# WHAT RETRIEVAL BUYS DEPENDS ON WHO READS: SPLITTING MULTI-HOP RAG ACCURACY INTO ASSEMBLY AND READING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multi-hop questions name their endpoints but not the passages that connect them. Retrieval-augmented systems reach those passages either by building an entity graph offline or by retrieving iteratively with a model in the loop, and the two are compared by end-to-end accuracy. Accuracy, however, cannot tell whether a system failed to assemble the evidence or its reader failed to use it. We separate the two steps with two controls: at a matched passage count only the retrieval varies, and with identical passages only the reader varies. Whether the evidence holds the whole chain is scored apart from the answer. Together the controls split each reader's accuracy gap to gold evidence into an assembly share and a reading share. Measuring the steps apart changes verdicts. On MuSiQue, end-to-end accuracy ranks a single read of a graph index at least as high as an iterative loop. Yet at the same passage count the loop completes far more chains, and its passages raise accuracy once read after the search. The split also shows that the bottleneck differs by reader. On MuSiQue's deep questions most of the 8B readers' gap is reading: the index's top-ranked wrong passages cost open readers up to 32B more than the frontier model, at least twice as much by point estimate. Training moves one share and not the other: a small model distilled call by call from the loop's search learns to search, with no detectable reduction in its reading share. Running alone, it outscores the published systems we compare on the same MuSiQue questions. Retrieval gains are thus relative to the reader, and we propose evaluating them at a matched passage count and per reader.

## 1 INTRODUCTION AND RELATED WORK

A multi-hop question names its endpoints but not its middle. To answer *who directed the film in which an actress made her debut*, a system must find her biography, learn the title of her first film, and only then retrieve the passage that names its director. Structured systems reach this unstated middle offline, extracting an entity graph that one retrieval can follow (Gutiérrez et al., 2024; 2025; Edge et al., 2024; Guo et al., 2025); iterative systems reach it online, letting a model read what came back and retrieve again (Trivedi et al., 2023; Press et al., 2023; Jin et al., 2025).

Which approach is better is argued with end-to-end accuracy (Zhuang et al., 2026; Guo et al., 2025), which mixes two failures with different remedies: never assembling the passages a chain needs, and failing to use the passages at hand. RAGChecker separates retriever from generator errors at a fixed top-$k$ (Ru et al., 2024), which cannot compare iterative systems that gather a different number of passages per question. Whole-chain recall is scored for retrievers (Xiong et al., 2021), but not across readers given the same evidence. For a given reader, it remains unclear whether accuracy stops at the evidence or at reading it.

We answer this with two controls, each varying one step while the other is fixed. HippoRAG 2, a deployed graph index, supplies single-pass evidence; a loop driven by Claude Opus 4.6 supplies iterative evidence (Figure 1). Both deliver the same number of passages per question, so better assembly should show as more complete chains and a gain for every reader. All readers, open models from 7B to 235B and three frontier models, see identical lists, so a gain that depends on



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.

reading should move with the reader. Both run on MuSiQue, 2WikiMultiHopQA, and HotpotQA (Trivedi et al., 2022; Ho et al., 2020; Yang et al., 2018), calling four-hop questions deep.

**Contributions.** HippoRAG 2's index, ranking, and corpus releases, IRCoT as a comparator (Trivedi et al., 2023), the three benchmarks, and the readers are existing work. We build the loop as our measuring instrument and teacher, not as a proposed method; Figure 4e shows where it and the searcher distilled from it stand among published systems. Our contributions are:

- **A measurement that splits accuracy into assembly and reading (§2).** At a matched passage count and with identical passages for every reader, a reader's gap to gold evidence splits into an assembly share and a reading share. It implies a standard: a claim that a retrieval method helps multi-hop question answering should show more complete chains at a matched passage count, with a gain that holds across readers. With the completeness gap, the split also yields a rule for what to pay for (§6).
- **Accuracy alone misranks retrieval (§3).** On MuSiQue, a single read of the graph index scores at least as well as the loop's own answers. Yet at the same passage count the loop completes far more deep chains (46.4% against 18.7%), and over all MuSiQue questions its passages raise every grid reader's accuracy once read after the search.
- **Where accuracy stops depends on the reader (§4).** On MuSiQue's deep questions, most of the 8B readers' gap is reading. The index's top-ranked wrong passages cost open readers up to 32B more accuracy than the frontier model, at least twice as much by point estimate.
- **Training moves one share, not the other (§5).** The smaller models that lose most to distractors are the ones a deployed system can afford, so we distill the loop into an 8B. Controls trained on identical inputs show that the 8B inherits the teacher's search decisions, with no detectable reduction in its reading share; after training it still loses more to wrong passages than the frontier model.
- **A call-by-call distillation framework and the searcher it yields (§5).** Each call of the frontier loop becomes one training example, using only the teacher's text outputs. Trained this way on MuSiQue alone, the 8B runs the loop at 47.4 exact match, above the published systems evaluated on the same questions and above a single read by the same-size Qwen3-8B. It transfers to 2WikiMultiHopQA without further training, and with a frontier reader over its searches it comes within a point of its teacher's own loop. Its training calls, code, and checkpoint are released with the supplementary material.

Retrieval gains are therefore relative to the reader: completing a chain is worth more to a reader that loses little to wrong passages than to one that loses much. A benchmark run with one reader on one corpus measures only one reader–corpus pair.

**Related work.** Iterative retrievers interleave reading with search (Trivedi et al., 2023; Press et al., 2023; Shao et al., 2023; Jiang et al., 2023), and graph indexes precompute the links a chain follows (Edge et al., 2024; Sarthi et al., 2024; Gutiérrez et al., 2025; Zhuang et al., 2026); both are compared by end-to-end accuracy, each at its own passage budget. Analyses of multi-hop benchmarks show that many questions can be answered without the full chain (Min et al., 2019; Chen & Durrett, 2019), which is why we score completeness apart from the answer. Studies of distracting context add irrelevant or random passages and read them with one or a few models (Shi et al., 2023; Cuconasu et al., 2024; Yoran et al., 2024); we pad with the index's own top-ranked passages and read them with ten. Distilled and reward-trained searchers (Kang et al., 2025; Jin et al., 2025; Zheng et al., 2025) report end-to-end gains, without controls that change only the training targets.

## 2 Measurement Protocol

The measurement changes one step at a time and scores the evidence apart from the answer. This section defines the conditions that vary retrieval, the readers that vary reading, and the quantities scored. §3–§5 compare them in one grid (Table 2), with conditions as columns and readers as rows.

**Corpora.** We use the MuSiQue, 2WikiMultiHopQA, and HotpotQA releases of HippoRAG 2, so that its index and ranking run as deployed, with 1,000 development questions each. Their four-hop questions, 166 on MuSiQue and 241 on 2WikiMultiHopQA, form each corpus's *deep stratum*; HotpotQA questions are all two-hop.

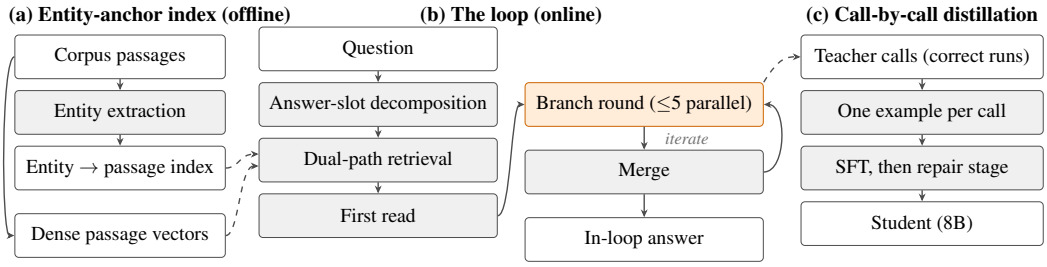



Figure 1: The loop and its distillation. (a) An entity-to-passage index without relations, beside dense vectors. (b) For each answer slot, up to five candidates retrieve in parallel and are kept or dropped. (c) Each teacher call becomes one training example (§5).

