# ECLIPSE: JOINT RISK CONTROL FOR LLM OUTPUT RECOVERY AND EVIDENCE COMPLETENESS

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

## ABSTRACT

Selecting fewer retrieved passages can reduce the amount of text needed to inspect a language model's existing output. However, reproducing that output from fewer passages does not establish that the supporting evidence remains available. We introduce Eclipse, a framework for evaluating output recovery together with evidence completeness, defined as retaining all benchmark-annotated supporting passages. Its search keeps a small initial set of retrieved passages and adds passages chosen by the likelihood of the original output. It stops when the model reproduces that output or reaches a preset size limit. We then evaluate the fixed search procedure on separate calibration samples. These provide upper confidence bounds on recovery failure across queries and missing evidence among returned sets. Certification requires both bounds to meet preset limits under the stated independence and sampling assumptions. On held-out 2Wiki queries, search with Qwen and a supervised retriever satisfies both requirements for 94.1% of queries, exceeding retrieval-order addition by 9.2 percentage points. This configuration also meets the preset limits for both calibration bounds. Across our experiments, search mainly helps when the initial passages already contain the supporting evidence; recovering the output often leaves missing evidence unresolved. Jointly checking output recovery and evidence completeness therefore helps assess smaller passage sets, while neither establishes answer correctness.

## 1 INTRODUCTION

Retrieval-augmented language models generate answers from retrieved passages (Lewis et al., 2020). These passages provide material for inspecting the generated answer (Gao et al., 2023b). Inspecting the full retrieved pool can require reading passages beyond those needed to support the answer. We therefore seek a smaller set for inspection. We study this selection after the model has answered, keeping that answer as a fixed reference. We assess the smaller set against two requirements. First, the model should still produce the original output from the smaller context, which we call *output recovery*. Second, the set should retain all supporting passages annotated for the benchmark question, which we call *evidence completeness*. Neither requirement establishes the other. A set can reproduce the output while omitting supporting passages, or retain all supporting passages yet lead the model to a different output. Checking only one requirement therefore leaves the other unverified; we evaluate both explicitly.

Existing work provides tools for examining each requirement. Rationale extraction and context attribution identify inputs that preserve or influence a prediction (Lei et al., 2016; DeYoung et al., 2020; Liu et al., 2025; Cohen-Wang et al., 2024). Citation evaluation assesses whether sources support generated statements (Gao et al., 2023b; Chuang et al., 2025; Wallat et al., 2024). We therefore assess both requirements on the same smaller passage set, keeping the original output fixed. Our question is when this procedure can be certified to meet preset limits on output-recovery failure across queries and missing support among returned contexts.

Figure 1 illustrates the distinction using two passage sets for the same question: where was Film A's director born? Passage $d_1$ identifies Mira as the director, and $d_2$ states that Mira was born in Lisbon. In (a), the model outputs Lisbon using only $d_1$, although the passage establishing the birthplace is missing. In (b), it outputs Lisbon with both $d_1$ and $d_2$, so both supporting passages remain available. An output-recovery check accepts both sets; checking evidence completeness



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.

distinguishes them. Even when both checks pass, they establish neither answer correctness nor the model's causal reliance on the retained passages.

To quantify the two requirements, we express each as a population risk. Recovery risk is the probability that the procedure fails to recover the original output within the allowed context size and returns nothing. Evidence risk is the probability that a returned context omits at least one benchmark-annotated supporting passage. The first risk is defined over all queries; the second is defined over queries for which the procedure returns a context.

We introduce Eclipse, a framework that combines context search with certification of the two risks. The name reflects how output recovery can eclipse, or obscure, the absence of supporting evidence. Its protected forward search retains a small initial set of retrieved passages. If the model does not recover the original output, it adds the passage that makes that output most likely, then checks again. It returns the first matching context or returns nothing if the size limit is reached. Search uses no support annotations; these are used to assess evidence completeness during calibration. Before calibration, we fix the model and the complete search procedure, including its size limit, as a *policy*. Separate calibration samples provide upper confidence bounds on recovery failure and evidence error (Clopper & Pearson, 1934; Angelopoulos et al., 2021). We certify the policy only if both bounds meet preset limits. Under the stated sampling assumptions, this certificate controls population risks rather than guaranteeing each returned context.

Our main studies use the Qwen and Phi-3 language models for multihop question answering: Qwen on 2Wiki and MuSiQue, and Phi-3 on MuSiQue. On the held-out 2Wiki test set, forward search with Qwen and a supervised retriever recovers the output with complete evidence on 94.1% of queries, exceeding retrieval-order addition by 9.2 percentage points. This policy also passes both calibration bounds. On a separate MuSiQue cohort, adapting the retriever improves evidence retention but yields no joint certificate. Historical MuSiQue certification claims remain conditional because independence of policy design from earlier calibration feedback is unestablished. Across our experiments, search mainly helps when the initial passages already contain complete support; recovering the output often leaves missing evidence unresolved. These findings show why smaller contexts should be assessed for both output recovery and evidence completeness. Neither check, alone or in combination, establishes that the answer is correct.

Our contributions are:

1. an empirical analysis showing why output recovery alone cannot establish evidence completeness;
2. a protocol for certifying fixed context-selection policies against both recovery failure and evidence error; and
3. an evaluation of protected forward search that identifies its benefits, computational costs, and dependence on the initial retrieved evidence.

