CHASE: CHAIN-GUIDED HYPOTHETICAL SUPPORTING-FACT GENERATION FOR GRAPH-FREE MULTI-HOP DENSE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph-based retrieval-augmented generation (RAG) outperforms conventional dense-retrieval RAG on multi-hop questions by capturing cross-document connections through an explicit graph index, at the cost of offline graph construction and, for some systems, task-specific training. We ask whether the reasoning structure such a graph supplies can instead be elicited from the question itself at query time. We propose CHASE (Chain-guided Hypothetical Supporting-fact gEneration), a training-free framework in which a single LLM call infers the latent reasoning chain of a multi-hop question and emits one *hypothetical supporting fact* (HSF) per hop, written in the declarative form a supporting passage would take. Each HSF is issued as an independent dense query and the per-anchor results are consolidated by max-score fusion, so the generator produces a chain while the retriever consumes a set and all vector searches run in parallel. Under a matched first-stage protocol that holds the encoder, corpus and index fixed, HSF anchors raise Recall@2 over question-only retrieval by 13.9 points on MuSiQue, 11.1 on 2WikiMultiHopQA and 7.0 on HotpotQA, and anchor-aware reranking adds a further 9.5, 12.2 and 4.0 points over question-only reranking. Against published graph-based and iterative systems CHASE is competitive with GFM-RAG and IRCoT+HippoRAG, although those systems use different encoders, so we report that comparison as context rather than as a controlled result. We also map where the method applies: anchor generation calls for a generator above a capability threshold that, at Recall@2, no open model we tested at 35B parameters or below reaches on two of the three benchmarks, which makes distillation of anchor generation the clearest route to a deployable version.

1 INTRODUCTION

Retrieval-augmented generation is only as reliable as the evidence made available to its generator. The dependency is acute in multi-hop question answering, where a retriever must recover a set of interdependent passages from a question that gives only partial cues to the underlying reasoning chain. Two mismatches make this hard: bridge entities identifying later-hop evidence are often absent from the question, and a question expresses a global, relational information need while each supporting passage states a local, declarative fact. Consider the three-hop MuSiQue question (Trivedi et al., 2022) *"In what region of the country of S-Fone is the place of birth of John Phan located?"*, which requires passages establishing that *S-Fone operates in Vietnam*, that *John Phan was born in Da Nang*, and that *Da Nang lies in the South Central Coast*. Neither bridge entity nor the answer appears in the question, and because each passage realizes a different segment of the latent chain, similarity to the question alone may not rank every required passage highly. We call this mismatch between a global question representation and heterogeneous, hop-local evidence the *cross-hop representation gap*.

Existing paradigms address this gap at different points in the retrieval–reasoning tradeoff. One-shot dense retrieval is efficient, but a single query does not expose missing bridge entities or hop-level information needs. Iterative and agentic retrieval condition each query on previously retrieved evidence, buying adaptivity at the cost of sequential retrieval depth and the risk that an incorrect intermediate decision propagates to later hops (Qi et al., 2019; Trivedi et al., 2023; Shi et al., 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Graph-based RAG externalizes entities and relations in a persistent index, giving an explicit substrate for traversal at the cost of offline graph construction and, for some systems, task-specific training (Edge et al., 2025; Gutiérrez et al., 2024; Luo et al., 2025). A complementary question follows. Can an LLM infer the local evidence needs from the question alone, turning multi-hop search into a set of parallel retrieval operations over an unmodified passage index?

We introduce CHASE (Chain-guided Hypothetical Supporting-fact gEneration), a training-free framework that uses an LLM to externalize a question-conditioned retrieval plan: no parameter updates, no task-specific retriever, generator, encoder and corpus index all fixed. In one retrieval-stage LLM call, CHASE produces a variable-length, dependency-ordered sequence of **hypothetical supporting facts** (HSFs), each expressing one inferred evidence need in the declarative form likely to occur in a supporting passage. CHASE issues all HSFs as independent dense queries and applies max-score fusion to preserve the strongest HSF–passage match. The LLM generates a chain, but the retriever consumes a set: once the HSFs exist, all vector searches execute in parallel. HSFs are speculative anchors, not evidence; only passages retrieved from the corpus are passed downstream.

We separate two kinds of evidence throughout. Our claims rest on a *matched* protocol in which the encoder (BGE-M3), corpus, index and candidate budget are identical across conditions and only the query representation changes. Under that protocol, HSF anchors improve Recall@2 over question-only retrieval by 13.9, 11.1 and 7.0 points on MuSiQue, 2WikiMultiHopQA and HotpotQA, and anchor-aware reranking adds 9.5, 12.2 and 4.0 points over question-only reranking. We additionally report a cross-system comparison with published graph-based and iterative systems. Those systems use different encoders and retrieval budgets, so we present it as context rather than as a controlled result and claim no state of the art from it. A generator-scale study then locates the regime in which the method works: at Recall@2, every open generator we evaluated at 35B parameters or below stays below the matched question-only baseline on 2Wiki and HotpotQA, though on 2Wiki the deficit disappears once the cutoff is relaxed to Recall@5.

The main contributions are as follows:

1. **A representation-gap account of multi-hop retrieval.** We formulate evidence acquisition as coverage of heterogeneous, hop-local information needs and measure the resulting gap directly in embedding space, not only through end-task recall.
2. **A plan-once, retrieve-in-parallel framework.** CHASE elicits HSF anchors in one LLM call, retrieves with all anchors in parallel, fuses by max score, and reranks each candidate against its most compatible anchor, leaving the passage index unchanged.
3. **A matched evaluation separating the method from its components.** We isolate anchor form, fusion rule and reranking strategy under a fixed encoder, and identify the generator-scale regime in which HSF augmentation does *not* pay off.

2 Related Work

Query-side approaches to dense retrieval. Dense passage retrievers map queries and passages into a shared embedding space, enabling efficient nearest-neighbor search over large corpora (Karpukhin et al., 2020; Lewis et al., 2020). Multi-Hop Dense Retrieval adapts this paradigm to compositional questions by conditioning later retrieval steps on evidence selected at earlier hops (Xiong et al., 2020; Izacard et al., 2022). Generative query expansion instead translates an underspecified information need into text that is easier to match. HyDE embeds a generated hypothetical document as the query, explicitly allowing it to be fictitious (Gao et al., 2023); related approaches generate context directly, augment a query with a hypothetical document, or steer the expansion with corpus statistics (Yu et al., 2022; Wang et al., 2023; Lei et al., 2024), while query rewriting and question decomposition produce alternative queries whose results are merged (Ma et al., 2023; Ammann et al., 2025). HyDE is the closest precedent but differs in granularity and objective: one global hypothetical document, against CHASE's multiple declarative anchors for distinct hops of a latent chain. CHASE also leaves indexing granularity unchanged, unlike proposition-level retrieval (Chen et al., 2024). Multi-query retrieval needs a consolidation rule, typically score aggregation or reciprocal rank fusion (Cormack et al., 2009), after which a cross-encoder can reassess the fused pool (Nogueira & Cho, 2020). Both are separable from HSF generation, so we evaluate them independently in Section 4.4.