Table 1: Comparators (§3, §5), MuSiQue; ALL = all questions, 4-hop = deep stratum. Top: a 200-passage read of the ranking and IRCoT, a second iterative system, read by the frontier model. Bottom: loops driven by GPT-5.5 and by the distilled 8B, with the frontier model's own completeness in parentheses. Strict EM, with 95% intervals in brackets; the three reader differences for the GPT-5.5 driver each exclude zero.

| | |
|---|---|
| **A deeper single read and a second iterative system** | |
| Top-200 read − loop re-read (ALL) | 0.0 $[-2.1, 2.1]$ |
| Input tokens, top-200 read / loop with re-read (ALL) | 30.7K / 31.3K |
| IRCoT re-read / IRCoT in-loop answer (ALL) | 63.5 / 42.5 |
| **Drivers other than the frontier model** | |
| GPT-5.5 driver, completeness (ALL) | 77.2 (78.5) |
| GPT-5.5 driver's Loop passages − frontier model's read by Opus 4.6 / GPT-5.5 / Gemini, ALL | $-3.3$ / $-3.0$ / $-3.7$ |
| dist. 8B driver, completeness (ALL / 4-hop) | 72.3 (78.5) / 30.7 (46.4) |
| its re-read − its in-loop answer (frontier model; ALL) | $+15.9$ $[13.0, 18.9]$ |

**Assembly: the same number of passages.** To compare how two retrievals assemble evidence, we give them the same passage budget. The single-pass condition cuts HippoRAG 2's deployed ranking at $k$ passages (*Top-$k$*; $k$=5 is its official depth, $k$=20 a shared budget). HippoRAG 2 (HR2) ranks passages by combining dense similarity with personalized PageRank over its entity graph. The iterative condition lists the passages the loop collected, in collection order, each text once (*Loop passages*). The loop collects a different number $K_q$ for each question, so the ranking is also cut at that question's own $K_q$ (*Top-$K_q$*). The two lists then differ only in how their passages were found. The budget matches passages, not compute.

**The loop answers during its search.** The loop (Figure 1) splits the question into answer slots and fills them in turn: for each slot it proposes candidate answers, retrieves passages for each, and keeps or drops them after reading. The model that runs it is its *driver*; Claude Opus 4.6 drives it unless another driver is named. Its record of queries, passages, and notes is its *trajectory*, and the answer it gives from inside that record is its *in-loop answer*. To read the same evidence after the search, the *Loop passages* condition hands the collected passages to readers as a plain list; reading a search's passages this way is a *re-read*.

**Reading: the same passages.** Within a column every reader receives identical lists, so a difference down a column is a difference in reading. Top-$k$ and loop passages are what a deployed system delivers; three diagnostic conditions bound what reading can do. *Closed book* gives no passages and shows what a reader knows without evidence. *Gold* gives the annotated passages of the chain and shows reading when the evidence is complete. *Gold+distractors* pads the gold passages to twenty with the ranking's highest-ranked non-gold passages, in ranking order, adding the noise a deployed index returns. §4 varies this noise to locate what in it costs a reader.

**Readers.** The readers span the scales and families a deployment might choose. Claude Opus 4.6 is *the frontier model* and, in §5, the *teacher*. Qwen3-8B, the official post-trained release, is paired with the distilled 8B, or *student*, which is Qwen3-8B-Base fine-tuned on the teacher's calls, so the

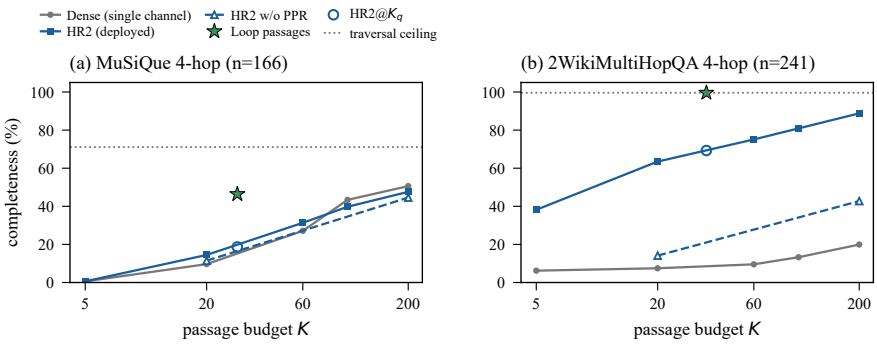



Figure 2: Completeness against passage budget $K$ on the deep strata. On MuSiQue, one pass of the ranking stays below half even at $K$=200; the loop's equal-budget comparison is the ranking cut at each question's own $K_q$ (HR2@$K_q$). The other single-pass curves are HR2 without personalized PageRank (w/o PPR, measured at $K$=20 and 200 only) and one dense pass over the loop's own index (Dense); dotted, the share of chains whose every hop is stated in some single passage.

two 8Bs share size and lineage; Qwen3-8B-Base also reads, as the student's starting point. GPT-5.5, Gemini 3.8 Flash, OLMo-3-7B-Instruct, and Qwen3-14B, 32B, and 235B-A22B complete the set. The three frontier models are the *frontier readers*, the open-weight ones the *open readers*.

**What is scored.** Evidence is scored apart from the answer, so that a condition is judged both on what it delivered and on what readers made of it. A question's evidence is *complete* if every gold passage appears in the list; *completeness* is the share of complete questions. Answers are scored by strict exact match (EM) under the official normalization. Two quantities split a reader's gap to gold evidence, gold minus top-20. Its *reading share* is its exact match on gold minus its exact match on gold+distractors: the chain is complete in both, so only the added passages differ. Its *assembly share* is its exact match on gold+distractors minus its exact match on the top-20 read: the count stays at twenty, but the chain is completed.

**Holding the answer format fixed.** Exact match rewards an answer's form as well as its content, so one prompt serves every instruction-following reader, condition, and corpus. Qwen3-8B-Base, which does not follow it, answers under its own minimal prompt and is compared only with itself.

**Statistics.** Reads of a fixed list decode greedily; the frontier model's loop and its closed-book reads sample at temperature 0.2. Differences are tested question by question, with paired bootstrap 95% intervals; contrasts that lose significance under Benjamini–Hochberg correction are marked.

## 3 What End-to-End Accuracy Hides

End-to-end accuracy is how graph indexes and iterative loops are compared (§1). This section applies the first control of §2, varying only the retrieval at a matched passage count, and asks whether that comparison survives once assembly is measured apart.

**End to end, the graph index looks at least as good as iteration.** Read by the same frontier model, HR2's top twenty passages score 61.4 over all MuSiQue questions and the loop's in-loop answer 57.8; on the deep stratum they score 41.0 and 40.4 (Table 2, Figure 4). Compared this way, iteration buys nothing on MuSiQue.

**Yet at the same passage count, iteration assembles far more of each chain.** Cut at each question's own $K_q$, the ranking completes 18.7% of MuSiQue's deep chains and the loop 46.4% (Table 2, COMP row; Figure 2); the gain recurs on 2WikiMultiHopQA's deep stratum and vanishes on HotpotQA, which one pass already completes. One dense pass over the loop's own index completes fewer deep chains than the ranking at twenty passages (Figure 2), so its gain does not come from its dense retriever. It saves passages rather than tokens, since a 200-passage read of the ranking catches up at a similar token cost (Table 1). On MuSiQue, the assembly gain is large where accuracy shows none.

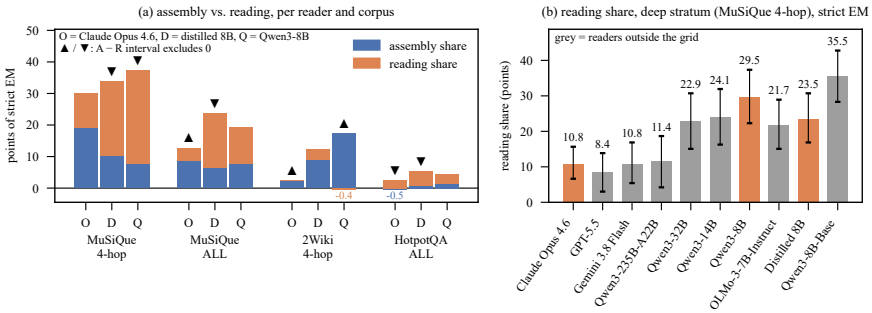



Figure 3: Where accuracy stops. (a) Each grid reader's gap to gold evidence split into its assembly share and reading share, on MuSiQue 4-hop, MuSiQue ALL, 2WikiMultiHopQA 4-hop, and HotpotQA; a mark shows where the difference between the two shares excludes zero (▲: assembly share larger; ▼: reading share larger). (b) Reading share on MuSiQue 4-hop for every reader, with bootstrap 95% intervals. Gray bars: readers outside the grid of Table 2. Qwen3-8B-Base answers under its own minimal prompt; its cost is shown for reference, not for comparison.