## 2 Two requirements for a context audit

For query $q$, let $E(q)$ be the retrieved passage pool and $D_q(S)$ the model's deterministic output from context $S$. The reference is $\hat{y}_q = D_q(E(q))$. A policy fixed before calibration, denoted $\pi$, searches for a smaller set $W_\pi(q) \subseteq E(q)$ that reproduces this reference within specified size and computation limits. Output equality compares only the stripped first generated line, with lowercasing, removal of ASCII punctuation and English articles as whole words, and whitespace collapse. It is normalized equality, not exact-string equality or semantic equivalence.

A *release* is a context returned to the user after output recovery. Returning nothing is called abstention. Let $F_\pi(q)$ indicate recovery failure and $A_\pi(q)$ indicate release. A policy could apply an additional rejection rule after recovery. Our central policy has no such rule, so $A_\pi = 1 - F_\pi$. The benchmark annotations define $L_{\mathrm{ev}}(q, W_\pi) = 1$ when a released set omits any annotated supporting paragraph, and zero otherwise. The two risks are

$$R_{\mathrm{rec}}(\pi) = \Pr(F_\pi = 1), \qquad R_{\mathrm{ev}}(\pi) = \Pr(L_{\mathrm{ev}} = 1 \mid A_\pi = 1). \tag{1}$$

The evidence risk is defined only for positive release probability. Joint certification requires $R_{\mathrm{rec}} \leq \alpha$ and $R_{\mathrm{ev}} \leq \beta$, with protocol targets .10 and .20 in our studies. These targets require

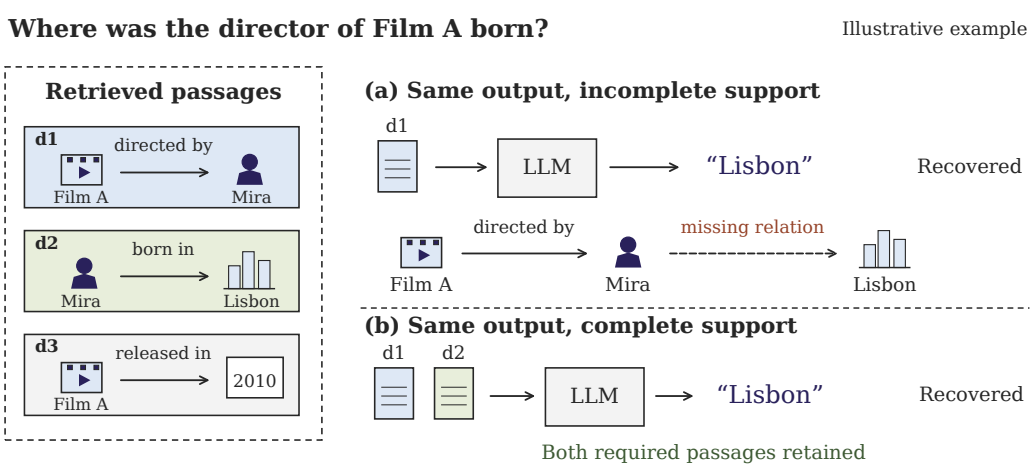


Figure 1: Output recovery does not establish evidence completeness. In this fictional example, the same answer can be generated with or without the passage linking the director to the birthplace. Both outputs match the full-context reference, but only (b) retains both required passages. This is a conceptual illustration, not a measured trajectory or an online evidence-label check.

at least 90% recovery and at most 20% evidence error among releases. They are experimental screening thresholds, not validated tolerances for human auditing; applications must set targets before calibration. We report *coverage*, the probability $c = \Pr(A_\pi = 1)$ of returning a context, and *complete matching coverage*, $\Pr(A_\pi = 1, L_{\text{ev}} = 0)$. The latter counts queries with both output recovery and complete evidence. Returned-set size and total search cost measure the associated resource burden. For the central policy, the recovery requirement is equivalently $c \geq .90$.

**Why neither requirement replaces the other.** A search can reproduce every output while returning poor evidence. Conversely, a rule can return excellent evidence on very few queries and fail recovery on the rest. Unconditional released-evidence error equals $cR_{\text{ev}}$ and can become small simply through abstention. Thus an evidence-only certificate does not establish recovery coverage, and a recovery-only certificate does not establish released-evidence quality.

**Benchmark-relative evidence.** Both main benchmarks require all annotated supporting paragraphs to be retained; extra paragraphs are allowed. MuSiQue uses `is_supporting` labels, and 2Wiki uses the official supporting facts. Throughout this paper, evidence completeness means retention of these annotations; it does not test semantic support for the model's answer. A wrong reference can be reproduced alongside complete support; neither risk certifies answer correctness or causal dependence on the passages.

**Scope of a reduced context.** Returning at most 16 of 20 passages bounds the output-set size, not the number of candidates scored or total computation. Returning the full pool would recover a deterministic reference trivially. We therefore evaluate smaller released sets while accounting for all search scoring and generation. The 2Wiki study similarly limits releases to eight of ten passages.

## 3 Context Search and Policy Certification

Eclipse first constructs contexts, then assesses the fixed policy on calibration samples. Output matching determines each return decision; certification assesses population risks.

### 3.1 Protected Forward Search

Protected forward search keeps the initial retrieved passages, even if the model could recover its output without some of them. This preserves existing support without guaranteeing that the initial set is complete.