Iterative and agentic retrieval. A second family closes the loop between reasoning and retrieval. Iterative query generation, Self-Ask and IRCoT expose intermediate reasoning states and use them to request additional evidence (Qi et al., 2019; Press et al., 2023; Trivedi et al., 2023); ReAct generalizes the pattern to actions in an external environment (Yao et al., 2023), and FLARE and Generate-then-Ground trigger retrieval from intermediate generations (Jiang et al., 2023; Shi et al., 2024). This feedback loop lets later queries incorporate newly discovered entities, but requires sequential model and retrieval steps, and an incorrect early step can redirect subsequent search. Learned multi-hop systems instead train a retriever to compose an evidence sequence (Zhuang et al., 2024; Zhang et al., 2024); their execution profiles differ from prompted LLM agents and should not be assigned a common latency cost. CHASE commits to all anchors before observing any retrieval result, which removes the dependency between retrieval calls but forgoes the ability to revise a poor anchor.

Graph-based RAG. Graph-based RAG externalizes relational structure in a persistent corpus index. GraphRAG constructs an entity graph and community summaries for query-focused summarization (Edge et al., 2025); HippoRAG combines a knowledge graph with Personalized PageRank, and GFM-RAG learns a graph foundation model for retrieval (Gutiérrez et al., 2024; Luo et al., 2025). Recent systems explore logic-aware graph construction, knowledge-graph-guided retrieval and graph-enhanced agents (Liu et al., 2025; Zhu et al., 2025; Shen et al., 2025). These systems differ in schema, extraction pipeline and supervision, so their costs and failure modes are implementation-dependent. CHASE does not seek to replace graph reasoning; it studies the complementary regime in which the passage index stays fixed and the LLM creates ephemeral, question-specific anchors instead of a persistent corpus graph.

3 METHOD

3.1 PROBLEM FORMULATION AND HYPOTHESES

Let q denote a question, $\mathcal{P} = \{p_1, \ldots, p_N\}$ a passage corpus, and $\mathcal{G}_q \subseteq \mathcal{P}$ the gold supporting passages for q. A retriever returns a ranked list whose top-k set is $\mathcal{R}(q; k)$. The dataset-level *Recall@*k metric is

$$\text{Recall@}k = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{|\mathcal{G}_q \cap \mathcal{R}(q; k)|}{|\mathcal{G}_q|}, \tag{1}$$

where $\mathcal{Q}$ is the evaluation question set.

Multi-hop retrieval presents two challenges: a *missing-entity gap*, since later-hop passages center on bridge entities absent from the question, and a *form gap*, since questions express relations or comparisons while passages state atomic declarative facts. In the S-Fone example, one dense query vector must match three passages whose local topics diverge from each other and from the global question. CHASE addresses both by generating hop-local anchors in the declarative form of a supporting passage, adaptively choosing how many; this number need not match the gold passage count or any annotated hop structure.

Two hypotheses motivate the design and are tested directly in Section 4. **H1 (representation gap).** For an appropriately generated anchor set H, the best anchor lies closer to the supporting passages than the question itself. Writing E_ϕ for the encoder, we define the gap

$$\Delta_{\text{sim}}(q) = \max_{h_i \in H} \mathbb{E}_{p \in \mathcal{G}_q}[\text{sim}(E_\phi(h_i), E_\phi(p))] - \mathbb{E}_{p \in \mathcal{G}_q}[\text{sim}(E_\phi(q), E_\phi(p))]\,, \tag{2}$$

and predict $\mathbb{E}_q[\Delta_{\text{sim}}(q)] > 0$ with $P(\Delta_{\text{sim}}(q) > 0) > 0.5$. This is an empirical claim about proximity in the encoder's space, not a requirement that anchors align one-to-one with gold passages. **H2 (capability threshold).** Producing useful anchors requires decomposing a compound information need, rendering each relation declaratively, and instantiating entities the question does not name. These abilities improve with generator capability. Following work on emergent abilities (Wei et al., 2022), we predict a threshold τ below which HSF augmentation gives no benefit over question-only retrieval with the same encoder. Appendix B gives the formal statements, and Appendix C reports a third, exploratory hypothesis on entity-level error tolerance that its own test supports only in part: retrieval degrades monotonically as HSF entities are corrupted, and anchors whose relational form is intact while their entities are all wrong retrieve worse than the question itself.

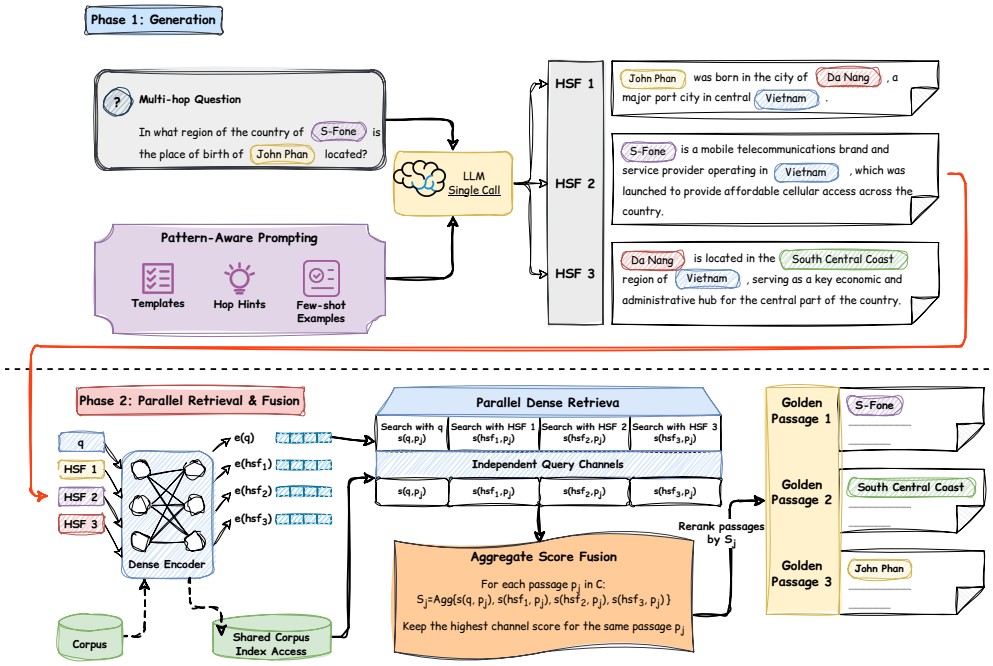


Figure 1: CHASE generates a dependency-ordered sequence of HSFs, retrieves independently with each HSF, and combines candidates by max-score fusion.

3.2 The CHASE Framework

CHASE separates question-conditioned planning from corpus-grounded evidence acquisition: an LLM proposes hop-level retrieval anchors, and a dense retriever executes the resulting searches independently and consolidates their candidates (Figure 1).

Stage 1: hypothetical supporting-fact generation. Given question q and prompt π, generator G_θ produces a variable-length, dependency-ordered sequence

$$H = G_\theta(q, \pi) = \{h_1, h_2, \ldots, h_m\}, \tag{3}$$

where each h_i describes one inferred local evidence need and m is chosen adaptively. The prompt supplies few-shot demonstrations and requests a dependency-consistent progression, so later HSFs may refer to entities hypothesized earlier; the generator receives no gold answer, gold passages, supporting-fact annotations or hop count. An HSF may posit a bridge entity or even the final answer, but such claims serve only as retrieval anchors and are never accepted as evidence (prompts in Appendix J).

