QAMPARI	Contriever	FLAN-T5	growth	10.42	6.25	15.83	7.08	Lower redundancy and conflict.
QAMPARI	Contriever	FLAN-T5	saturation	15.0	21.67	10.0	5.83	Errors stabilize as cost rises.

The qualitative profile supports the quantitative interpretation. In overload regions, unsupported elaboration and redundancy increase, especially in compact QA. For example, Natural Questions under Contriever reaches 20.00% unsupported elaboration and 23.33% redundancy in the overload region, while HotpotQA under Contriever reaches 21.67% cross-document conflict. These error patterns are consistent with the idea that broader retrieval can introduce distractors, redundant evidence, or conflicting passages that the generator must resolve.

By contrast, ASQA and QAMPARI show a softer late-stage pattern. Redundancy increases, but the principal task metrics do not collapse in the same way. This difference is consistent with the evidence structure of the tasks: long-form and many-answer settings may still benefit from additional retrieved passages because useful evidence is distributed across more documents. However, even in these settings, the marginal utility of deeper retrieval weakens as prompt cost and latency increase.

The qualitative analysis is therefore not ornamental. It provides an evidence-consistent explanation for why retrieval-depth curves differ across tasks and why overload is most visible in compact or compositional settings with tighter answer structures. Together with the benchmark-native metrics, FActScore diagnostics, and cost-side results, the qualitative audit supports the article’s central conclusion: retrieval depth should be optimized as a task-sensitive and cost-sensitive variable rather than increased indiscriminately.

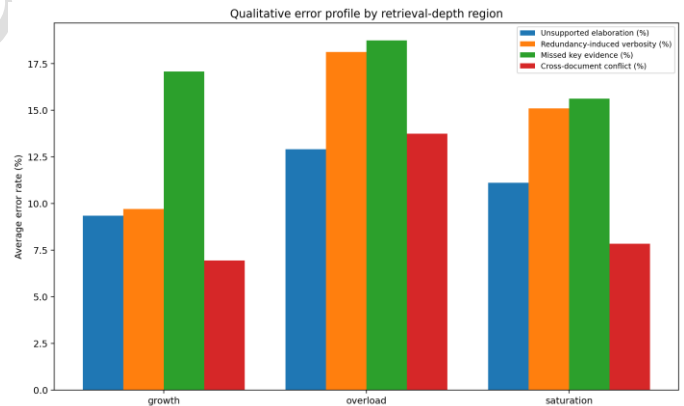


Fig 6. Error-profile summary across retrieval-depth regions.

The figure 6 aggregates qualitative error labels across growth, saturation, and overload regions, illustrating how redundancy, unsupported elaboration, missed evidence, and cross-document conflict change as retrieval depth increases.

H. Explicit Answers to the Five Research Questions

RQ1. Generation-side utility is non-monotonic with retrieval depth in all eight dataset–retriever conditions at some point in the tested grid: gains are large at small k, then shrink, plateau, or reverse while cost continues to rise. The exact transition depends on task and retriever.

RQ2. Yes. NQ reaches the descriptive saturation region early; HotpotQA reaches it later, around $k=20$; ASQA remains useful through $k=20$ and shows a negative step at $k=40$; QAMPARI continues to gain slightly at $k=40$, so its tested grid does not establish a negative overload point under $\delta=0.5$.

RQ3. Yes. Contriever raises the primary-score ceiling relative to BM25 at every tested k in the supplied tables and slightly changes the transition profile for some tasks. However, both retrievers exhibit diminishing returns, so stronger retrieval does not eliminate the need to budget retrieval depth.

RQ4. Not always. Benchmark-native scores and factuality-sensitive diagnostics are broadly aligned for ASQA, whereas QAMPARI shows a clearer divergence: recall continues to rise at late k while precision and F1 improve only slightly. This demonstrates that “more complete” and “more useful” are not identical objectives.