**Read after the search, the loop's passages raise every grid reader's accuracy.** Over all MuSiQue questions, a re-read of the loop's passages lifts each grid reader above the ranking at the same passage count, the frontier model from 61.4 to 65.1 ($+3.7$ $[1.4, 6.1]$). The same holds for a second iterative system: IRCoT's passages, re-read, score 63.5 against 42.5 for its in-loop answer (Table 1). Better assembly thus shows both as more complete chains and as a gain for every grid reader. What separates this comparison from the end-to-end one is answering during the search versus reading after it; §5 locates the loss. On the deep stratum only Qwen3-8B's gain survives Benjamini–Hochberg correction (Table 2). The loop's list differs from the ranking in its wrong passages as well as its completeness, so §4 separates the two with the noise drawn from the index in both lists.

**Completeness alone misjudges value the other way.** With GPT-5.5 as driver, completeness over all MuSiQue questions stays within noise of the frontier model's own loop (77.2% against 78.5%, $[-3.6, 0.9]$; Table 1). Yet every frontier reader, GPT-5.5 included, scores three to four points lower on GPT-5.5's Loop passages over the same questions (Table 1). Neither accuracy nor completeness alone ranks a retrieval, which is why the standard of §1 asks for both at a matched passage count.

**Takeaway.** Measured end to end, iteration buys nothing on MuSiQue; measured apart, it assembles far more of each chain, and its passages raise accuracy once read after the search. Accuracy can hide an assembly gain and completeness can hide a loss in value, so a retrieval comparison needs both at a matched passage count. How much a reader loses to wrong passages even when the chain is complete is its reading share, which §4 measures reader by reader.

## 4 Where Accuracy Stops Depends on the Reader

§3 left open how much of the assembled value reaches each reader, because the loop's list changes its wrong passages along with its completeness. This section holds the noise to the index's own passages and splits each reader's gap to gold evidence into its assembly and reading shares (§2).

**On MuSiQue's deep questions, the 8B readers lose most of their gap to reading.** Their reading share exceeds their assembly share by 13.3 points for the distilled 8B and by 21.7 for Qwen3-8B (Table 3, top); with the gold passages ranked or shuffled, by at least 16.9, all intervals excluding zero. For the frontier model the assembly share is the larger one over all MuSiQue questions ($+4.8$ $[1.6, 8.0]$), but not detectably so on the deep stratum. For the 8B readers, a more complete chain closes less of the gap than a cleaner list would.

**The split follows reader strength.** The three frontier readers lose at most 10.8 points to the index's wrong passages and every open reader up to 32B at least 21.7, at least twice as much by point estimate. Each of these open readers' differences from the frontier model excludes zero (Table 3, bottom; Figure 3b), and within Qwen3 the reading share falls with scale until, at 235B, no

Table 2: The grid (§2), strict exact match: conditions as columns, readers as rows. COMP = completeness (%); CB = closed book; Loop = loop passages as a plain list; Gold+d. = gold+distractors. †The closed-book cells of the Qwen3-8B rows are Qwen3-8B-Base's, under its own minimal prompt (§2), shown for reference.

| | **CB** | **Top-5** | **Top-20** | **Top-$K_q$** | **Loop** | **Gold** | **Gold+d.** |
|---|---|---|---|---|---|---|---|
| *MuSiQue 4-hop (n=166)* | | | | | | | |
| COMP | 0 | 0.6 | 14.5 | 18.7 | 46.4 | 100 | 100 |
| Opus 4.6 | 13.3 | 27.7 | 41.0 | 42.2 | 50.6 | 71.1 | 60.2 |
| dist. 8B | 4.8 | 19.9 | 17.5 | 19.3 | 24.7 | 51.2 | 27.7 |
| Qwen3-8B | 7.2† | 19.3 | 19.9 | 21.1 | 33.1 | 57.2 | 27.7 |
| *MuSiQue ALL (n=1,000)* | | | | | | | |
| COMP | 0 | 34.6 | 51.0 | 49.5 | 78.5 | 100 | 100 |
| Opus 4.6 | 29.2 | 52.4 | 61.4 | 61.4 | 65.1 | 74.2 | 70.2 |
| dist. 8B | 8.2 | 37.7 | 38.8 | 38.4 | 42.7 | 62.6 | 45.4 |
| Qwen3-8B | 8.3† | 38.6 | 41.6 | 41.7 | 47.4 | 60.9 | 49.2 |
| *2WikiMultiHopQA 4-hop (n=241) / HotpotQA ALL (n=1,000)* | | | | | | | |
| COMP | 0/0 | 38.2/92.6 | 63.5/98.5 | 69.3/98.8 | 99.6/98.6 | 100/100 | 100/100 |
| Opus 4.6 | 84.2/47.5 | 95.9/69.3 | 96.3/71.5 | 97.1/71.0 | 98.3/71.0 | 98.8/73.6 | 98.8/71.0 |
| dist. 8B | 36.5/22.4 | 76.8/61.5 | 80.9/61.7 | 85.9/61.3 | 90.9/60.3 | 93.4/67.1 | 90.0/62.4 |
| Qwen3-8B | 39.0†/21.3† | 61.4/62.6 | 75.9/62.6 | 79.7/62.2 | 93.4/63.6 | 92.9/66.8 | 93.4/63.9 |

difference from the frontier model is detected. Completing the chain is worth 11.4 points more to the frontier model than to Qwen3-8B ($[3.0, 19.9]$); against the distilled 8B the difference does not survive Benjamini–Hochberg correction (Table 3, top). What a reader loses to the same passages therefore moves with the reader.

**Small and frontier readers lose to different passages.** A non-gold passage is *answer-bearing* if it contains the gold answer, and *chain-naming* if it names an entity used by a later hop of the gold chain. To locate what in the *padding*, the wrong passages added to the gold ones, costs a reader, *padding variants* replace it with random corpus passages, with ranks 101–200, or with top-ranked passages that exclude both kinds. Random padding alone costs the 7–8B readers 17 to 25 questions and the frontier readers at most 5 (Table 3, bottom). Swapping part of the padding for chain-naming passages of similar rank lowers the frontier model's accuracy by 6.6 points ($[-11.4, -1.8]$). Part of the 7–8B readers' loss therefore comes from the added passages themselves, while the frontier model's comes from passages that name the chain.

**The pattern appears only on the corpus whose index names the chain.** On 2WikiMultiHopQA's deep stratum and on HotpotQA, no grid reader's reading share exceeds 4.7 points (Table 2), although on HotpotQA it is the larger share of a small gap for the frontier model and the distilled 8B (Figure 3a). On 2WikiMultiHopQA's deep stratum the larger share is assembly for the frontier model and Qwen3-8B (Figure 3a). The corpora differ in what the index can return: a median MuSiQue deep question has 80 chain-naming passages in the corpus, against 1 on 2WikiMultiHopQA's deep questions with a bridge entity and 0 on HotpotQA's bridge questions. Where accuracy stops is therefore a property of the reader–corpus pair, not of the retrieval alone.

**Takeaway.** With the index's noise held fixed, most of the 8B readers' gap on MuSiQue's deep questions is reading, and completing the chain is worth more to the frontier model than to Qwen3-8B, which loses far more to the same passages. The split follows reader strength and appears only on MuSiQue, whose index returns passages that name the chain. Where accuracy stops is thus a property of the reader–corpus pair; §5 asks whether training on the frontier model's search moves either share.

## 5 Training Moves One Share, Not the Other

§4 found that, on MuSiQue's deep questions, the readers a deployed system can afford lose most of their gap to reading. If assembly and reading are separate abilities, training such a reader on a frontier model's search should move the first and leave the second. We test this by distilling the loop into an 8B, call by call. The teacher runs it on 1,313 MuSiQue training questions, disjoint from the development sets, and each of about 10,000 calls from its correct runs becomes one fine-

Table 3: MuSiQue 4-hop unless noted; ALL = all MuSiQue questions. Top: each grid reader's assembly share $A$ and reading share $R$, their difference $S$, and the frontier model's assembly share minus the reader's, with 95% intervals; *not significant after Benjamini–Hochberg correction. Bottom: questions lost as the padding moves from random passages to ranks 101–200 (+rank), to top ranks without chain-naming or answer-bearing passages (+topical), to the index's own, which adds chain-naming and answer-bearing passages (+index); last column, cost in points minus the frontier model's, with 95% intervals.