Stage 2: parallel dense retrieval and max fusion. For each HSF h_i and passage p we compute $s(h_i, p) = \mathrm{sim}(E_\phi(h_i), E_\phi(p))$ and take the top-k passages per anchor, $\mathcal{R}(h_i; k)$, with $k = 32$ by default. The candidate pool is the deduplicated union $\mathcal{C} = \bigcup_{i=1}^{m} \mathcal{R}(h_i; k)$; because no anchor depends on another anchor's results, all m vector searches are independent and may be issued in parallel. Fusion assigns each candidate the max-fusion score

$$\mathrm{score}(p) = \max_{1 \le i \le m} s(h_i, p), \tag{4}$$

so a passage enters the top of the list if it matches *any* anchor strongly. This disjunctive objective matches the structure of multi-hop evidence and is robust to a weak anchor, which yields low scores that cannot drag down passages matched by a better anchor, whereas sum or mean fusion lets one outlier depress the aggregate. It remains vulnerable when a bad anchor produces confident false positives that displace true evidence at truncation, an exposure that Appendix C measures directly by corrupting anchor entities. Section 4.4 quantifies the comparison.

Stage 3: anchor-aware cross-encoder reranking. An optional cross-encoder R_ψ reranks $\mathcal{C}$. Standard reranking scores candidates against the question alone, $r_\psi(q,p) = R_\psi(q \oplus p)$, which works only when all relevant passages are semantically proximate to q. Later-hop passages are not: for *"What is the nationality of the director of film X?"*, the director passage may score low because it never mentions film X, while scoring highly against the second-hop anchor. We therefore score each candidate against its most compatible query,

$$r_{\text{anchor}}(p) = \max\Big(r_\psi(q,p),\ \max_{1\le i\le m} r_\psi(h_i,p)\Big), \tag{5}$$

and rank by r_{anchor} in descending order. This is CHASE's default. Because anchor-max can over-promote candidates that match one anchor but have marginal global relevance, we optionally blend it with the dense ranking by reciprocal rank fusion (Cormack et al., 2009),

$$s_{\text{RRF}}(p) = \frac{1}{k_0 + \text{rank}_{\text{dense}}(p)} + \frac{1}{k_0 + \text{rank}_{\text{anchor}}(p)}, \tag{6}$$

with smoothing constant $k_0 = 60$, which yields a more conservative ranking.

4 Experiments

4.1 Setup

Datasets and metric. We evaluate on MuSiQue (Trivedi et al., 2022), which composes 2–4 hop questions and explicitly avoids reasoning shortcuts; HotpotQA (Yang et al., 2018), with bridge and comparison questions over Wikipedia; and 2WikiMultiHopQA (Ho et al., 2020), built from Wikidata with compositional, comparison, bridge-comparison and inference types. Following Gutiérrez et al. (2024) we use their curated 1,000-question evaluation subsets. We report Recall@k (Eq. 1), following Trivedi et al. (2023) and Gutiérrez et al. (2024). Dataset statistics are in Appendix H.

CHASE configuration. The generator is Qwen3.8-Max at temperature 0.1, prompted with five few-shot examples per dataset; the dense encoder is BGE-M3 (Chen et al., 2025) with $k = 32$ candidates per anchor; the cross-encoder is bge-reranker-v2-m3 applied to 80 fused candidates. Full settings are in Appendix I.

Comparison protocol. We report two kinds of comparison and label them separately throughout. A *matched* comparison varies only the query representation between question, sub-query and HSF, holding the encoder, corpus, index, candidate budget and reranker fixed. These rows support our claims about what HSF generation contributes. A *cross-system* comparison places CHASE beside published numbers for BM25 (Robertson & Walker, 1994), Contriever, GTR, ColBERTv2, RAPTOR, Proposition, HippoRAG, GFM-RAG, Adaptive-RAG, FLARE and IRCoT (Izacard et al., 2022; Ni et al., 2022; Santhanam et al., 2022; Sarthi et al., 2024; Chen et al., 2024; Gutiérrez et al., 2024; Luo et al., 2025; Jeong et al., 2024; Jiang et al., 2023; Trivedi et al., 2023). Those systems use different encoders and retrieval budgets, so that block is descriptive context: we do not attribute its margins to HSF generation and make no state-of-the-art claim from it.

4.2 Retrieval Performance

Matched comparison. The upper block of Table 1 isolates the contribution of the query representation. With the encoder and index held fixed and no reranking, HSF anchors improve Recall@2 over question-only retrieval by 13.92 points on MuSiQue (58.04 vs. 44.12), 11.09 on 2Wiki (74.97 vs. 63.88) and 7.00 on HotpotQA (78.30 vs. 71.30), with Recall@5 gains of 17.41, 17.87 and 4.00 points. Adding anchor-aware reranking to the same pipeline improves Recall@2 by a further 9.47, 12.20 and 3.95 points over question-only reranking of the same candidate pool. The gain is largest on MuSiQue and 2Wiki and smallest on HotpotQA, consistent with HotpotQA's shorter chains (2.00 average hops against 2.65 for MuSiQue) leaving less representation gap to close.

Cross-system context. In the lower block, CHASE's reranked configurations reach 60.31 Recall@2 on MuSiQue, 78.63 on 2Wiki and 87.00 on HotpotQA, above the strongest iterative baseline IRCoT+HippoRAG (45.3 / 75.8 / 67.0) and the graph foundation model GFM-RAG (49.1 / — /

Table 1: Retrieval performance on multi-hop QA benchmarks. The upper block is a **matched** comparison: encoder (BGE-M3), corpus, index and candidate budget are identical across rows and only the query representation or reranking strategy changes. The lower block reports published numbers for systems using different encoders and is descriptive context, not a controlled comparison. *: exceeds the protocol's theoretical upper bound of 88.25 for Recall@2 on 2Wiki; excluded (see text). Six further cross-system baselines are in Appendix A.

Method	MuSiQue		2Wiki		HotpotQA	
	R@2	R@5	R@2	R@5	R@2	R@5
Matched first stage: BGE-M3 encoder, no reranking						
Question only	44.12	56.92	63.88	71.28	71.30	84.95
CHASE (HSF, max fusion)	**58.04**	**74.33**	**74.97**	**89.15**	**78.30**	**88.95**
Matched reranked: same candidate pool, bge-reranker-v2-m3						
Question-only reranking	50.25	63.65	66.43	74.78	83.05	92.20
CHASE (anchor-aware, max)	59.72	75.98	**78.63**	90.78	**87.00**	**94.45**
CHASE (anchor-aware, RRF)	**60.31**	**77.38**	78.25	**91.15**	83.80	94.00
Cross-system context (different encoders; not controlled)						
BM25	32.3	41.2	51.8	61.9	55.4	72.2
ColBERTv2	37.9	49.2	59.2	68.2	64.7	79.3
HippoRAG (ColBERTv2)	40.9	51.9	70.7	89.1	60.5	77.7
GFM-RAG	49.1	58.2	90.8*	95.6*	78.3	87.1
FLARE	44.3	55.1	67.1	73.1	73.1	81.3
IRCoT + HippoRAG (ColBERTv2)	45.3	57.6	75.8	93.9	67.0	83.0

78.3). Because these systems use different encoders, the margins conflate query representation with retrieval stack; the matched block is the evidence for CHASE's contribution.