A passage ranker selects the four starting passages. Our main configuration uses BGE-reranker-base trained with support labels, termed supervised BGE. For the 20-passage setting, the deterministic search proceeds as follows:

1. Set $S$ to these four passages and generate $D_q(S)$. If it matches $\hat{y}_q$, release $S$ immediately.
2. Otherwise, score every remaining passage $e$ by the mean token log probability of the reference's first line given $S \cup \{e\}$. Add the highest-scoring passage, with source-index tie breaking.
3. Generate on the enlarged set. Release the first matching set; otherwise repeat until size 16. If size 16 still fails, abstain.

All four starting passages remain protected. Generation determines recovery; likelihood only chooses the next candidate. No evidence-score rejection rule or supporting-fact labels are used during inference. Passages are serialized in source-index order, and prompts are not silently truncated. Figure 2 shows the complete stopping logic.

In this 20-passage setting, the maximum search uses 13 generation checks and scores 126 candidate sequences in 66 batches of at most two. Shared ranking and full-context reference generation are recorded separately. The procedure uses standard greedy search. We compare it with ordered addition, which follows retrieval rank, and recursive protected deletion, which removes passages while retaining output recovery. All are evaluated with the same evidence-completeness criterion.

### 3.2 Joint certification of fixed policies

Calibration assesses the fixed procedure across queries, using one outcome per query regardless of the number of search calls. For a frozen policy, count $x_{\mathrm{rec}}$ recovery failures among $n_{\mathrm{rec}}$ calibration queries. On an independent evidence sample, count $x_{\mathrm{ev}}$ evidence errors among the $m_{\mathrm{ev}}$ releases. Let $U(x, n; \eta)$ be the one-sided Clopper–Pearson upper bound (Clopper & Pearson, 1934) with tail $\eta$. It is the $(1-\eta)$ quantile of $\mathrm{Beta}(x+1, n-x)$ for $x < n$, and equals one when $x = n$. We denote the recovery and evidence bounds by $U_{\mathrm{rec}}$ and $U_{\mathrm{ev}}$. Our implementation requires at least 40 releases for evidence certification.

**Theorem 1** (Joint certificate for frozen policies)**.** *Suppose $M$ policies, models, prompts, and evidence rules are fixed independently of the calibration samples. Assume independent and identically distributed (IID) queries from the target distribution, with independent recovery and evidence roles. Let $\delta$ be the allowed probability of any false certificate across the $M$ policies. Use tail $\delta/(2M)$ for each of the $2M$ exact-binomial bounds. Certify a policy only when its recovery upper bound is at most $\alpha$ and its evidence upper bound is at most $\beta$. Then*

$$\Pr_{\mathcal{C}}[\exists \pi \textit{ certified}: \; R_{\mathrm{rec}}(\pi) > \alpha \;\vee\; R_{\mathrm{ev}}(\pi) > \beta] \leq \delta. \tag{2}$$

**Proof.** Recovery failures are binomial for each fixed policy. Conditional on $m_{\mathrm{ev}} = m > 0$, evidence errors are binomial with parameter $R_{\mathrm{ev}}(\pi)$. Exact coverage, averaged over positive release counts, limits the probability that an upper bound falls below its true risk to $\delta/(2M)$. A union bound over policies and risks proves the result. Rejecting candidates with fewer than 40 releases cannot increase the false-certificate probability. Policies need not be independent (Appendix A).

Here $\mathcal{C}$ denotes the calibration samples. The bound controls the probability of returning any false certificate over repeated calibration samples. It does not assign a posterior probability to a certificate already obtained. With $A_\pi = 1 - F_\pi$, recovery control also gives a lower bound on coverage. That equivalence would fail if another rule rejected contexts after recovery.

**Protocol scope.** Calibration selects neither search depths nor thresholds. The guarantee assumes a fixed pipeline and query distribution; hash-based role separation alone establishes neither IID sampling nor document independence.

## 4 Recovery, evidence retention, and certification

We first compare search procedures under both requirements, then examine computation costs, dependence on initial evidence, and performance in new cohorts.

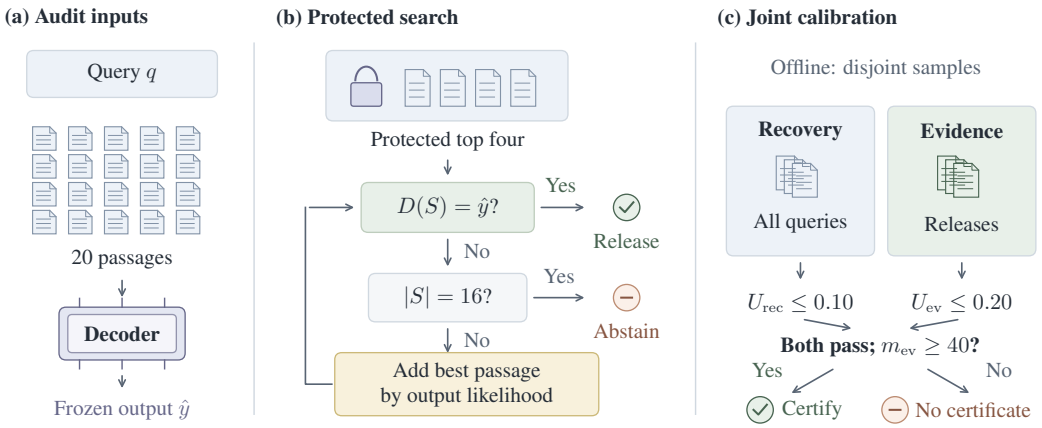


Figure 2: Eclipse in the 20-passage setting. (a) Inputs and frozen reference. (b) Protected search returns the first normalized first-line match; failure at size 16 causes abstention. Additions maximize reference-output likelihood. (c) Separate samples bound recovery failure and evidence error among releases; certification also requires 40 evidence-calibration releases. No evidence labels are used online. For 2Wiki, the pool and cap are ten and eight passages.