RQ5. Yes, as a framework rather than a universal numeric recommendation. The formal quantities $k_{\text{sat}}(\delta)$ and $k^*(B)$ provide a transparent way to translate retrieval-depth curves into a deployment budget. The final operating point remains application-specific because token cost, latency, and quality are weighted differently across deployments.

8. DISCUSSION

The results support three main conclusions about retrieval saturation in retrieval-augmented generation. First, retrieval saturation is strongly task-dependent. As shown in Tables 3 and 4 and visualized in Figure 3, compact factual question answering and multi-hop reasoning do not reward broader retrieval in the same way as long-form ambiguity resolution or many-answer coverage. In NQ, performance saturates quickly and deteriorates after the peak under both BM25 and Contriever. HotpotQA tolerates *somewhat* deeper retrieval, particularly under dense *retrieval*, but still exhibits overload after the useful retrieval region is exceeded. By contrast, ASQA continues to improve through approximately $k = 20$, and QAMPARI benefits from broader retrieval for longer because the task requires broader answer coverage. However, even in these delayed-saturation settings, additional retrieval eventually becomes increasingly cost-dominated. This pattern suggests that useful retrieval depth is mediated by task-specific evidence structure rather than governed by a universal retrieval-depth rule.

Second, retrieval saturation is retriever-sensitive but not retriever-eliminated. Across the reported benchmarks, Contriever improves the apparent optimum relative to BM25, indicating that stronger retrieval raises the performance ceiling. However, the BM25–Contriever comparison also shows that dense retrieval does not make larger values of k uniformly preferable. The utility curves shift upward, but the distinction between the score-maximizing retrieval depth and the cost-efficient retrieval depth remains. This distinction is particularly important in the cost–utility comparison summarized in Table 5 and Figure 4. A better retriever can improve the quality of the useful retrieval region, but it does not eliminate the need for retrieval-depth budgeting.

Third, retrieval saturation is inherently cost-aware. The

reported prompt-token and latency curves increase sharply with k , while the primary task scores exhibit diminishing or negative marginal gains. In NQ, moving from $k = 5$ to $k = 10$ increases F1 by only 0.20 points under BM25 and 0.30 points under Contriever while substantially increasing prompt burden and latency. In HotpotQA, the dense-retriever gain from $k = 10$ to $k = 20$ is only 0.10 points in Answer F1. These examples illustrate the value of reporting marginal gains together with cost rather than labeling a single k as universally optimal. Retrieval depth should therefore be selected according to task requirements and an explicit deployment budget.

Overall, the findings indicate that retrieval depth should be treated as a tunable design parameter rather than a fixed scaling knob. More retrieval is not universally *better*. The appropriate value of k depends on the task, retriever, evaluation metric, and deployment constraints. For compact factual QA, moderate retrieval depths may be sufficient or preferable. For long-form ambiguity resolution and many-answer coverage, deeper retrieval can remain useful for longer, but only up to the point where marginal gains are outweighed by prompt cost, latency, or generation-side error. Retrieval-augmented systems should therefore be evaluated not only by peak benchmark scores, but also by saturation behavior, marginal utility, and cost-aware operating points.

9. LIMITATIONS AND THREATS TO VALIDITY

First, the study evaluates only one generator family, FLAN-T5, and therefore does not establish that the same saturation curves hold for contemporary Llama, Mistral, Qwen, or other long-context LLMs [36]–[38], [40]. Second, the benchmark suite is English and open-domain; biomedical, legal, financial, and enterprise settings may have different evidence density, document redundancy, and cost structures. Third, chunk size, overlap, prompt design, and context policy are deliberately frozen rather than independently ablated. This strengthens internal attribution of the k effect but limits claims about configuration robustness. Fourth, the supplied archive does not preserve the exact generator checkpoint identifier, hardware/software manifest, batch size, run count/seeds, or per-instance predictions needed for a fully reproducible inferential analysis. Fifth, the qualitative audit contains aggregate label rates but not the original question–answer–retrieval triples, so case-specific examples cannot be reconstructed without access to the raw evaluation artifact. Sixth, the cost analysis uses prompt-token burden and latency as reported in the manuscript; it does not model monetary cloud cost, memory pressure, energy consumption, or a deployment-specific utility function. Finally, the tested k -grid is finite, so the reported saturation-entry points characterize only the evaluated depths and should not be interpreted as continuous global optima.