We exclude GFM-RAG's reported 90.8 Recall@2 on 2Wiki because it exceeds the 88.25% theoretical maximum of this evaluation protocol, whose 235 four-hop questions cap per-question recall at 0.5; we use IRCoT+HippoRAG (75.8) as the 2Wiki reference instead.[1]

4.3 End-to-End Question Answering

Table 2 reports Exact Match and F1 when the retrieved passages are handed to the same answer generator (GPT-4o-mini) across methods. CHASE reaches 36.1 / 47.99 on MuSiQue and 59.76 / 72.96 on HotpotQA, ahead of every baseline listed, including GFM-RAG (30.2 / 40.4 and 51.6 / 66.9). The MuSiQue margin is the largest, consistent with that benchmark's removal of reasoning shortcuts: covering the whole chain in one round gives the answer generator all required evidence at once. On 2Wiki, however, CHASE reaches 59.32 / 68.66 and *does not* match GFM-RAG (69.8 / 77.7), though it leads the iterative baselines. 2Wiki is built from Wikidata, so its questions are the closest in this suite to explicit relational traversal, the structure a persistent entity graph encodes and a query-time anchor set does not. We read this as a boundary of the graph-free approach rather than noise (Section 5).

4.4 Ablations: Anchor Form, Fusion Rule, and Reranking

Table 3 varies one design choice at a time under the matched protocol.

Anchor form. Question-only retrieval is the lower bound everywhere. Declarative HSFs give the best MuSiQue result, 2.22 points above interrogative sub-queries (58.04 vs. 55.82), and the two are within 0.30 points on HotpotQA; on 2Wiki, however, *sub-queries beat HSFs* by 2.25 points (77.22

[1]Bound: $(765 \times 1.0 + 235 \times 0.5)/1000$, from counting supporting facts per question.

Table 2: End-to-end QA performance (EM / F1). All rows use the same answer generator over each method's retrieved passages.

Retriever	MuSiQue		2Wiki		HotpotQA	
	EM	F1	EM	F1	EM	F1
No retrieval	12.5	24.1	31.0	39.0	30.4	42.8
ColBERTv2	15.5	26.4	33.4	43.3	43.4	57.7
GraphRAG (MS)	13.4	29.5	28.3	46.9	35.3	54.6
LightRAG	18.1	27.5	45.1	49.5	36.8	48.3
HippoRAG (ColBERTv2)	19.2	29.8	46.6	59.5	41.8	55.0
GFM-RAG	30.2	40.4	**69.8**	**77.7**	51.6	66.9
Adaptive-RAG	13.8	25.6	48.9	62.8	45.5	59.6
FLARE	16.2	28.4	46.7	65.4	48.7	60.6
IRCoT (ColBERTv2)	19.1	30.5	35.4	45.1	45.5	58.4
IRCoT + HippoRAG (ColBERTv2)	21.9	33.3	47.7	62.7	45.7	59.2
CHASE	**36.1**	**47.99**	59.32	68.66	**59.76**	**72.96**

Table 3: Ablations under the matched protocol (BGE-M3 encoder, identical corpus, index and candidate budget). Each block varies one factor; the HSF / max / no-reranking configuration is shared across blocks and repeated for reference. Best in each block is bold. Recall@10 is in Appendix E.

Configuration	MuSiQue		2Wiki		HotpotQA	
	R@2	R@5	R@2	R@5	R@2	R@5
(a) Anchor form (max fusion, no reranking)						
Question only	44.12	56.92	63.88	71.28	71.30	84.95
Sub-query	55.82	73.94	**77.22**	**90.53**	78.00	**89.20**
HSF	**58.04**	**74.33**	74.97	89.15	**78.30**	88.95
(b) Fusion rule (HSF anchors, no reranking)						
Max	**58.04**	**74.33**	**74.97**	**89.15**	**78.30**	**88.95**
Mean	47.99	64.56	73.25	88.25	73.65	87.40
Sum	25.48	42.83	54.98	71.92	65.05	75.65
RRF	24.07	42.38	47.20	67.40	52.05	68.00
CombMNZ	22.73	40.88	51.90	69.90	64.35	74.60
(c) Reranking strategy (HSF anchors, max fusion)						
No reranking	58.04	74.33	74.97	89.15	78.30	88.95
Question-only	50.25	63.65	66.43	74.78	83.05	92.20
Anchor-aware + max	59.72	75.98	**78.63**	90.78	**87.00**	**94.45**
Anchor-aware + RRF	**60.31**	**77.38**	78.25	**91.15**	83.80	94.00

vs. 74.97). The pattern tracks how much of the chain the question already exposes: MuSiQue hides bridge entities, so a declarative form that names them helps, whereas 2Wiki's Wikidata-derived questions often state both endpoints, leaving less for a hypothesized fact to add and more risk that a wrong entity guess narrows the query. Anchor form is benchmark-dependent, not a uniform win.

Fusion rule. Max fusion dominates every alternative on every benchmark: mean fusion costs 10.05 Recall@2 points on MuSiQue, and sum, RRF and CombMNZ collapse to 22–26 there and 47–55 on 2Wiki. The size of the gap supports the disjunctive reading of Eq. 4: when evidence needs are heterogeneous across hops, any rule that averages across anchors penalizes a passage for failing to match anchors it was never meant to match.

Reranking. Anchor-aware reranking is best everywhere, and question-only reranking is actively harmful on the two harder benchmarks, costing 7.79 points on MuSiQue and 8.54 on 2Wiki against no reranking while helping HotpotQA by 4.75: a cross-encoder that only ever sees the global question demotes the later-hop passages that the first stage surfaced. Between the anchor-aware variants neither dominates: RRF leads on MuSiQue, plain anchor-max on HotpotQA by 3.20 points.

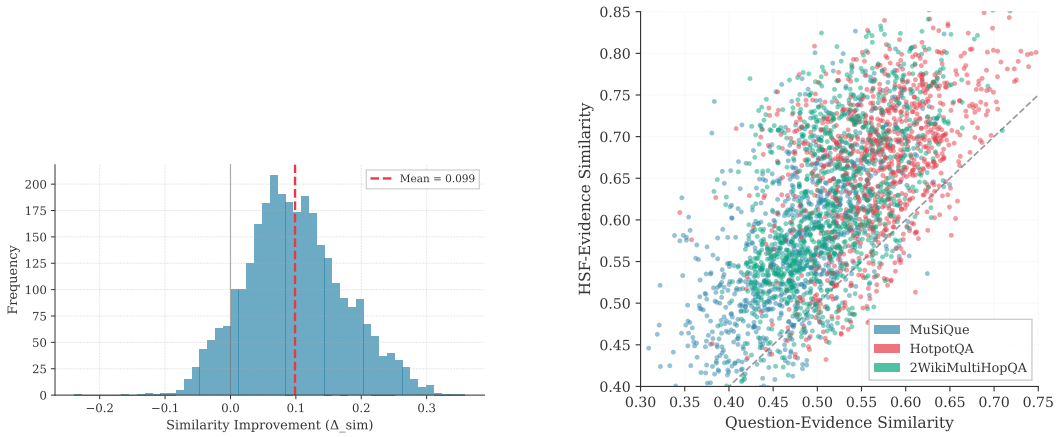


Figure 2: **Left:** distribution of the representation gap Δ_{sim} over 3,000 question–passage pairs (mean 0.099, median 0.094; 90.9% positive). **Right:** question-only similarity against best-HSF similarity. Points above the diagonal are cases where an anchor aligns better than the question; the margin is largest where question-passage similarity is initially low.