**Main study design.** We evaluate MuSiQue (Trivedi et al., 2022) with Qwen2.5-7B and Phi-3, and 2Wiki (Ho et al., 2020) with Qwen. Each study uses 1,000 recovery-calibration, 560 evidence-calibration, and 1,000 test queries, with eight separate engineering checks. We combine three rankers (BM25 lexical scoring, base BGE without dataset tuning, and supervised BGE) with four search rules: top-four checking, ordered addition, protected forward search, and recursive protected deletion. Top-four checks only the initial set; addition checks each successive retrieval depth. Recursive deletion starts from the full pool and subdivides a deletion block when removing it prevents recovery. All policies protect the initial four passages. Release caps are 16/20 passages for MuSiQue and 8/10 for 2Wiki. The three rankers and four searches give 12 policies, each with two bounds evaluated before test inference. For MuSiQue, historical feedback leaves policy–calibration independence unestablished; numerical passes are reported conditionally, not as independent confirmation. The newly reserved 2Wiki cohort is the primary certification study. Table 2 compares the three search policies under supervised BGE; the all-ranker admission analysis appears in Section 6.

MuSiQue reuses 2,560 historical queries after selection and training/development exclusions; roles are disjoint within each study and shared across decoders. 2Wiki uses newly reserved training-split queries after excluding historical queries and ranker-training inputs. These checks concern queries, not document or pretraining overlap (Appendix B.1). Historical diagnostic cohorts are documented separately in Appendix E.

**Joint certification and complete releases.** Supervised forward and recursive search pass both numerical bounds in the Qwen studies; addition fails recovery despite passing evidence. No policy passes both bounds in the Phi-3 study. Its supervised forward recovery-failure count is 80/1,000, but the upper bound .1076 exceeds .10. The calibration sample is therefore insufficient to certify this

Table 1: Roles within each of the three main studies. MuSiQue query roles are shared across decoders; they do not constitute independent-query replication. Historical diagnostics use separate cohorts (Appendix E).

| quantity | value per study | purpose |
|---|---|---|
| recovery calibration | 1,000 | test recovery risk |
| evidence calibration | 560 | test released-evidence risk |
| test | 1,000 | evaluate frozen policies |
| tail per risk bound | $.05/24$ | 12 policies, two risks |

Table 2: Main comparison with supervised BGE: 1,000 test queries per study. Release (rel.) and complete matching coverage (CM) use all test queries; evidence error (err.) conditions on release. All three are percentages. $U_{\mathrm{rec}}$ and $U_{\mathrm{ev}}$ are calibration upper bounds, not test estimates. A numerical pass requires bounds at most .10/.20 and at least 40 evidence-calibration releases. Each study uses tail $.05/24$ for 12 policies and two risks. The pass column records whether these numerical criteria are met. MuSiQue retains unresolved historical dependence; the theorem's assumptions are not established by these numbers. Nonpassing policies remain in the diagnostic evaluation. Seconds per query include search, ranking, and reference generation; these are recorded cumulative costs, not service latency or equal-compute budgets.

| study | policy | rel.% | CM% | err.% | $U_{\mathrm{rec}}$ | $U_{\mathrm{ev}}$ | pass | s/query |
|---|---|---|---|---|---|---|---|---|
| MuSiQue / Qwen | BGE addition | 80.6 | 77.8 | 3.47 | 0.2097 | 0.0702 | no | 1.861 |
| | Protected forward | 96.1 | 91.6 | 4.68 | 0.0476 | 0.0869 | yes | 3.524 |
| | Recursive deletion | 91.9 | 88.0 | 4.24 | 0.0706 | 0.0722 | yes | 4.491 |
| 2Wiki / Qwen | BGE addition | 85.3 | 84.9 | 0.47 | 0.1892 | 0.0218 | no | 0.661 |
| | Protected forward | 94.6 | 94.1 | 0.53 | 0.0694 | 0.0190 | yes | 0.868 |
| | Recursive deletion | 93.8 | 93.3 | 0.53 | 0.0847 | 0.0194 | yes | 1.143 |
| MuSiQue / Phi-3 | BGE addition | 72.3 | 69.8 | 3.46 | 0.3182 | 0.0690 | no | 2.454 |
| | Protected forward | 90.6 | 86.0 | 5.08 | 0.1076 | 0.0830 | no | 4.867 |
| | Recursive deletion | 89.2 | 84.8 | 4.93 | 0.1334 | 0.0720 | no | 6.131 |

policy; it does not show that population risk exceeds the target. The historical Phi-3 certificate used a different cohort and confidence allocation (Appendix E), so it does not overturn this decision.

Among test queries, forward improves complete matching coverage over addition by 13.8, 9.2, and 16.2 percentage points for MuSiQue/Qwen, 2Wiki/Qwen, and MuSiQue/Phi-3. Descriptive paired 95% bootstrap intervals are [11.5,16.2], [7.4,11.1], and [13.6,18.8]; all three comparisons pass within-study Holm correction. Against recursive deletion, differences are 3.6, 0.8, and 1.2 points. Only MuSiQue/Qwen passes the corresponding adjusted test ($p < .001$); the other two differences do not establish superiority or equivalence. Forward has lower recorded search cost than recursive deletion, but higher cost than addition. The replay below compares returns under deadlines.