**Split of the gap to gold evidence**

| | $A$ | $R$ | $S{=}A{-}R$ | $S$, **ALL** | **frontier** $A-A$ |
|---|---|---|---|---|---|
| Claude Opus 4.6 | 19.3 | 10.8 | $+8.4$ $[-0.6, 16.9]$ | $+4.8$ $[1.6, 8.0]$ | — |
| Distilled 8B | 10.2 | 23.5 | $-13.3$ $[-23.5, -3.0]$ | $-10.6$ $[-15.1, -6.2]$ | $+9.0$ $[0.6, 17.5]^*$ |
| Qwen3-8B | 7.8 | 29.5 | $-21.7$ $[-33.1, -10.2]$ | $-4.1$ $[-8.6, 0.4]$ | $+11.4$ $[3.0, 19.9]$ |

**Distractor cost by padding source**

| | **random** | **+rank** | **+topical** | **+index** | **total** | **cost** | **− frontier model** |
|---|---|---|---|---|---|---|---|
| Claude Opus 4.6 | $-3$ | 9 | 1 | 11 | 18 | 10.8 | — |
| GPT-5.5 | 0 | 5 | 6 | 3 | 14 | 8.4 | — |
| Gemini 3.8 Flash | 5 | 9 | 6 | $-2$ | 18 | 10.8 | — |
| Qwen3-235B-A22B | 0 | 7 | 12 | 0 | 19 | 11.4 | $+0.6$ $[-7.2, 8.4]$ |
| Qwen3-32B | 4 | 12 | 16 | 6 | 38 | 22.9 | $+12.0$ $[3.6, 20.5]$ |
| Qwen3-14B | 11 | 9 | 11 | 9 | 40 | 24.1 | $+13.3$ $[4.8, 21.7]$ |
| Qwen3-8B | 20 | 6 | 15 | 8 | 49 | 29.5 | $+18.7$ $[10.2, 27.1]$ |
| OLMo-3-7B-Instruct | 25 | 0 | 11 | 0 | 36 | 21.7 | $+10.8$ $[3.0, 19.3]$ |
| Distilled 8B | 17 | 9 | 23 | $-10$ | 39 | 23.5 | $+12.7$ $[4.8, 20.5]$ |

tuning example for Qwen3-8B-Base. A second stage retrains on the student's own rollouts after the teacher repairs their earliest error. Unlike agent distillation, which trains on whole trajectories (Kang et al., 2025), each call is its own example; unlike on-policy distillation from the teacher's token probabilities (Agarwal et al., 2024), the repair stage uses only the teacher's text.

**End to end, the student learns the loop, but that number mixes two abilities.** Running alone, the student closes 70% of the gap between its untrained base and the teacher on MuSiQue, and the same MuSiQue-trained checkpoint closes 81% and 79% on the other two corpora (Figure 4a). The repair stage adds no detected gain ($+0.5$ $[-1.6, 2.6]$), and the closed-book score barely moves (Table 2), so memorized answers do not explain it. As in §3, this number mixes what the student learned to find with how it reads what it finds, so we separate the two.

**What transfers is the teacher's decisions of where to search.** Four controls train the same base for one stage on identical inputs and change only the targets: the teacher's decisions permuted across questions (*DR*), reasoning removed (*NR*), or gold answers with the call format kept (*FP*) or not (*bare*). With permuted decisions no search is learned, while removing the reasoning costs little against the one-stage student ($-3.9$ $[-6.7, -1.2]$), and the gold-target controls score near zero (Figure 4b). As a driver, the student completes most of the chains its teacher completes over all MuSiQue questions, though far fewer on the deep stratum (Table 1).

**The reading share does not move with them.** If only the search transfers, the student should lose to the index's wrong passages what any model fine-tuned on the same inputs loses. No difference is detected: on MuSiQue's deep stratum the student loses 23.5 points and each gold-target control 24.1 ($-0.6$ $[-8.4, 7.2]$ against the bare control). Against the untuned Qwen3-8B the difference also spans zero ($-6.0$ $[-15.1, 3.0]$; Figure 4c). After training, the student still loses 12.7 points more than the frontier model to the same passages ($[4.8, 20.5]$; Table 3).[1]

**The loop loses by answering during the search.** A finished search can be answered by its driver in the loop, by one *fresh call* over its trajectory, or by a re-read of its passages. For the teacher over all MuSiQue questions these give 57.8, 64.3, and 65.1 (Figure 4d), so one fresh call recovers most of the in-loop loss, in a contrast that also changes decoding (§2). The frontier model's re-read of the student's passages scores 15.9 points above the student's own in-loop answer (Table 1): the student's search outruns its reading, the same split seen from the student's side.

[1]The untrained base's higher cost in Figure 3b is measured under its own minimal prompt; the gold-target controls, fine-tuned on the same inputs, already reach the student's cost.

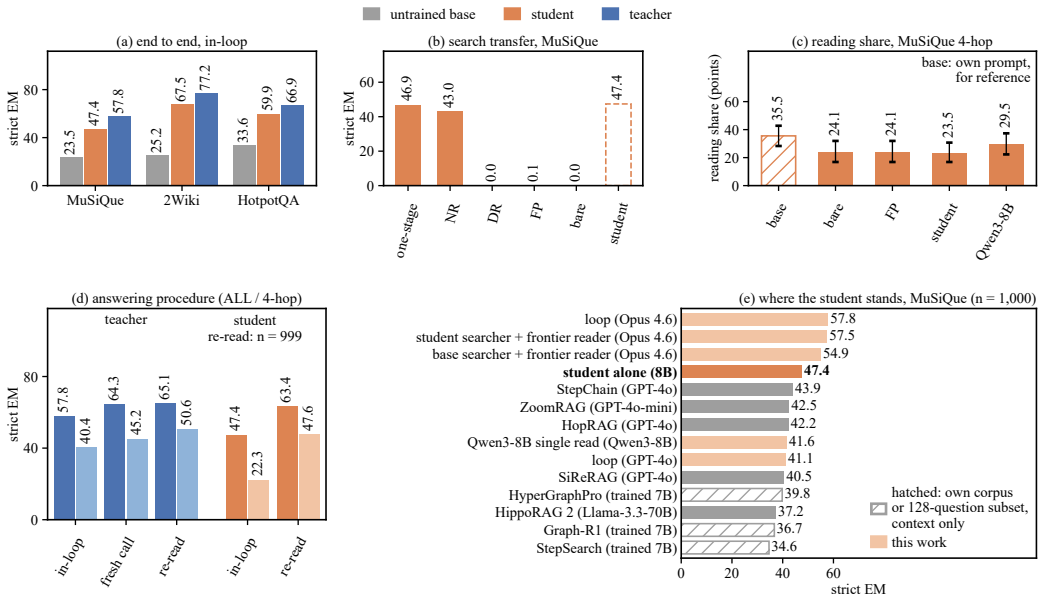



Figure 4: Distillation (§5), strict EM. (a) Running the loop alone: the untrained base, the student, and the teacher on the three corpora. (b) The one-stage student and the four controls on MuSiQue. (c) Reading share on MuSiQue's deep stratum, with 95% intervals; the base answers under its own prompt and is shown for reference. (d) Answering a finished search, over all MuSiQue questions and its deep stratum. (e) MuSiQue; each bar names the model that reads its evidence. Published systems use the same corpus and questions (Gutiérrez et al., 2025; Zhang et al., 2025; Liu et al., 2025; Hu et al., 2026; Ni et al., 2025); hatched bars are trained 7B searchers on their own corpora or question subsets, shown for context only (Zheng et al., 2025; Luo et al., 2026; Park et al., 2026).

**Running alone, the student outscores the published systems on the same questions.** With no frontier call, the student scores 47.4 on MuSiQue, above every published system in Figure 4e evaluated on the same corpus and questions. It also scores 5.8 points above a single read by the same-size Qwen3-8B ($[2.8, 8.9]$), so the gain is not only a newer base model's. On HotpotQA, which one pass already completes, a single read remains better.

**Its searches are best read by a stronger reader.** Since the student acquires the search with no detectable change in its reading share, the reading step belongs with a stronger reader. A frontier model reading the student's trajectories scores 57.5, against the student's own 47.4 and the teacher's own loop at 57.8 (Figure 4e). Even the untrained base reaches 54.9 as a searcher for the same reader, so at this end most of the score is the reader's.[2]

**Takeaway.** End to end, the student closes most of the gap to its teacher. Measured apart, it has learned where to search, but its reading share is no detectably lower than its controls', and it still loses more to wrong passages than the frontier model. Training on a frontier model's search thus moves assembly and leaves the reading share.