4.5 H1: Does the Representation Gap Close?

Recall gains show that HSFs help, but not *why*. We test H1 by measuring Δ_{sim} (Eq. 2) in BGE-M3 space over 3,000 question–passage pairs. The gap is positive for 90.9% of examples, with mean 0.099 and median 0.094; Figure 2 shows the distribution and the per-example scatter, where points cluster above the diagonal most strongly in the low-similarity region, so anchors help most where question-only retrieval is weakest, and the 9.1% of negative cases concentrate where the question's phrasing already resembles the passage's. By question type the gap is largest for *inference* (0.157, N=108) and *compositional* (0.139, N=413) and smallest for *comparison* (0.067, N=433), whose questions already name both targets, with *bridge* in between (0.100, N=811). That ordering matches the recall results (Appendix D).

4.6 H2: Generator Capability Is a Precondition

CHASE's anchors are only as good as the generator that writes them. We re-ran the pipeline with six open generators from 0.8B to 35B, holding the encoder, corpus, index and prompt fixed. The result (Table 4) is largely negative and cutoff-dependent: only the 27B and 35B models clear the matched question-only baseline at all, on MuSiQue at every k but on 2Wiki only for $k \geq 5$, while HotpotQA stays 1.1–1.5 points short throughout. Our reported results therefore rest on a frontier-scale generator. That the 2Wiki deficit is confined to $k = 2$ places the failure at the top of the ranking rather than in coverage, the exposure max fusion carries.

This makes H2's threshold τ a function of the cutoff: the crossing lies between 9B and 27B on MuSiQue and on 2Wiki for $k \geq 5$, but above 35B on 2Wiki at $k = 2$ and on HotpotQA throughout. Scale alone does not predict it—on MuSiQue gains are steep from 2B to 4B, flat to 9B and steep again to 27B, and the 35B mixture-of-experts model (3B active) trails the dense 27B model—so the operative capacity is active rather than nominal.

4.7 Cost

CHASE uses one generation call per question, but it is not cheap: 15,901 tokens per question on MuSiQue, 14,811 on HotpotQA and 9,207 on 2Wiki, dominated by the few-shot prompt, which is identical across questions within a dataset and thus a natural target for prefix caching we do not exploit. The tradeoff is one large call instead of several sequential ones against iterative methods, and per-query token cost instead of one-time index construction against graph-based ones: favorable on corpora that change often, unfavorable on static corpora with high query volume (Appendix G).

Table 4: Generator scale against the matched question-only baseline (italicized row). Cells at or above the baseline are bold. The two largest open generators clear the baseline on MuSiQue at both cutoffs and on 2Wiki at $k = 5$ (the 27B model only by 0.01), but no open generator reaches it on 2Wiki at $k = 2$ or on HotpotQA at either cutoff. Recall@10, in Appendix F, follows the $k = 5$ pattern.

Generator	Params (B)	MuSiQue		2Wiki		HotpotQA	
		R@2	R@5	R@2	R@5	R@2	R@5
Question only (no generator)	—	*44.12*	*56.92*	*63.88*	*71.28*	*71.30*	*84.95*
Small	0.8	32.16	43.17	43.88	52.27	53.01	68.04
Small	2.0	33.79	46.01	50.48	60.60	58.62	73.75
Mid	4.0	41.94	55.51	56.17	65.38	63.91	78.18
Mid	9.0	42.59	55.97	54.54	66.22	64.26	78.51
Large	27.0	**48.95**	**63.94**	60.90	**71.29**	70.00	83.87
Large (MoE, 3B active)	35.0	**47.28**	**63.26**	62.70	**72.88**	69.22	82.90
Frontier (Qwen3.8-Max)	—	**58.04**	**74.33**	**74.97**	**89.15**	**78.30**	**88.95**

5 Conclusion and Limitations

We asked whether the reasoning structure a retrieval graph supplies offline can instead be elicited from the question at query time. Under a matched protocol that holds the encoder, corpus and index fixed, it largely can. Declarative hop-local anchors raise first-stage Recall@2 by 13.9, 11.1 and 7.0 points over question-only retrieval, anchor-aware reranking adds 9.5, 12.2 and 4.0 points more, and end-to-end QA improves on MuSiQue and HotpotQA. The embedding-space measurement of Section 4.5 shows the mechanism at work: the anchor–passage gap is positive for 90.9% of examples and widens as question-only similarity falls. CHASE obtains this from a single generation call per question, with no parameter updates and no change to the corpus index, so the searches for all anchors can be issued at once.

Two results mark the boundary of the approach. The first is a capability requirement: at Recall@2 no open generator we tested at 35B parameters or below reached the question-only baseline on 2Wiki or HotpotQA, and the crossing moves with the cutoff (Section 4.6), so distilling anchor generation into a model that clears it is the most direct route to a deployable version. The second is a dependence on question type: on 2Wiki, whose Wikidata-derived questions name both endpoints, interrogative sub-queries retrieve better than declarative HSFs and GFM-RAG keeps an end-to-end lead, which suggests choosing the anchor form per question rather than fixing it in advance.

The design also accepts two costs by construction. Committing to every anchor before any retrieval result is what makes the searches independent, and it gives up the chance to revise an anchor once evidence arrives; a repair loop that regenerates only low-confidence anchors would recover part of that at limited serial cost. Token use per question is likewise the price of planning ahead, though the few-shot prefix is fixed within a dataset and invites prefix caching. A third dependence is empirical rather than structural: Appendix C shows that corrupting the entities of every HSF while leaving its relational form intact removes CHASE's entire advantage over question-only retrieval, so the gain rests on instantiating entities correctly and not on declarative form alone, which makes anchor factuality itself a target for improvement. Our evaluation covers three English Wikipedia-derived benchmarks with one encoder family; all of these are natural targets for follow-up work.

AI use statement

Large language models enter this work in two distinct roles, which we separate here. As an object of study, an LLM is a component of the proposed method: the generator of Section 3.2 produces the hypothetical supporting facts, and the models in Table 4 were run as experimental subjects. As a writing aid, the authors used an LLM assistant to edit and improve the readability of the manuscript. All research questions, method design, experimental protocol, analysis and interpretation are the

authors' own, and the authors take full responsibility for the content of this paper, including any errors.

Reproducibility statement

Section 4.1 specifies the generator, encoder, reranker, candidate budgets and evaluation subsets; Appendix I gives the full hyperparameters, Appendix H the dataset statistics, and Appendix J the verbatim prompts. All benchmarks are public and we use the curated 1,000-question subsets released by Gutiérrez et al. (2024). The matched protocol of Section 4.1 fixes the encoder, corpus, index and candidate budget across every condition we compare, so the retrieval rows can be reproduced without access to our generation outputs. Anonymized source code is provided as supplementary material.

References

Paul J. L. Ammann, Jonas Golde, and Alan Akbik. Question Decomposition for Retrieval-Augmented Generation. In Jin Zhao, Mingyang Wang, and Zhu Liu (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 497–507, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-254-1. doi: 10.18653/v1/2025.acl-srw.32.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, December 2025.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense X Retrieval: What Retrieval Granularity Should We Use? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15159–15177, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.845.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 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, pp. 758–759, New York, NY, USA, July 2009. Association for Computing Machinery. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572114.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From Local to Global: A Graph RAG Approach to Query-Focused Summarization, February 2025.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise Zero-Shot Dense Retrieval without Relevance Labels. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.99.