**Time budgets change the preferred search.** To compare returns before search completion, a post-hoc replay includes scoring, ranking, and reference generation under the same per-query time limits (Appendix C.1). Recursive deletion may return its last verified, size-admissible set before finishing compression. At two seconds on MuSiQue/Qwen, complete matching coverage is 55.6% for forward, 65.9% for addition, and 67.2% for recursive deletion. At 16 seconds, these become 90.4%, 77.8%, and 87.6%. Forward benefits from a larger time budget, while addition and recursive deletion perform better at the shorter deadline in this replay. Replayed stopping rules have no new certificate or measured online latency guarantee.

**Size of the released audit context.** We next measure how much text remains after search. Forward releases average 5.41, 4.49, and 6.11 passages for MuSiQue/Qwen, 2Wiki/Qwen, and MuSiQue/Phi-3. Median prompt-token reductions against each query's full-context prompt are 68.2%, 36.3%, and 66.4%. These descriptive summaries condition on release, whose coverage is reported in Table 2; they include uncertified diagnostics. On queries released by both forward and recursive search, forward uses 1.52, 0.18, and 1.55 fewer passages on average. Table 17 reports all four supervised policies and paired token comparisons. Prompt tokens include instructions and the question. We measure fewer input tokens, without measuring human diagnosis time.

**Dependence on initial evidence.** To assess whether search repairs missing evidence, we grouped supervised-BGE forward trajectories by initial support completeness (Table 10). Protection necessarily preserves complete initial support; among initially incomplete releases, evidence error was about 70% for both MuSiQue decoders. Initial support was complete on 930/1,000 MuSiQue queries (93.0%) and 993/1,000 2Wiki queries (99.3%); the seven incomplete 2Wiki queries limit subgroup

Table 3: Fusion-ranker validation on 2Wiki/Qwen. Bounds are from calibration; complete matching coverage (CM) uses 1,000 test queries. Targets are .10/.20.

| policy | $U_{\mathrm{rec}}$ | $U_{\mathrm{ev}}$ | CM (%) | joint |
|---|---|---|---|---|
| Addition | .1555 | .1967 | 75.0 | no |
| Forward | .0627 | .1982 | 80.7 | yes |
| Recursive | .0804 | .1923 | 80.9 | yes |

interpretation. These annotation-based strata have no separate certificates and do not provide an online rule for detecting missing evidence.

**Separating recovery from evidence retention.** To test whether recovery can remain high when initial evidence deteriorates, a fixed-seed ablation replaces the supervised-BGE starting set with four random passages. It uses 256 historical Phi-3 test queries and keeps subsequent search unchanged. Coverage changes from 94.14% to 92.58%, while conditional evidence error rises from 5.81% to 82.28%. Complete, matching releases fall from 227 to 42. This diagnostic supports separating output recovery from evidence completeness and identifies the retrieved starting set as important in this configuration. It does not establish semantic support or model reliance on those passages.

**Intermediate retrieval-quality validation.** The near-complete initial support in the main 2Wiki study motivates a test with weaker retrieval. A separately frozen study used 256 selection queries to combine base and supervised BGE rankings by reciprocal-rank fusion with weight $w = .3$, followed by independent 1,000/560/1,000 recovery, evidence, and test roles. Nine policies used simultaneous tails $.05/18$. Initial test support completeness was 79.1%, 83.6%, and 99.6% for base BGE, fusion, and supervised BGE. Fusion forward and recursive search both passed certification (Table 3); base-BGE policies failed the evidence bound. Forward gained 5.7 percentage points over fusion addition; its comparison with recursive deletion did not establish superiority (Holm-adjusted $p = 1$). The evidence bound .1982 is close to .20. This configuration passes with less complete initial retrieval; open-domain performance and document-independent replication remain untested (Appendix C.2).

**Document-isolated finite-population validation.** To test beyond query-level exclusions, a further 2Wiki experiment retained $N = 6{,}495$ queries with disjoint candidate documents after training-document and historical-query exclusions. Both decoders shared randomly allocated 1,000/560/1,000 roles. All 24 ranker–search–decoder combinations passed simultaneous hypergeometric bounds, which account for sampling without replacement from this finite population. Supervised-BGE forward reached 99.2% complete matching coverage on both decoders (0.30%/0.20% conditional evidence error); fusion forward reached 95.2%/94.4%. Comparison questions account for 88.0% of this test set, versus 32.5% in the intermediate-retrieval study; the higher rates cannot establish improvement across the original question mix. This restricted-population extension is separate from the IID theorem (Appendices B.5 and C.4).

**Adaptive search versus simple controls.** To assess the contribution of likelihood-guided selection, we compared protected forward search with fixed top-four, fixed top-eight, and protected random addition on the same document-isolated test queries. With supervised BGE, complete matching coverage was 86.3%/80.2% for top-four, 91.9%/88.9% for top-eight, and 96.3%/95.3% for random addition (Qwen/Phi-3). Protected forward reached 99.2% for both decoders with the same initial set and candidate pool. These descriptive comparisons assess the search rule with a common starting set; the controls received no new certificate (Appendix C.5).

**Target-domain adaptation on MuSiQue.** We next tested whether adapting retrieval enables certification on MuSiQue. On 500 shared test queries from a separately frozen 1,346-query population, MuSiQue-adapted BGE raises forward complete matching over base BGE from 58.0% to 81.8% (Qwen) and 55.8% to 75.6% (Phi-3). For this configuration, test evidence error falls to 16.5%/17.6%, but the calibration bounds (.2451/.2430) exceed .20. None of the 18 evaluated policies passes joint certification. Most net gains occur where initial support becomes complete in a paired descriptive analysis. Recursive deletion retains slightly higher complete matching (82.4%/76.4%) with larger sets; adaptation does not establish cross-dataset certification (Appendix C.6).