## 6 From the Split to a Choice

Two measured quantities place each reader–corpus pair and so say what is worth paying for (Figure 5). The completeness gap of a single read bounds what better assembly can add; the reader's reading share is what it loses to wrong passages even when the evidence is complete.

[2]The same reader scores the teacher's trajectories higher (64.3, Figure 4d), under a different prompt and decoding, so the student's search is not shown to match the teacher's for this reader.

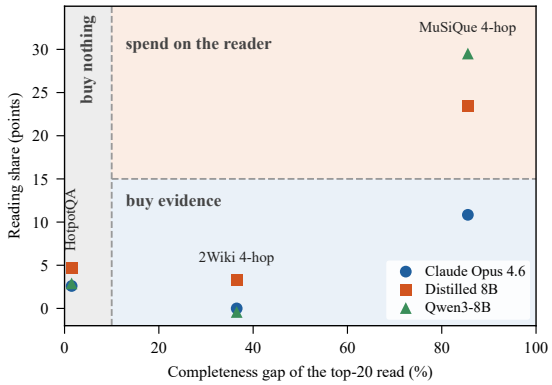



Figure 5: A rule from the split (§6). Each point is a grid reader on one corpus, placed at the completeness gap of its top-20 read (100 minus COMP, Table 2) and at the reader's reading share. Boundaries are drawn for legibility, not fitted.

**Where one pass already completes the chains, buy nothing.** On HotpotQA a single read leaves almost no chain incomplete, and it scores 71.5 against 71.0 for the loop's re-read (Table 2), so the loop's added cost buys no accuracy.

**Where the gap is large and the reader keeps its evidence, buy evidence.** For the frontier model on MuSiQue, the loop with a re-read lifts accuracy from 61.4 to 65.1 at the same passage count (§3). A 200-passage single read reaches the same accuracy at similar token cost (Table 1), so evidence can be bought either way. On 2WikiMultiHopQA's deep stratum the same branch applies but little is left to win: complete evidence adds only 2.5 points to the frontier model's single read (Table 2).

**Where the reader loses much to noise, spend on the reader.** On MuSiQue's deep questions most of an 8B reader's gap is reading (§4). Over all MuSiQue questions, re-reading the loop's passages lifts Qwen3-8B by 5.7 points at the same passage count, yet leaves it 17.7 points below the frontier model reading the same passages (Table 2). Training the 8B base on the loop's search lifts its own loop by 23.9 points but does not detectably lower its reading share (§5), so the reading step belongs to the strongest reader available.

## 7 Limitations and Conclusion

**Limitations.** We compare one graph index with one loop and one frontier teacher; indexes built without relations or around communities (Zhuang et al., 2026; Edge et al., 2024) may split differently. The split by reader is resolved on MuSiQue's deep questions, and for the frontier model only over all MuSiQue questions. The student comes from one teacher, one base model, and one training corpus. The split is measured over sets of questions, so it guides choices per reader and corpus, not per question.

**Conclusion.** End-to-end accuracy cannot tell whether a system failed to assemble its evidence or its reader failed to use it. Holding one step fixed while the other varies splits each reader's gap into the two, which changes verdicts. On MuSiQue, iteration assembles far more of each chain than a graph index at the same passage count, and its passages raise accuracy when read after the search. On the deep questions the 8B readers lose most of their gap to reading, which training on a frontier model's search does not detectably reduce. Where accuracy stops is thus a property of the reader–corpus pair: retrieval claims should report completeness at a matched passage count and gains across readers, and deployments should give the reading step to the strongest affordable reader.

### AI use statement

Large language models are both objects of study and research tools in this work. As objects of study, frontier and open models serve as drivers, teachers, and readers; every claim about them

is an experimental result. As tools, AI assistants helped frame the research questions in writing, drafted experimental designs, wrote the retrieval, training, evaluation, and analysis code and the data-processing scripts, drafted interpretations of results, and drafted and edited text. The authors selected and approved every experimental design, tested the code, ran every experiment, checked every reported number, and revised all text. The student's training data are synthetic: they are the frontier model's own calls when it runs the loop on MuSiQue training questions (§5). AI was not used to generate or alter the benchmarks or to prove mathematical claims. The authors take full responsibility for the content of this work.

### Reproducibility statement

All corpora and question sets are the public HippoRAG releases. The supplementary material contains the code and prompts for the loop, every reading condition, the distillation and its controls, and every analysis and figure, together with the training configuration and the per-question scores behind every reported number. The trained checkpoints are withheld only while the review is anonymous and will be released publicly once it ends; until then they can be reproduced from the released training code and configuration.

## References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Jifan Chen and Greg Durrett. Understanding dataset design choices for multi-hop reasoning. In *NAACL*, 2019.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 719–729, 2024.

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. *arXiv preprint arXiv:2404.16130*, 2024.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10746–10761, 2025.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *International Conference on Machine Learning (ICML)*, 2025.

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

Xianming Hu, Jingyang Chen, Bin Tang, Yihe Liu, Yihong Huang, Hongbo Zhao, Nuoyi Chen, Jie Zhang, Ping Li, and Kai Zhang. Zoomrag: Hierarchical random-walk zooming across multi-scale information graphs for fast and accurate rag. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32840–32859, 2026.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *EMNLP*, 2023.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Conference on Language Modeling (COLM)*, 2025.

Minki Kang, Jongwon Jeong, Seanie Lee, Jaewoong Cho, and Sung Ju Hwang. Distilling llm agent into small models with retrieval and code tools. *Advances in Neural Information Processing Systems*, 38:106501–106538, 2025.

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 *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1897–1913, 2025.

Haoran Luo, Haihong E, Guanting Chen, Qika Lin, Yikai Guo, Fangzhi Xu, Zemin Kuang, Meina Song, Xiaobao Wu, Yifan Zhu, and Luu Anh Tuan. Graph-R1: Towards agentic GraphRAG framework via end-to-end reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2026.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. Compositional questions do not necessitate multi-hop reasoning. In *ACL*, 2019.

Tengjun Ni, Xin Yuan, Shenghong Li, Kai Wu, Ren Ping Liu, Wei Ni, and Wenjie Zhang. Stepchain graphrag: Reasoning over knowledge graphs for multi-hop question answering. *arXiv preprint arXiv:2510.02827*, 2025.

Jinyoung Park, Sanghyeok Lee, Omar Zia Khan, Hyunwoo J Kim, and Joo-Kyung Kim. Hypergraphpro: Progress-aware reinforcement learning for structure-guided hypergraph rag. *arXiv preprint arXiv:2601.17755*, 2026.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of EMNLP*, 2023.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pp. 32628–32649, 2024.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of EMNLP*, 2023.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227, 2023.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 10014–10037, 2023.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations (ICLR)*, 2021.

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 *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations (ICLR)*, 2024.

Nan Zhang, Prafulla Kumar Choubey, Alexander Fabbri, Gabriel Bernadett-Shapiro, Rui Zhang, Prasenjit Mitra, Caiming Xiong, and Chien-Sheng Wu. Sirerag: Indexing similar and related information for multihop reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 100413–100432, 2025.

Xuhui Zheng, Kang An, Ziliang Wang, Yuhang Wang, and Yichao Wu. Stepsearch: Igniting llms search ability via step-wise proximal policy optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21816–21841, 2025.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. Linearrag: Linear graph retrieval augmented generation on large-scale corpora. In *International Conference on Learning Representations*, volume 2026, pp. 147053–147075, 2026.

# A Appendix

## A.1 Setup

**Corpora and evidence lists.** §2 uses the HippoRAG 2 releases of MuSiQue, 2WikiMultiHopQA, and HotpotQA, 1,000 development questions each; the deep stratum is each corpus's four-hop subset (166 and 241 questions). A question's evidence is complete if every gold passage appears in the list, with passages matched by text. Top-$k$ cuts the deployed ranking at $k$. $K_q$ is the number of distinct passages, by text, that any call of the loop read for question $q$, including branches later pruned, and Loop passages lists them once each, in order of first appearance; $K_q$ has median 18 on MuSiQue and 25 on its deep stratum. Per MuSiQue question, the loop's calls take 25.8K input tokens and its re-read 5.5K, against 4.0K for a top-20 read. Gold lists follow annotation order. Gold+distractors places the gold passages first and fills the list to twenty with the ranking's highest-ranked non-gold passages, in ranking order.