Bernal J. Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569, December 2024. doi: 10.52202/079017-1902.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised Dense Information Retrieval with Contrastive Learning, August 2022.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active Retrieval Augmented Generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550.

Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. Corpus-Steered Query Expansion with Large Language Models. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 393–401, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-short.34.

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. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020.

Hao Liu, Zhengren Wang, Xi Chen, Zhiyu Li, Feiyu Xiong, Qinhan Yu, and Wentao Zhang. HopRAG: Multi-Hop Reasoning for Logic-Aware Retrieval-Augmented Generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1897–1913, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.97.

Linhao Luo, Zicheng Zhao, Reza Haffari, Dinh Phung, Chen Gong, and Shirui Pan. GFM-RAG: Graph Foundation Model for Retrieval Augmented Generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 36371–36405. Curran Associates, Inc., 2025.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query Rewriting in Retrieval-Augmented Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5303–5315, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.322.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1864–1874, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.146.

Rodrigo Nogueira and Kyunghyun Cho. Passage Re-ranking with BERT, April 2020.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and Narrowing the Compositionality Gap in Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. Answering Complex Open-domain Questions Through Iterative Query Generation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2590–2602, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1261.

S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In Bruce W. Croft and C. J. van Rijsbergen (eds.), *Sigir '94*, pp. 232–241, London, 1994. Springer London. ISBN 978-1-4471-2099-5.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.272.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval, January 2024.

Zhili Shen, Chenxin Diao, Pavlos Vougiouklis, Pascual Merita, Shriram Piramanayagam, Enting Chen, Damien Graux, Andre Melo, Ruofei Lai, Zeren Jiang, Zhongyang Li, Ye Qi, Yang Ren, Dandan Tu, and Jeff Z. Pan. GeAR: Graph-enhanced Agent for Retrieval-augmented Generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 12049–12072, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.624.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. Generate-then-Ground in Retrieval-Augmented Generation for Multi-hop Question Answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7339–7353, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.397.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop Questions via Single-hop Question Composition. In Brian Roark and Ani Nenkova (eds.), *Transactions of the Association for Computational Linguistics*, volume 10, pp. 539–554, Cambridge, MA, 2022. MIT Press. doi: 10.1162/tacl_a_00475.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.557.

Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query Expansion with Large Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9414–9423, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.585.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models, October 2022.

Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. In *International Conference on Learning Representations*, October 2020.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models, March 2023.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Generate rather than Retrieve: Large Language Models are Strong Context Generators. In *The Eleventh International Conference on Learning Representations*, September 2022.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Liu Yong, and Shen Huang. End-to-End Beam Retrieval for Multi-Hop Question Answering. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1718–1731, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.96.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. Knowledge Graph-Guided Retrieval Augmented Generation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 8912–8924, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.449.

Ziyuan Zhuang, Zhiyang Zhang, Sitao Cheng, Fangkai Yang, Jia Liu, Shujian Huang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Qi Zhang. EfficientRAG: Efficient Retriever for Multi-Hop Question Answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3392–3411, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.199.

A Additional Cross-System Baselines

Table 5 lists the cross-system baselines omitted from Table 1 for space. As in the main text, these systems use different encoders and retrieval budgets from ours, so the block is descriptive context rather than a controlled comparison.

Table 5: Cross-system baselines omitted from Table 1 (different encoders; not a controlled comparison).

Method	**MuSiQue**		**2Wiki**		**HotpotQA**	
	R@2	R@5	R@2	R@5	R@2	R@5
Contriever	34.8	46.6	46.6	57.5	57.2	75.5
GTR	37.4	49.1	60.2	67.9	59.4	73.3
RAPTOR (ColBERTv2)	36.9	46.5	57.3	64.7	63.1	75.6
Proposition (ColBERTv2)	37.8	50.1	55.9	64.9	63.9	78.1
Adaptive-RAG	35.1	44.7	44.7	61.4	61.0	76.4
IRCoT + ColBERTv2	41.7	53.7	64.1	74.4	67.9	82.0

B Formal Statement of Hypotheses

H1 (representation gap). For a supporting passage set $\mathcal{G}_q$ and generated anchor set H,

$$\max_{h_i \in H} \mathbb{E}_{p \in \mathcal{G}_q} \left[\text{sim} \left(E_\phi(h_i), E_\phi(p) \right) \right] > \mathbb{E}_{p \in \mathcal{G}_q} \left[\text{sim} \left(E_\phi(q), E_\phi(p) \right) \right], \tag{7}$$

equivalently $\Delta_{\text{sim}}(q) > 0$ under Eq. 2. We predict $\mathbb{E}_q[\Delta_{\text{sim}}(q)] > 0$ and $P(\Delta_{\text{sim}}(q) > 0) > 0.5$ across $\mathcal{Q}$. Tested in Section 4.5.

H2 (capability threshold). There exists a generator capability threshold τ, operationalized through parameter count or benchmark performance, such that

$$\text{Capability}(G_\theta) > \tau \implies \text{Recall}_{\text{CHASE}}(G_\theta) > \text{Recall}_{\text{question-only}}, \tag{8}$$

where both sides use the same encoder, corpus and index. Tested in Section 4.6: at $k = 2$ the estimated τ exceeds 35B parameters on 2Wiki and HotpotQA, but on 2Wiki it falls below 27B for $k \geq 5$.

H3 (structure-conditioned error tolerance; exploratory). Each HSF has two properties that can vary independently: factuality, whether its instantiated entities are correct, and relational compatibility, whether it expresses the relation the hop requires. An HSF stating *"John Phan was born in Hanoi, Vietnam"* is factually wrong but relationally compatible with passages about birthplaces in Vietnam. H3 predicted that relational compatibility is the primary determinant of retrieval success and factual accuracy a secondary enhancement, so that with $\mathcal{Q}_{\text{low}} = \{q : \text{EntityAcc}(H_q) \leq 40\%\}$,

$$\text{Corr}\big(\text{EntityAcc}(H), \text{Recall@}k\big) > 0 \quad \text{and} \quad \mathbb{E}_{q \in \mathcal{Q}_{\text{low}}}[\text{Recall@}k(q)] > \text{Recall@}k_{\text{question-only}}. \tag{9}$$

Appendix C reports the test. The correlation came out positive as predicted, but the second prediction fails on two of three benchmarks, and anchors whose relational form is intact while every entity is wrong retrieve worse than the question alone, so the premise that factual accuracy is only a secondary enhancement does not hold.

C Entity Perturbation Study

Protocol. To test H3 we corrupted the entities of generated HSFs at five perturbation levels. At perturbation level $X\%$, a randomly sampled $X\%$ of HSF records have their person names, locations, dates and organizations replaced by plausible but incorrect alternatives, using an LLM prompted to preserve the surrounding relational structure and sentence form (Appendix J); the remaining $(100 - X)\%$ of records keep their original HSFs. Sub-queries and sentence structure are retained, so that relational compatibility is nominally held fixed while factual content degrades. Entity accuracy in the sense of Eq. 9 is therefore $100 - X$: level 0% is unperturbed CHASE and level 100% corrupts every record. Retrieval is otherwise the matched first-stage configuration of Table 1 (BGE-M3 encoder, max fusion, no reranking), so the 0% and question-only rows of Table 6 are reproduced from that table as reference points.