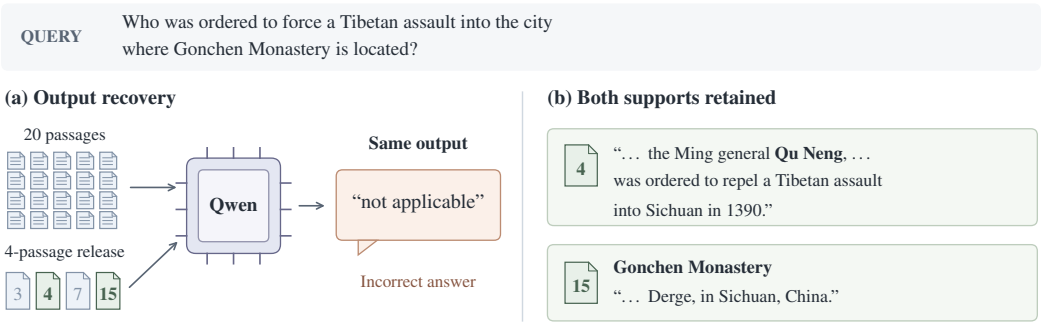


Figure 3: A real Qwen audit case, selected by the fixed hash rule in Appendix D.2. The four-passage release reproduces the incorrect reference "not applicable" while retaining both annotated support passages. Passage 15 locates the monastery in Sichuan; passage 4 names Qu Neng in an assault into Sichuan. Both are benchmark supports. This makes the discrepancy available for inspection; it does not show answer correction, causal dependence, or a measured improvement in human diagnosis. Passage indices are zero-based.

**Answer correctness and an inspection case.** A separate question is whether the recovered output is correct. Reference-answer accuracy in the main tests is 29.9% for MuSiQue/Qwen, 53.5% for 2Wiki/Qwen, and 17.3% for MuSiQue/Phi-3 (Table 19); neither certificate guarantees answer correctness. Figure 3 shows an incorrect output reproduced from four passages that retain both annotated supports.

## 5 Related work

**Output attribution and context reduction.** Rationale and attribution methods identify inputs that preserve or influence predictions (Lei et al., 2016; DeYoung et al., 2020; Ribeiro et al., 2016; Lundberg & Lee, 2017). Language-model methods accelerate ablation, learn citations, or estimate context influence (Liu et al., 2025; Chuang et al., 2025; Cohen-Wang et al., 2024). We combine output recovery with an evidence check, without claiming causal attribution. Prompt compression targets input cost and downstream performance (Jiang et al., 2023; 2024); context position can also affect model behavior (Liu et al., 2024). We instead assess retained support while keeping an existing output fixed; constructing the reduced context can be costly.

**Citation quality and factual support.** ALCE evaluates answer correctness and citation quality (Gao et al., 2023b), while RARR retrieves evidence and revises unsupported content (Gao et al., 2023a). RAGTruth annotates hallucination spans (Niu et al., 2024); RAGAs, ARES, and RAGChecker assess retrieval and generation quality (Es et al., 2024; Saad-Falcon et al., 2024; Ru et al., 2024). Related work examines abstention under insufficient evidence (Peng et al., 2025) and disagreement between citation-faithfulness measures (Zhang et al., 2024). We assess retention of every annotated supporting paragraph, even for incorrect outputs. Evidence criteria differ by benchmark: FEVER labels claims and support (Thorne et al., 2018), whereas MuSiQue and 2Wiki annotate multihop support (Trivedi et al., 2022; Ho et al., 2020). Our criterion can reject alternative valid evidence and does not measure factual correctness.

**Risk control and selective prediction.** Selective prediction studies coverage and errors among accepted predictions; SelectiveNet learns prediction and rejection jointly (Geifman & El-Yaniv, 2019). We assess coverage and evidence error for fixed context-selection policies. Conformal and related methods control risks in retrieval-augmented generation (RAG), structured outputs, reasoning prefixes, and temporal events (Angelopoulos et al., 2024; Li et al., 2024; Kotte, 2026; Cheung et al., 2026; Hu et al., 2026). Learn then Test already supports multiple risks and selective prediction (Angelopoulos et al., 2021). We use this established perspective with exact binomial bounds (Clopper & Pearson, 1934) to assess the two requirements of context auditing. Extensions beyond exchange-

ability address other sampling conditions (Barber et al., 2023); our certificate instead assumes IID calibration and independently fixed policies, an unresolved premise for historical MuSiQue.

## 6 When to use the dual audit

**Separate requirements change policy selection.** We compared recovery-only, evidence-only, and joint policy selection (*admission*) on the same 12 policies. Each rule used 95% familywise confidence, with tails $.05/12$ for single-risk and $.05/24$ for joint bounds. Targets (.10/.20), the 40-release minimum, and policy outputs were unchanged. This post-hoc reanalysis reproduced the original admission sets (Appendix C.9).

On 2Wiki, recovery-only admission retained six policies; four failed the evidence bound and had 20.2–33.4% test evidence error. Evidence-only admission retained four; two failed the recovery bound and had 14.7% and 32.7% test recovery failure. Joint admission retained supervised-BGE forward and recursive search, with 93.8–94.6% coverage and 0.53% evidence error each. Each exclusion arose from the omitted risk, not confidence allocation alone. Joint assessment changes policy eligibility, not the outputs themselves.