**Padding.** A non-gold passage is *answer-bearing* if it contains the normalized gold answer, and *chain-naming* if it contains, as a normalized substring, an entity consumed by a later hop of the gold chain, the final answer excluded. Every padding variant keeps the gold passages first and fills to twenty passages: *random* draws non-gold corpus passages under a fixed seed; *ranks 101–200* takes the ranking's non-gold passages from that band, in order; the top-ranked variants scan the ranking from the top, excluding answer-bearing passages and either excluding chain-naming ones (D0) or holding exactly four of them (D4). D0 and D4 place their padding at mean ranking positions 15.4 and 14.6, so the swap of §4 holds rank about equal.

**Stated hops.** A hop of a gold chain is *stated* if some single passage mentions both of its entities under deterministic alias normalization; the dotted lines of Figure 2 give the share of deep chains with every hop stated (71.1% on MuSiQue, 99.6% on 2WikiMultiHopQA), where chains without a measurable bridging hop count as stated.

## A.2 The loop and its three answers

**Structure.** The loop (Figure 1) decomposes the question into answer slots with anchor entities and retrieves an initial list of passages. A first read fills the slots the passages support and proposes at most five candidates for the earliest open slot. Each candidate retrieves its own top ten passages for the candidate and its sub-question, and one read decides whether it resolves the slot, is dropped, or rewrites the next query. Survivors merge and are pruned to five, and rounds repeat until a verbatim quote resolves the final slot or a round cap is reached; resolved entities are not searched again.

Table 4: Models (§2, §5).

| Role | Model |
|---|---|
| Frontier model, loop driver, teacher | claude-opus-4-6 |
| Other frontier readers | gpt-5.5; gemini-3.8-flash |
| Student base | Qwen3-8B-Base |
| Open readers | Qwen/Qwen3-8B; allenai/Olmo-3-7B-Instruct; Qwen/Qwen3-14B, Qwen3-32B, Qwen3-235B-A22B |
| Loop driver in Figure 4e | gpt-4o |
| Loop index embedding | embed-v-4-0 (1536-d) |
| HippoRAG 2 embedding and extraction | nvidia/NV-Embed-v2; gpt-4o-mini |

Retrieval runs over the loop's own index, which pairs a dense retriever with an entity-to-passage map without relations. Prompt templates for all six call types are in the supplementary material.

**Three ways to answer a finished search.** The *in-loop answer* is the driver's final call inside the trajectory, sampled at temperature 0.2. A *fresh call* reads the whole trajectory in one new call at temperature 0 under the answer prompt of §A.3. A *re-read* gives a reader the Loop passages as a plain list under the same prompt. The trajectory reads of §5 that score 57.5 and 54.9 use a fixed trajectory rendering and prompt, sampled at temperature 0.2, which is why they differ in prompt and decoding from the fresh call behind 64.3.

### A.3 Readers, prompt, and scoring

Table 4 lists the models. Every instruction-following reader answers under one prompt whose final sentence reads: *Your final answer after* `Answer:` *must be the exact entity or minimal phrase copied verbatim from the passages, as short as possible, with no extra words.* Qwen3-8B-Base answers under its own minimal prompt, included in the supplementary material. Answers are scored by strict exact match under the official normalization; an empty answer, blank or not in the required format, scores zero. A list longer than an 8B reader's context window is scored incorrect; this affects only the 8B readers' Loop passages lists, 8 of 166 on MuSiQue's deep stratum.

### A.4 Statistics

Paired contrasts are tested question by question: two-sided exact McNemar on the discordant pairs, and 95% intervals from resampling questions with replacement ($B$=10,000). A difference between two readers' differences resamples questions jointly and recomputes both differences in each replicate. The share differences $S$ of Table 3 use the same resampling, with a sign-flip permutation test for their $p$ values. Benjamini–Hochberg correction ($q$=0.05) is applied to one family: the family comprises every paired contrast whose significance the main text states, 45 in all (Table 5). Two lose significance after correction, marked $^*$ in the table.

### A.5 Completeness by passage budget

Figure 2 plots completeness on the deep strata against the passage budget; Table 6 gives its values. The loop and the ranking cut at each question's own $K_q$ sit at the mean $K_q$, 28.4 on MuSiQue and 34.9 on 2WikiMultiHopQA.

### A.6 Comparators

Table 1 gives the comparisons §3 and §5 rely on; Table 7 gives them on both strata, with the frontier-model pools that GPT-5.5 as driver changes.

### A.7 Padding sources and gold placement

§4 reports each reader's reading share and what in the padding costs it; Table 8 gives every reader's strict EM under each padding source on MuSiQue's deep stratum, and Table 9 the share split under three gold placements. Gold minus random padding reads $-1.8$ $[-4.8, 1.2]$ for Claude Opus 4.6,

Table 5: The Benjamini–Hochberg family (§A.4): every paired contrast whose significance the main text states. Estimates in points of strict EM, or of completeness for #7; $S$ = assembly share minus reading share; 95% intervals; $q$ after correction; *not significant after correction.