Result. Table 6 shows monotone degradation: each additional 20% of corrupted records costs Recall@2 on all three benchmarks, and corrupting every record costs 15.46 points on MuSiQue (58.04 to 42.58), 17.45 on HotpotQA (78.30 to 60.85) and 18.64 on 2Wiki (74.97 to 56.33), between 22 and 27% of the unperturbed value. Read as a dose–response over the six levels, entity accuracy and Recall@2 are positively related on every benchmark ($r > 0.998$, Spearman $\rho = 1$), the direction Eq. 9 predicts.

H3 holds in its first clause and fails in its second. The second prediction asks that questions with entity accuracy at or below 40%, here the levels corrupting at least 60% of records, still beat question-only retrieval. Averaged over those three levels, Recall@2 is 45.82 on MuSiQue against a question-only 44.12, but 64.25 on HotpotQA against 71.30 and 60.19 on 2Wiki against 63.88: the prediction holds on MuSiQue alone, and by 1.70 points. At full corruption every benchmark falls below question-only retrieval, by 1.54 points on MuSiQue, 10.45 on HotpotQA and 7.55 on 2Wiki. Relational compatibility alone therefore does not carry retrieval. Since CHASE's unperturbed margins over question-only are 13.92, 7.00 and 11.09 points on the same three benchmarks and none of

that margin survives entity corruption, correct entity instantiation accounts for essentially the whole of the measured gain, and H3's ordering of the two properties is backwards. This also bounds the exposure noted in Section 3.2: max fusion protects against an anchor that scores low everywhere, not against one that confidently retrieves the wrong entity's passages.

An error budget rather than error tolerance. The useful reading is not that factual errors are harmless but that the method has a budget for them. Interpolating between levels, CHASE stays at or above the question-only baseline up to roughly 90% corruption on MuSiQue, 59% on 2Wiki and 41% on HotpotQA. That ordering follows the size of the unperturbed gain, so the benchmarks with the largest representation gap also absorb the most generator error, which is what one would expect if HSFs help by closing a gap whose width sets the tolerance.

What the design cannot show. Because a record is either fully corrupted or untouched, level-X recall is close to the mixture $(1 - X)R_{0\%} + XR_{100\%}$: that prediction matches every intermediate level to within 0.61 points. The intermediate levels therefore add little beyond the two endpoints, and the near-linearity is a property of the sampling design rather than evidence against a threshold in how much corruption a single HSF can absorb. For the same reason Eq. 9's correlation, stated per question, can only be evaluated as an aggregate dose–response across conditions. Two further limits bear on interpretation. The perturbation both removes the correct entity and substitutes a competing real one, so it does not separate the cost of losing the right anchor from the cost of being pulled toward a wrong one; replacing entities with type placeholders (*"a city in Vietnam"*) would degrade specificity without introducing a competitor, and is a control we did not run. And adversarial substitution is not natural hallucination: a generator's entity errors are correlated with question difficulty and are plausibly closer to the truth than a uniformly wrong replacement, so these numbers bound the cost of wrong entities rather than estimate the cost of the errors CHASE actually makes.

Table 6: Recall@2 under entity perturbation. Level $X\%$ corrupts the entities of $X\%$ of HSF records, so entity accuracy is $100 - X$. Recall falls monotonically with corruption and drops below question-only retrieval once most records are corrupted. The 0% and question-only rows are the matched first-stage rows of Table 1.

Records perturbed	**MuSiQue**	**HotpotQA**	**2WikiMQA**
0% (unperturbed CHASE)	58.04	78.30	74.97
20%	54.87	74.20	71.62
40%	52.17	71.45	67.55
60%	49.33	67.80	63.73
80%	45.55	64.10	60.50
100% (all records)	42.58	60.85	56.33
Question-only reference	44.12	71.30	63.88

D Representation Gap by Question Type

Table 7 gives the per-type breakdown summarized in Section 4.5, and Figure 3 the corresponding distributions. Types are ranked by gap magnitude. Note the small sample sizes for the 4-hop subsets ($N = 27$ and $N = 31$ for 4hop2 and 4hop3), which we do not treat as evidence of a trend in chain length.

E Full Ablation Results

Table 8 repeats the ablations of Section 4.4 including Recall@10.

F Generator Scale: Full Results

Table 9 gives the Recall@10 column omitted from Table 4. It follows the Recall@5 pattern: the 27B and 35B generators clear the question-only baseline on MuSiQue (71.23 and 70.31 vs. 64.21) and

Table 7: Representation gap (Δ_{sim}) by question type. Positive values indicate the best anchor is more similar to the supporting passages than the original question.

Question type	N	Δ_{sim}	sim_q	sim_{hsf}
Inference	108	0.157	0.552	0.709
Compositional	413	0.139	0.528	0.667
2hop	518	0.111	0.500	0.611
Bridge	811	0.100	0.565	0.665
3hop1	243	0.082	0.452	0.534
Bridge-Comparison	235	0.077	0.461	0.538
4hop1	108	0.075	0.422	0.497
3hop2	73	0.071	0.497	0.568
Comparison	433	0.067	0.557	0.624
4hop2	27	0.047	0.449	0.496
4hop3	31	0.041	0.423	0.464

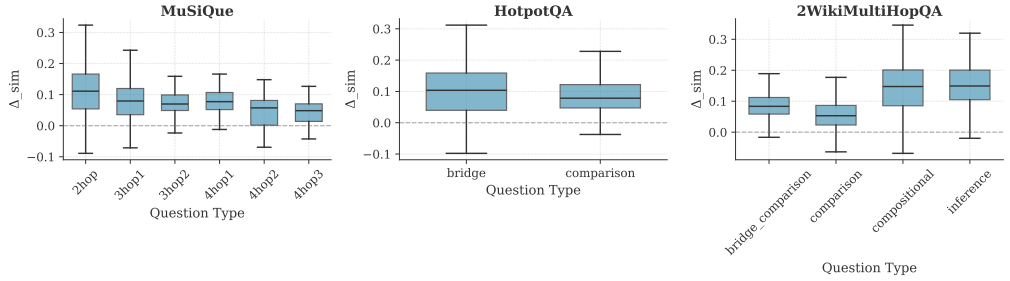


Figure 3: Distribution of similarity gaps (Δ_{sim}) across question types. Inference and compositional questions exhibit the largest median gaps; comparison questions show tighter distributions with smaller median improvements.

on 2Wiki (74.85 and 76.08 vs. 74.38), while every open generator remains below it on HotpotQA (best 89.30 vs. 90.75). The cutoff at which the 2Wiki crossing occurs is therefore between $k = 2$ and $k = 5$.

G Token Usage

H Dataset Details

HotpotQA (Yang et al., 2018) contains questions requiring reasoning over multiple Wikipedia paragraphs, with two main types: *bridge* (finding a shared entity between documents) and *comparison* (comparing properties across entities).

2WikiMultiHopQA (Ho et al., 2020) is constructed from Wikidata and includes four question types: *compositional*, *comparison*, *bridge-comparison* and *inference*.

MuSiQue (Trivedi et al., 2022) is synthesized by composing single-hop questions, featuring 2–4 reasoning hops, and explicitly avoids reasoning shortcuts.

Following Gutiérrez et al. (2024) we use their curated 1,000-sample development splits rather than the original official splits, enabling comparison with prior graph-based and iterative methods.