For historical MuSiQue, independence between policy design and calibration remains unestablished. Failure to certify does not prove excessive risk: supervised Phi-3 forward failed despite 9.4% test recovery failure.

**Choosing a search under resource limits.** Forward search improves complete matching over addition at greater cost; recursive deletion can return verified contexts sooner under short deadlines. The budget and returned size therefore also guide policy choice.

**What retained evidence permits.** Search helps most with complete initial support, without reliably repairing missing evidence. With dataset-supervised retrieval and 83.6% initial completeness, fusion passes certification, but its evidence-risk bound is close to the target. Complete contexts can retain irrelevant passages; answer correction and faster human diagnosis remain untested (Appendix D.2).

**Where the certificate applies.** Hash splitting does not establish IID sampling or independently frozen policies. Each study has its own confidence family; the document-isolated guarantee covers only its selected finite population. Document exclusion does not establish pretraining independence; shared decoder roles do not provide independent-query replication. Open-domain retrieval remains untested. Changed policies or distributions need new justification.

## 7 Conclusion

Eclipse jointly assesses output recovery and annotated evidence completeness. The reserved 2Wiki studies support policy certification under the stated sampling conditions. MuSiQue adaptation improves evidence retention without a joint certificate; historical results remain conditional. Search mainly helps when initial support is complete, with benefits depending on the budget and comparator. Neither requirement establishes answer correctness or causal attribution.

## Reproducibility Statement

Appendix B details the experimental protocol and reproduction requirements. The CPU supplement verifies reported results from saved records; it does not rerun training or inference. Trained ranker weights are not included. Appendix B.6 reports a separate author-run inference check.

## AI Use Statement

During the preliminary literature search, generative AI was used to help organize Chinese and English search keywords. It was also used to polish the language of the final manuscript. The authors are responsible for the final manuscript.

## REFERENCES


Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. URL https://arxiv.org/abs/2110.01052v4.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

Matt Y. Cheung, Ashok Veeraraghavan, Hanjie Chen, and Guha Balakrishnan. Conformal certification of reasoning trace prefixes. *arXiv preprint arXiv:2605.30085*, 2026. arXiv:2605.30085.

Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, and Wen-tau Yih. SelfCite: Self-supervised alignment for context attribution in large language models. In *International Conference on Machine Learning (ICML)*, 2025.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. ContextCite: Attributing model generation to context, 2024.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, 2024. doi: 10.18653/v1/2024.eacl-demo.16.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023a. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023b. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

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. doi: 10.18653/v1/2020.coling-main.580.

Cheng Hu, Cong Cao, Fangfang Yuan, Diandian Guo, Pin Xu, Yu Liu, and Yanbing Liu. Conformal event prediction with temporal knowledge graph. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5233–5248. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.258. URL https://aclanthology.org/2026.findings-acl.258/.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, 2023. doi: 10.18653/v1/2023.emnlp-main.825.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 1658–1677, 2024. doi: 10.18653/v1/2024.acl-long.91.

Varun Kotte. When can conformal risk control certify LLM outputs? bounds, impossibility, and adaptation for structured generation. *arXiv preprint arXiv:2606.29054*, 2026. arXiv:2606.29054.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, 2016. doi: 10.18653/v1/D16-1011.

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, 2020.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2024. doi: 10.18653/v1/2024.naacl-long.210.

Fengyuan Liu, Nikhil Kandpal, and Colin Raffel. AttriBoT: A bag of tricks for efficiently approximating leave-one-out context attribution. In *International Conference on Learning Representations (ICLR)*, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774, 2017.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10868–10882, 2024. doi: 10.18653/v1/2024.acl-long.585.

Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. Unanswerability evaluation for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025. doi: 10.18653/v1/2025.acl-long.415.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778.

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*, volume 37, 2024. doi: 10.52202/079017-0692.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024*

*Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, 2024. doi: 10.18653/v1/2024.naacl-long.20.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. Correctness is not faithfulness in RAG attributions, 2024.

Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-hong Huang, and Evangelos Kanoulas. Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics. In *Proceedings of the 17th International Natural Language Generation Conference*, 2024. doi: 10.18653/v1/2024.inlg-main.35.

# A STATISTICAL GUARANTEE AND SCOPE

This section proves the IID certificate in Theorem 1. The finite-population extension uses a different sampling model and is derived in Appendix B.5. Fix all $M$ policies, model interfaces and evidence rules before calibration. For policy $\pi$, let $F_\pi$ be its recovery-failure indicator and $A_\pi$ its release indicator. Write $r_\pi = \Pr(F_\pi = 1)$ and $e_\pi = \Pr(L_{\text{ev}} = 1 \mid A_\pi = 1)$ when $\Pr(A_\pi = 1) > 0$. These are population risks for the eligible-query distribution, not risks conditional on a particular test question.

**Proof of Theorem 1.** For IID recovery queries, $X_r \sim \text{Binomial}(n_r, r_\pi)$. The one-sided exact upper bound satisfies $\Pr\{U(X_r, n_r; \eta) < r_\pi\} \le \eta$. On the independent evidence sample, condition on the release indicators. The released queries are independent draws from the distribution conditional on $A_\pi = 1$. Thus, conditional on their count $N_e = m > 0$, $X_e \sim \text{Binomial}(m, e_\pi)$, and

$$\Pr\{N_e > 0,\ U(X_e, N_e; \eta) < e_\pi\} \le \sum_{m>0} \Pr(N_e = m)\eta \le \eta.$$

If release probability is zero, positive release counts occur with probability zero and the rule receives no evidence certificate. The additional requirement $N_e \ge 40$ can only remove certificates. A certified policy with $r_\pi > \alpha$ or $e_\pi > \beta$ must have at least one upper bound below its true risk. Taking $\eta = \delta/(2M)$ and a union bound over the $2M$ bounds proves the claim. Policies may share calibration questions; independence between policies is unnecessary. The probability concerns repeated calibration samples, not a posterior confidence conditional on having observed a passing certificate.