| # | Where | Contrast | $n$ | Est. | 95% interval | $p$ | $q$ |
|---|---|---|---|---|---|---|---|
| 1 | §3 | re-read − top-$K_q$, Opus 4.6, ALL | 1000 | $+3.7$ | $[1.4, 6.1]$ | 0.0020 | 0.0064 |
| 2* | §3 | re-read − top-$K_q$, Opus 4.6, 4-hop | 166 | $+8.4$ | $[1.2, 15.7]$ | 0.044 | 0.059 |
| 3 | §3 | re-read − top-$K_q$, distilled 8B, ALL | 1000 | $+4.3$ | $[1.5, 7.1]$ | 0.0037 | 0.0084 |
| 4 | §3 | re-read − top-$K_q$, distilled 8B, 4-hop | 166 | $+5.4$ | $[-0.6, 12.0]$ | 0.14 | 0.16 |
| 5 | §3 | re-read − top-$K_q$, Qwen3-8B, ALL | 1000 | $+5.7$ | $[3.0, 8.4]$ | 3.4e-5 | 2.6e-4 |
| 6 | §3 | re-read − top-$K_q$, Qwen3-8B, 4-hop | 166 | $+12.0$ | $[5.4, 18.7]$ | 8.2e-4 | 0.0028 |
| 7 | §3 | completeness, GPT-5.5 − frontier driver, ALL | 1000 | $-1.3$ | $[-3.6, 0.9]$ | 0.29 | 0.33 |
| 8 | §3 | GPT-5.5's − frontier's pools, read by Opus 4.6 | 1000 | $-3.3$ | $[-5.4, -1.3]$ | 0.0022 | 0.0067 |
| 9 | §3 | read by GPT-5.5 | 1000 | $-3.0$ | $[-5.0, -1.1]$ | 0.0032 | 0.0076 |
| 10 | §3 | read by Gemini 3.8 Flash | 1000 | $-3.7$ | $[-5.7, -1.7]$ | 3.9e-4 | 0.0015 |
| 11 | §3 | top-200 read − loop re-read, ALL | 1000 | 0.0 | $[-2.1, 2.1]$ | 1.0 | 1.0 |
| 12 | §3 | IRCoT re-read − in-loop answer, ALL | 1000 | $+21.0$ | $[18.3, 23.8]$ | 7.0e-50 | 1.6e-48 |
| 13 | §4 | $S$, Opus 4.6, MuSiQue 4-hop | 166 | $+8.4$ | $[-0.6, 16.9]$ | 0.086 | 0.11 |
| 14 | §4 | $S$, Opus 4.6, MuSiQue ALL | 1000 | $+4.8$ | $[1.6, 8.0]$ | 0.0043 | 0.0092 |
| 15 | §4 | $S$, distilled 8B, MuSiQue 4-hop | 166 | $-13.3$ | $[-23.5, -3.0]$ | 0.020 | 0.030 |
| 16 | §4 | $S$, distilled 8B, MuSiQue ALL | 1000 | $-10.6$ | $[-15.1, -6.2]$ | <1e-4 | 5.6e-4 |
| 17 | §4 | $S$, Qwen3-8B, MuSiQue 4-hop | 166 | $-21.7$ | $[-33.1, -10.2]$ | 3.0e-4 | 0.0012 |
| 18 | §4 | $S$, Qwen3-8B, MuSiQue ALL | 1000 | $-4.1$ | $[-8.6, 0.4]$ | 0.081 | 0.11 |
| 19* | §4 | frontier $A - A$, distilled 8B, 4-hop | 166 | $+9.0$ | $[0.6, 17.5]$ | 0.045 | 0.059 |
| 20 | §4 | frontier $A - A$, Qwen3-8B, 4-hop | 166 | $+11.4$ | $[3.0, 19.9]$ | 0.010 | 0.016 |
| 21 | Table 3 | cost − frontier, Qwen3-235B-A22B | 166 | $+0.6$ | $[-7.2, 8.4]$ | 0.93 | 0.96 |
| 22 | Table 3 | cost − frontier, Qwen3-32B | 166 | $+12.0$ | $[3.6, 20.5]$ | 0.0054 | 0.011 |
| 23 | Table 3 | cost − frontier, Qwen3-14B | 166 | $+13.3$ | $[4.8, 21.7]$ | 0.0030 | 0.0075 |
| 24 | Table 3 | cost − frontier, Qwen3-8B | 166 | $+18.7$ | $[10.2, 27.1]$ | <1e-4 | <1e-3 |
| 25 | Table 3 | cost − frontier, OLMo-3-7B-Instruct | 166 | $+10.8$ | $[3.0, 19.3]$ | 0.010 | 0.016 |
| 26 | Table 3 | cost − frontier, distilled 8B | 166 | $+12.7$ | $[4.8, 20.5]$ | 0.0030 | 0.0075 |
| 27 | §4 | chain-naming swap D4 − D0, Opus 4.6, 4-hop | 166 | $-6.6$ | $[-11.4, -1.8]$ | 0.013 | 0.020 |
| 28 | Figure 3a | $S$, Opus 4.6, 2Wiki 4-hop | 241 | $+2.5$ | $[0.8, 4.6]$ | 0.030 | 0.043 |
| 29 | Figure 3a | $S$, distilled 8B, 2Wiki 4-hop | 241 | $+5.8$ | $[-1.2, 12.4]$ | 0.13 | 0.15 |
| 30 | Figure 3a | $S$, Qwen3-8B, 2Wiki 4-hop | 241 | $+17.8$ | $[10.8, 24.9]$ | <1e-4 | 5.6e-4 |
| 31 | Figure 3a | $S$, Opus 4.6, HotpotQA ALL | 1000 | $-3.1$ | $[-5.5, -0.7]$ | 0.016 | 0.024 |
| 32 | Figure 3a | $S$, distilled 8B, HotpotQA ALL | 1000 | $-4.0$ | $[-6.8, -1.3]$ | 0.0057 | 0.011 |
| 33 | Figure 3a | $S$, Qwen3-8B, HotpotQA ALL | 1000 | $-1.6$ | $[-4.8, 1.7]$ | 0.37 | 0.40 |
| 34 | §4 | $S$, distilled 8B, 4-hop, gold ranked | 166 | $-18.1$ | $[-28.9, -7.2]$ | 0.0025 | 0.0070 |
| 35 | §4 | $S$, distilled 8B, 4-hop, gold shuffled | 166 | $-16.9$ | $[-28.9, -4.8]$ | 0.0072 | 0.014 |
| 36 | §4 | $S$, Qwen3-8B, 4-hop, gold ranked | 166 | $-22.9$ | $[-33.7, -12.0]$ | 2.0e-4 | 9.3e-4 |
| 37 | §4 | $S$, Qwen3-8B, 4-hop, gold shuffled | 166 | $-16.9$ | $[-28.3, -5.4]$ | 0.0081 | 0.014 |
| 38 | §5 | DR − one-stage student, ALL | 1000 | $-46.9$ | $[-50.0, -43.7]$ | 1.3e-141 | 1.2e-140 |
| 39 | §5 | NR − one-stage student, ALL | 1000 | $-3.9$ | $[-6.7, -1.2]$ | 0.008 | 0.014 |
| 40 | §5 | reading share, student − bare, 4-hop | 166 | $-0.6$ | $[-8.4, 7.2]$ | 0.94 | 0.96 |
| 41 | §5 | reading share, student − Qwen3-8B, 4-hop | 166 | $-6.0$ | $[-15.1, 3.0]$ | 0.22 | 0.25 |
| 42 | §5 | repair stage, two-stage − one-stage, ALL | 1000 | $+0.5$ | $[-1.6, 2.6]$ | 0.71 | 0.76 |
| 43 | §5 | teacher, fresh call − in-loop answer, ALL | 1000 | $+6.5$ | $[4.8, 8.2]$ | <1e-6 | <1e-5 |
| 44 | §5 | frontier re-read − student in-loop, ALL | 999 | $+15.9$ | $[13.0, 18.9]$ | 1.7e-24 | <1e-22 |
| 45 | §5 | student alone − Qwen3-8B single read, ALL | 1000 | $+5.8$ | $[2.8, 8.9]$ | 2.1e-4 | 9.3e-4 |

Table 6: Completeness (%) against passage budget $K$ on the deep strata (Figure 2); — = not plotted.

| | | $K$ | | | | | | | at $K_q$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| Curve | Stratum | 5 | 10 | 20 | 50 | 60 | 100 | 200 | HR2 | Loop |
| HR2 | MuSiQue 4-hop | 0.6 | 4.8 | 14.5 | 26.5 | 31.3 | 39.8 | 47.6 | 18.7 | 46.4 |
| | 2Wiki 4-hop | 38.2 | 52.7 | 63.5 | 73.0 | 75.1 | 80.9 | 88.8 | 69.3 | 99.6 |
| HR2 w/o PPR | MuSiQue 4-hop | — | — | 11.4 | — | — | — | 44.6 | | |
| | 2Wiki 4-hop | — | — | 14.1 | — | — | — | 42.7 | | |
| Dense | MuSiQue 4-hop | 0.6 | — | 9.6 | — | 27.1 | 43.4 | 50.6 | | |
| | 2Wiki 4-hop | 6.2 | — | 7.5 | — | 9.5 | 13.3 | 19.9 | | |

$0.0$ $[-3.6, 3.6]$ for GPT-5.5, $3.0$ $[-1.2, 7.2]$ for Gemini 3.8 Flash, $0.0$ $[-6.6, 6.6]$ for Qwen3-235B-A22B, $2.4$ $[-4.2, 9.0]$ for Qwen3-32B, $6.6$ $[-0.6, 13.9]$ for Qwen3-14B, $12.0$ $[5.4, 18.7]$ for Qwen3-8B, $15.1$ $[9.0, 21.7]$ for OLMo-3-7B-Instruct, and $10.2$ $[4.2, 16.3]$ for the distilled 8B. The swap D4 minus D0 reads $-6.6$ $[-11.4, -1.8]$ for the frontier model and $+8.4$ $[1.8, 15.1]$ for the distilled 8B; for every other reader its interval spans zero.

**Supply of chain-naming passages.** Counting non-gold passages anywhere in the corpus that name an entity of the chain, the median question has 80 on MuSiQue's deep stratum (quartiles 56 and 142), 1 on 2WikiMultiHopQA's deep stratum (0 and 2; the 225 questions with a bridge entity), and 0 on HotpotQA's bridge questions (0 and 2; 705 questions).

Table 7: Comparators in full (§3, Table 1); MuSiQue, strict EM, 95% intervals in brackets. The last block splits all questions, and the deep stratum in parentheses, by which driver's Loop passages are complete; entries are GPT-5.5's pools minus the frontier model's, read by the frontier model, GPT-5.5, and Gemini 3.8 Flash (every 95% interval in the "only the frontier model's" row excludes zero); — = not reported.

| | ALL | 4-hop |
|---|---|---|
| Top-200 read − loop re-read | $0.0\ [-2.1, 2.1]$ | $+2.4\ [-3.0, 8.4]$ |
| Input tokens, top-200 read / loop with re-read | 30.7K / 31.3K | 31.3K / 52.7K |
| IRCoT: distinct passages per question | 76.6 | 109.4 |
| IRCoT: completeness (%) | 84.0 | 44.6 |
| IRCoT: re-read / in-loop answer | 63.5 / 42.5 | 46.4 / 15.1 |
| GPT-5.5 driver: completeness / frontier model's (%) | 77.2 / 78.5 | 39.2 / 46.4 |
| GPT-5.5's − frontier model's pools, read by Opus 4.6 | $-3.3\ [-5.4, -1.3]$ | $-8.4\ [-15.1, -1.8]$ |
| read by GPT-5.5 | $-3.0\ [-5.0, -1.1]$ | $-7.8\ [-13.9, -1.8]$ |
| read by Gemini 3.8 Flash | $-3.7\ [-5.7, -1.7]$ | $-7.8\ [-13.9, -1.8]$ |
| Both complete (713; 48) | $+1.8$ / $+0.4$ / $+0.1$ | |
| Neither complete (156; 72) | $-3.8$ / $-7.7$ / $-5.8$ | |
| Only the frontier model's complete (72; 29) | $-55.6$ / $-31.9$ / $-41.7$ ($-51.7$ / $-34.5$ / $-41.4$) | |
| Only GPT-5.5's complete (59; 17) | $0.0$ / $+3.4$ / $+1.7$ | |