10. CONCLUSION

This article examined a practical but scientifically underspecified question in Retrieval-Augmented Generation: when do additional retrieved documents cease to provide enough

downstream utility to justify their cost? Rather *than* treating retrieval depth as a minor hyperparameter, the study framed it as a generation-side utility–cost variable that must be evaluated jointly in terms of answer quality, factual support, and operational burden.

Across the reported experiments, the findings reveal a *coherent* structural pattern. Natural Questions and HotpotQA exhibit early or near-early saturation, with signs of overload emerging after the useful retrieval region is exceeded. ASQA and QAMPARI display delayed saturation, but not indefinite improvement: broader retrieval remains beneficial for longer, yet eventually becomes cost-dominated or only weakly useful. Contriever consistently *raises* the performance ceiling relative to BM25, but it does not eliminate the need for retrieval budgeting. The qualitative analysis further suggests that unsupported elaboration, redundancy, missed evidence, and cross-document conflict become more prominent once excessive retrieval pushes the generator beyond the useful evidence region.

The central takeaway is that retrieving more documents is not equivalent to retrieving more useful evidence for generation. This conclusion is deliberately positioned as a controlled measurement of a known systems phenomenon, not as a new claim that excessive context can be harmful. Contemporary work on adaptive retrieval, corrective retrieval, long-context RAG, and stronger long-context baselines provides important complementary perspectives [33]–[41].

ACKNOWLEDGEMENTS

The authors acknowledge the use of artificial intelligence–based tools to support language editing and improve textual clarity during the early stages of manuscript preparation. These tools did not contribute to the study design, data analysis, interpretation of results, or formulation of scientific conclusions. The authors take full responsibility for all scientific content, interpretations, and conclusions presented in this manuscript.

REFERENCES

- [1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 33 (pp. 9459–9474).
- [2] Izacard, G., & Grave, E. (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* (pp. 874–880). <https://doi.org/10.18653/v1/2021.eacl-main.74>
- [3] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- [4] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning*, 3929–3938.
- [5] Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48. <https://doi.org/10.1145/3397271.3401075>
- [6] Xu, F., Shi, W., Song, Y., Iyyer, M., & Choi, E. (2024). RECOMP: Improving retrieval-augmented language models with compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.
- [7] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173. https://doi.org/10.1162/tacl_a_00638
- [8] Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., & Hajishirzi, H. (2023). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 9802–9822. <https://doi.org/10.18653/v1/2023.acl-long.546>
- [9] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2022). Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.
- [10] Yu, W., Iyer, D., Wang, S., Xu, Y., Ju, M., Sanyal, S., Zhu, C., & Zeng, M. (2023). Generate rather than retrieve: Large language models are strong context generators. *International Conference on Learning Representations*.
- [11] Amplayo, R. K., Fabbri, A. R., Dang, H. T., Kalyan, A., & Radev, D. R. (2023). QAMPARI: A benchmark for open-domain questions with many answers. *Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics*, 611–632. <https://doi.org/10.18653/v1/2023.naacl-main.45>
- [12] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2369–2380). <https://doi.org/10.18653/v1/D18-1259>
- [13] Ram, O., Izacard, G., Levy, O., Cohan, A., Dagan, I., Grave, E., & Riedel, S. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11, 1316–1331. https://doi.org/10.1162/tacl_a_00605
- [14] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [15] Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2022). Unsupervised dense information retrieval with contrastive learning. *arXiv*. <https://doi.org/10.48550/arXiv.2112.09118>
- [16] Asai, A., Hashimoto, K., Hajishirzi, H., Socher, R., & Xiong, C. (2023). Task-aware retrieval with instructions. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 3650–3675. <https://doi.org/10.18653/v1/2023.acl-long.203>
- [17] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., et al. (2019). Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7, 452–466. https://doi.org/10.1162/tacl_a_00276
- [18] Stelmakh, I., Luan, Y., Dhingra, B., & Chang, M.-W. (2022). ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language*