I Implementation Details

LLM generator. Qwen3.8-Max generates HSFs in all main experiments, producing structured JSON with one HSF per reasoning step, guided by five few-shot examples per dataset. Temperature is 0.1.

Table 8: Full ablation results including Recall@10, under the matched protocol.

Configuration	MuSiQue			HotpotQA			2WikiMQA		
	R@2	R@5	R@10	R@2	R@5	R@10	R@2	R@5	R@10
Anchor form (max fusion, no reranking)									
Question only	44.12	56.92	64.21	71.30	84.95	90.75	63.88	71.28	74.38
Sub-query	55.82	73.94	80.91	78.00	89.20	93.00	77.22	90.53	91.35
HSF	58.04	74.33	80.42	78.30	88.95	92.80	74.97	89.15	91.22
Fusion rule (HSF anchors, no reranking)									
Max	58.04	74.33	80.42	78.30	88.95	92.80	74.97	89.15	91.22
Mean	47.99	64.56	74.19	73.65	87.40	92.40	73.25	88.25	90.75
Sum	25.48	42.83	58.06	65.05	75.65	81.15	54.98	71.92	81.08
RRF	24.07	42.38	57.29	52.05	68.00	78.75	47.20	67.40	79.20
CombMNZ	22.73	40.88	56.89	64.35	74.60	80.50	51.90	69.90	79.90
Reranking strategy (HSF anchors, max fusion)									
No reranking	58.04	74.33	80.43	78.30	88.95	92.80	74.97	89.15	91.23
Question-only	50.25	63.65	72.04	83.05	92.20	95.75	66.43	74.78	80.23
Anchor-aware + max	59.72	75.98	81.21	87.00	94.45	96.95	78.63	90.78	92.03
Anchor-aware + RRF	60.31	77.38	83.45	83.80	94.00	96.60	78.25	91.15	92.93

Table 9: Generator scale analysis, all metrics. The question-only row uses the same encoder, corpus and index.

Generator	B	MuSiQue			HotpotQA			2WikiMQA		
		R@2	R@5	R@10	R@2	R@5	R@10	R@2	R@5	R@10
Question only	—	44.12	56.92	64.21	71.30	84.95	90.75	63.88	71.28	74.38
Small	0.8	32.16	43.17	51.03	53.01	68.04	75.55	43.88	52.27	56.35
Small	2.0	33.79	46.01	53.16	58.62	73.75	79.56	50.48	60.60	65.25
Mid	4.0	41.94	55.51	63.07	63.91	78.18	84.73	56.17	65.38	69.62
Mid	9.0	42.59	55.97	63.34	64.26	78.51	85.59	54.54	66.22	71.04
Large	27.0	48.95	63.94	71.23	70.00	83.87	89.30	60.90	71.29	74.85
Large (MoE)	35.0	47.28	63.26	70.31	69.22	82.90	88.53	62.70	72.88	76.08
Frontier	—	58.04	74.33	80.42	78.30	88.95	92.80	74.97	89.15	91.22

Dense retriever. BGE-M3 (Chen et al., 2025) encodes queries and passages into a shared 1024-dimensional space. For each anchor we retrieve the top-32 candidates, fused by max-score aggregation (Eq. 4).

Cross-encoder reranker. bge-reranker-v2-m3, applied with the anchor-aware (anchor-max) strategy of Eq. 5 by default. We fuse 80 candidates before reranking, then truncate to the final top-$K \in \{2, 5, 10\}$ for evaluation.

Table 10: Average token usage per question for HSF generation (Qwen3.8-Max).

Dataset	Prompt	Completion	Total
HotpotQA	14,047	764	14,811
2WikiMultiHopQA	8,604	602	9,207
MuSiQue	15,392	509	15,901

Table 11: Statistics of the three multi-hop question answering datasets.

Dataset	Questions	Corpus docs	Avg. hops
2WikiMultiHopQA	1,000	6,119	2.47
HotpotQA	1,000	9,811	2.00
MuSiQue	1,000	11,656	2.65

Question-type distributions: *2WikiMultiHopQA*: compositional (413), comparison (244), bridge_comparison (235), inference (108); *HotpotQA*: bridge (811), comparison (189); *MuSiQue*: 2-hop (518), 3-hop (316), 4-hop (166).

J Prompt Templates

J.1 HSF Generation System Prompt

```
You are an expert Wikipedia editor and multi-hop reasoning
specialist. For every atomic reasoning step needed to answer
a multi-hop question, generate a pair of retrieval anchors:

1. sub_query: an atomic, low-assumption information need using
   EXPLICIT entity names from the question.
2. hypothetical_supporting_fact (hsf): a concrete, factual
   sentence written like a REAL Wikipedia excerpt that would
   answer the sub-query.

CRITICAL RULES:
- Use ACTUAL entity names, dates, and locations | NEVER use
  placeholders like "Director A", "[date]", "[year]", or
  "[profession]".
- If you don't know a specific fact (e.g., a director's name),
  make a plausible guess based on the question context or
  historical patterns.
- Each hsf must be a complete, standalone Wikipedia-style
  sentence with concrete details.

Infer the number and order of steps from the question. Each
pair must describe one step, and both fields must be non-empty.
```

J.2 Entity Perturbation Prompt

For error tolerance analysis (Section C), we use a controlled perturbation prompt to corrupt entity mentions while preserving relational structure:

```
You are a controlled-fact-perturbation editor. You are given
the reasoning steps of a multi-hop question, each with a
sub_query and a hypothetical supporting fact (hsf). Rewrite
EVERY hsf so that its entities are factually WRONG, while the
sentence keeps reading like a real Wikipedia excerpt.

CRITICAL RULES:
- Corrupt ENTITIES ONLY: person names, place names,
  organisations, dates, years, and quantities. Keep the
  sentence's syntax, length, register, and level of detail.
- The rewritten hsf must remain a plain, confident, standalone
  declarative sentence. Never hedge it: no "reportedly",
  "possibly", "some sources claim", "it is said".
- Keep every step's sub_query EXACTLY as given, character for
  character.
- Corrupt CONSISTENTLY across steps: if an entity appears in
  more than one step, replace it with the same wrong entity
  everywhere. The chain must remain internally coherent.
```

```
Example transformation:
original:  "Diego Maradona was signed by FC Barcelona in 1982
            for a then-world record transfer fee from Boca
            Juniors."
rewritten: "Diego Maradona was signed by Real Madrid in 1979
            for a then-world record transfer fee from River
            Plate."
```

J.3 Question Answering Prompt

For end-to-end QA evaluation (Section 4.3), we use a reasoning-then-answer prompt format:

```
As an advanced reading comprehension assistant, analyze the
passages and question carefully. Start reasoning after
"Thought: " and finish with "Answer: " followed by a concise
answer without further explanation.

[Few-shot example for HotpotQA]
Wikipedia Title: Walls and Bridges
Walls and Bridges is the fifth studio album by English
musician John Lennon. It was issued by Apple Records in 1974
and recorded during his separation from Yoko Ono.

Wikipedia Title: Nobody Loves You (When You're Down and Out)
Nobody Loves You is a song written by John Lennon and released
on Walls and Bridges.

Question: Nobody Loves You was written by John Lennon and
released on what album that was issued by Apple Records and
recorded during his separation from Yoko Ono?

Thought: The album is Walls and Bridges.
Answer: Walls and Bridges.
```