Table 8: Strict EM on MuSiQue's deep stratum ($n$=166) under the gold list, the index's own padding (Gold+distractors), and the padding variants of §A.1.

| Reader | Gold | Index padding | Random | 101–200 | D0 | D4 |
|---|---|---|---|---|---|---|
| Claude Opus 4.6 | 71.1 | 60.2 | 72.9 | 67.5 | 66.9 | 60.2 |
| GPT-5.5 | 68.7 | 60.2 | 68.7 | 65.7 | 62.0 | 62.7 |
| Gemini 3.8 Flash | 70.5 | 59.6 | 67.5 | 62.0 | 58.4 | 55.4 |
| Qwen3-235B-A22B | 59.6 | 48.2 | 59.6 | 55.4 | 48.2 | 45.8 |
| Qwen3-32B | 60.2 | 37.3 | 57.8 | 50.6 | 41.0 | 38.0 |
| Qwen3-14B | 56.6 | 32.5 | 50.0 | 44.6 | 38.0 | 37.3 |
| Qwen3-8B | 57.2 | 27.7 | 45.2 | 41.6 | 32.5 | 28.9 |
| OLMo-3-7B-Instruct | 41.0 | 19.3 | 25.9 | 25.9 | 19.3 | 16.9 |
| Distilled 8B | 51.2 | 27.7 | 41.0 | 35.5 | 21.7 | 30.1 |

### A.8 DISTILLATION

**Training.** §5 distills the loop call by call. The teacher runs the loop on 1,313 MuSiQue training questions, none of which appears in the three development sets. Each of the 9,994 calls from its correct runs, across the decomposition, first-read, branch, and merge call types, becomes one training example: the call's input as the prompt and the teacher's output as the target. Qwen3-8B-Base is fully fine-tuned on these examples, ordered by hop count, with over-length examples dropped (9,863 remain). In the second stage the student runs the loop on the training questions, the teacher repairs the earliest erroneous step of each failed run with at most five repairs per run (2,925 repair calls), and the student is retrained on the repaired runs (11,650 examples). Training is full-parameter in bf16 with a cosine schedule, 3% linear warmup, weight decay 0.01, gradient clipping at 1.0, and batches of 32,768 tokens with two accumulation steps; examples longer than 16,384 tokens are dropped. The first stage runs three epochs at a peak learning rate of $10^{-5}$ (1,545 steps), the second two epochs at $5 \times 10^{-6}$ (1,176 steps). Configuration files are in the supplementary material. Unlike agent distillation, which fine-tunes on whole teacher trajectories (Kang et al., 2025), each call here is its own example; unlike on-policy distillation, which matches the teacher's token probabilities on student samples (Agarwal et al., 2024), the repair stage uses only the teacher's text.

**Controls.** All four controls start from Qwen3-8B-Base and train one stage on the student's stage-one examples under its configuration, with every input identical and only the targets changed. DR permutes the first-read, branch, and merge targets across questions within each call type, so no example keeps its own question's target. NR empties every reasoning string in the targets and keeps every other field. FP keeps the teacher's call format, empties the retrieved values and cited evidence in its targets, and writes the gold answer into the merge slot. Bare writes the gold answer as the whole target.

Table 9: The share split of §4 on MuSiQue's deep stratum with the gold passages first, at their ranked positions, or shuffled; $A$ = assembly share, $R$ = reading share, $S = A - R$ with 95% intervals.

| Reader | Gold placement | $A$ | $R$ | $S$ |
|---|---|---|---|---|
| Distilled 8B | first | 10.2 | 23.5 | $-13.3$ $[-23.5, -3.0]$ |
| | ranked | 7.8 | 25.9 | $-18.1$ $[-28.9, -7.2]$ |
| | shuffled | 8.4 | 25.3 | $-16.9$ $[-28.9, -4.8]$ |
| Qwen3-8B | first | 7.8 | 29.5 | $-21.7$ $[-33.1, -10.2]$ |
| | ranked | 7.2 | 30.1 | $-22.9$ $[-33.7, -12.0]$ |
| | shuffled | 10.2 | 27.1 | $-16.9$ $[-28.3, -5.4]$ |

Table 10: Distillation in full (§5, Figure 4); MuSiQue unless noted, strict EM, 95% intervals in brackets. Self-contained runs use no frontier call.

| | |
|---|---|
| **End to end (ALL; MuSiQue / 2WikiMultiHopQA / HotpotQA)** | |
| Base / student / teacher, self-contained | 23.5 / 47.4 / 57.8; 25.2 / 67.5 / 77.2; 33.6 / 59.9 / 66.9 |
| Repair stage (student − one-stage), ALL / 4-hop | $+0.5$ $[-1.6, 2.6]$ / $-1.8$ $[-6.6, 3.0]$ |
| Closed book, Qwen3-8B-Base / student (ALL) | 8.3 / 8.2 |
| **Controls, self-contained** | |
| One-stage student / NR / DR / FP / bare, ALL | 46.9 / 43.0 / 0.0 / 0.1 / 0.0 |
| DR / NR − one-stage student, ALL | $-46.9$ $[-50.0, -43.7]$ / $-3.9$ $[-6.7, -1.2]$ |
| Final-hop recovery − base, DR / NR (495 questions) | $-32.5$ $[-37.0, -28.1]$ / $+8.5$ $[4.2, 12.5]$ |
| **Reading (student / FP / bare)** | |
| Gold, ALL / 4-hop | 62.6 / 58.6 / 55.0; 51.2 / 50.6 / 47.0 |
| Gold+distractors, ALL / 4-hop | 45.4 / 38.5 / 38.9; 27.7 / 26.5 / 22.9 |
| Reading share (4-hop) | 23.5 / 24.1 / 24.1 |
| Student − bare, reading share (4-hop) | $-0.6$ $[-8.4, 7.2]$ |
| Student − Qwen3-8B, reading share (4-hop) | $-6.0$ $[-15.1, 3.0]$ |
| Student alone − Qwen3-8B single read (ALL) | $+5.8$ $[2.8, 8.9]$ |
| **Answering a finished search (ALL / 4-hop)** | |
| Teacher: in-loop / fresh call / re-read | 57.8 / 64.3 / 65.1; 40.4 / 45.2 / 50.6 |
| Student in-loop / frontier re-read of it | 47.4 / 63.4; 22.3 / 47.6 |
| Frontier re-read − student in-loop | $+15.9$ $[13.0, 18.9]$ / $+25.3$ $[17.5, 33.1]$ |
| Frontier reads student / base / teacher trajectories | 57.5 / 54.9 / 64.3 |

### A.9 PUBLISHED SYSTEMS

Figure 4e places the student among published systems; Table 11 gives each system's reader, setting, and source.

Table 11: MuSiQue, strict EM as reported by each source; "matched" = the same corpus and 1,000 questions.

| System | Reader | Setting | Source | EM |
|---|---|---|---|---|
| HippoRAG 2 (Gutiérrez et al., 2025) | Llama-3.3-70B | matched | ICML'25 | 37.2 |
| SiReRAG (Zhang et al., 2025) | GPT-4o | matched | ICLR'25 | 40.5 |
| HopRAG (Liu et al., 2025) | GPT-4o | matched | ACL'25 Findings | 42.2 |
| ZoomRAG (Hu et al., 2026) | GPT-4o-mini | matched | ACL'26 Findings | 42.5 |
| StepChain (Ni et al., 2025) | GPT-4o | matched | preprint | 43.9 |
| StepSearch (Zheng et al., 2025) | trained 7B | own corpus | EMNLP'25 | 34.6 |
| Graph-R1 (Luo et al., 2026) | trained 7B | own graph, $n$=128 | ICML'26 | 36.7 |
| HyperGraphPro (Park et al., 2026) | trained 7B | $n$=128 | preprint | 39.8 |
| Qwen3-8B single read | Qwen3-8B | matched | this work | 41.6 |
| Loop, GPT-4o driver | GPT-4o | matched | this work | 41.1 |
| Student alone | distilled 8B | matched | this work | 47.4 |
| Base searching, frontier reading | Opus 4.6 | matched | this work | 54.9 |
| Student searching, frontier reading | Opus 4.6 | matched | this work | 57.5 |
| Loop | Opus 4.6 | matched | this work | 57.8 |