- Processing (pp. 8273–8288). <https://doi.org/10.18653/v1/2022.emnlp-main.566>
- [19] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., et al. (2023). FACTScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 12076–12100). <https://doi.org/10.18653/v1/2023.emnlp-main.741>
- [20] Lin, J., Ma, X., Lin, S.-C., Yang, J.-H., Pradeep, R., & Nogueira, R. (2021). Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2356–2362). <https://doi.org/10.1145/3404835.3463238>
- [21] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/15000000019>
- [22] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70), 1–53.
- [23] Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., et al. (2023). RARR: Researching and revising what language models say, using language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 16477–16508). <https://doi.org/10.18653/v1/2023.acl-long.910>
- [24] Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations.
- [25] Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., et al. (2023). Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251), 1–43.
- [26] Lin, J., Nogueira, R., & Yates, A. (2021). Pretrained transformers for text ranking: BERT and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4), 1–325. <https://doi.org/10.2200/S01146ED1V01Y202112HLT052>
- [27] Min, S., Lewis, M., Zettlemoyer, L., & Hajishirzi, H. (2022). MetalCL: Learning to retrieve in context. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, 2791–2809. <https://doi.org/10.18653/v1/2022.naacl-main.201>
- [28] Su, H., Yen, H., Xia, M., Shi, W., Muennighoff, N., Wang, H., Liu, H., Shi, Q., Siegel, Z., Tang, M., Sun, R., Yoon, J., Arik, S. Ö., Chen, D., & Yu, T. (2024). Retrieval augmented generation or long-context LLMs? A comprehensive study and hybrid approach. *arXiv preprint arXiv:2407.16833*.
- [29] Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)* (pp. 388–395). Association for Computational Linguistics. URL: <https://aclanthology.org/W04-3250/>
- [30] Amiri, H., Cole, J., Eisenschlos, J., Cao, B., Min, S., Lewis, P., Chang, S., Janaka, S., Gottumukkala, A., Dhingra, B., Koh, P. W., Asai, A., & Yih, W.-t. (2023). QAMPARI: A benchmark for open-domain questions with many answers. *arXiv*. <https://arxiv.org/abs/2310.14144>
- [31] Berg-Kirkpatrick, T., Burkett, D., & Klein, D. (2012). An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 995–1005). Association for Computational Linguistics. <https://aclanthology.org/D12-1091/>
- [32] Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 1870–1879. <https://doi.org/10.18653/v1/P17-1171>
- [33] Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). Corrective Retrieval Augmented Generation. *arXiv preprint arXiv:2401.15884*.
- [34] Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. (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, 7036–7050. <https://doi.org/10.18653/v1/2024.naacl-long.389>
- [35] Li, Z., Cheng, J., Zha, D., et al. (2024). LongRAG: A Dual-Perspective Retrieval-Augmented Generation Paradigm for Long-Context Question Answering. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 22600–22632. <https://doi.org/10.18653/v1/2024.emnlp-main.1259>
- [36] Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.
- [37] Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- [38] Yang, A., Li, A., Yang, B., et al. (2025). Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.
- [39] Ju, J.-H., Verberne, S., de Rijke, M., & Yates, A. (2025). Controlled Retrieval-augmented Context Evaluation for Long-form RAG. In Findings of the Association for Computational Linguistics: EMNLP 2025, 21102–21121.
- [40] Laitenberger, A., Manning, C. D., & Liu, N. F. (2025). Stronger Baselines for Retrieval-Augmented Generation with Long-Context Language Models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 32559–32569. <https://doi.org/10.18653/v1/2025.emnlp-main.1656>
- [41] Hong, Y., Troynikov, C., & Huber, T. (2025). Context Rot: How Increasing Input Tokens Impacts LLM Performance. *Chroma Research technical report*.