**Allocations actually used.** The historical studies use different policy families from the main studies. The initial Phi-3 validation evaluates one prespecified forward policy with two tails .025. Its comparison policies are descriptive. The later Qwen study evaluates three prespecified policies with six tails $.05/6$. The 768-query Phi-3 comparison does not supply new calibration certificates. Each of the three main studies covers 12 policies with 24 tails $.05/24$, using 1,000 recovery and 560 evidence-calibration queries. These guarantees are not simultaneous over every historical experiment, dataset, endpoint, or subsequently chosen model.

Hash-based role separation does not establish IID or eliminate document dependence and distribution shift. A population certificate does not certify every subgroup. The descriptive strata are not separately calibrated.

# B REPRODUCIBLE EXPERIMENTAL PROTOCOL

## B.1 PROTOCOL FOR THE THREE MAIN STUDIES

Each study used 8 engineering checks, 1,000 recovery-calibration queries, 560 evidence-calibration queries, and 1,000 test queries. MuSiQue reused 2,560 queries from an earlier study of retrieval-order prefixes: 480 previously assigned to each calibration role and 1,600 to testing. Earlier test outputs were unopened after failed evidence certification; earlier calibration outcomes had been observed. Its 640 selection queries were excluded. The eight MuSiQue engineering checks came from a separate historical smoke set and are outside the 2,560 statistical queries. Source IDs and normalized questions were checked against enumerated ranker training, development, and current validation cohorts. Hash ordering reassigned the 2,560 queries to the new roles. Phi-3 reused the exact Qwen role inputs, with new references and calibration. This reuse is not untouched independent confirmation. The recorded protocol explicitly states that the design was fixed before the new outputs, but not independent of all historical research decisions. Therefore, the independent-policy premise is not established for MuSiQue. We report its bounds as a conditional, retrospective application of the rule; 2Wiki is the primary new-reserve certification study.

For 2Wiki, 6,684 of 167,454 training queries overlapped enumerated history, including all 4,096 supervised-ranker training queries. Fixed hash ordering selected 2,568 eligible queries without question-type quotas or outcome filtering; normalized questions were unique. Source and label files were separate. No evaluation-query support label entered ranking, scoring, or generation as

an inference-time input. Both datasets required retaining every annotated supporting paragraph; additional passages were allowed.

BM25, base BGE, and dataset-supervised BGE each supplied four protected passages for top-four checking, ordered addition, forward search, and recursive deletion. Caps were 8/10 passages for 2Wiki and 16/20 for MuSiQue. Recursive deletion tried reversed-relevance blocks of at most four passages, subdividing rejected blocks in depth-first order with rollback. Maximum generation checks were 10 and 28, respectively. Native chat, greedy 64-token generation, normalization, and source-index presentation use the prompt and output rules below. Candidate scoring uses batches of two and explicit left-padding position IDs. Phi-3 end-of-sequence (EOS) token IDs are 32000, 32001, and 32007; Qwen EOS IDs are 151645 and 151643. Both use bfloat16 (bf16) precision, one beam, and no silent prompt truncation. References were regenerated per study. Twelve arms ran in hash-randomized order per query on GPUs 3, 4, 6, and 7.

Each study used tail $.05/24$ for its 24 bounds, targets .10/.20, and at least 40 evidence-calibration releases. No threshold was selected. Certification preceded test inference; uncertified policies were still tested diagnostically. Per-study bounds do not imply simultaneous 95% coverage over all studies. IID sampling and independently fixed policies remain assumptions, not consequences of provenance checks. Six comparisons per study paired forward with addition and recursive deletion under each ranker. Two-sided exact tests use discordant pairs: queries where only one compared policy returns complete matching evidence. Holm correction covers all six comparisons within a study. Paired bootstrap intervals used 10,000 query-level resamples; they are descriptive and not simultaneous. Shared queries across decoders are not independent replications. Main-table costs include ranking, reference generation, and search, including candidate scoring, but exclude loading and scheduling. Hardware contention limits latency interpretation.

Table 4: Main and historical study roles. Numbers count queries. Reused roles across decoders are not independent-query replications. Historical cohorts support only the explicitly identified supplementary comparisons, ablation, and cases.

| study | interface / token cap | calibration | evaluation |
|---|---|---|---|
| Main studies (each) | native chat / 64 | 1,000 + 560 | 1,000 |
| Initial Phi-3 | native chat / 64 | 1,200 + 600 | 1,000 |
| Phi-3 stronger controls | native chat / 64 | no new certificate | 768 |
| Qwen transfer | native chat / 64 | same 1,200 + 600 IDs | same 768 IDs |

### B.2 Ranker training and data separation

The supervised ranker determines which four passages the search protects. The following recipe describes the main-study rankers; the later MuSiQue adaptation is specified in Appendix C.6. The main ranker starts from `BAAI/bge-reranker-base`, an XLM-RoBERTa cross-encoder with 278,044,417 parameters and one output logit. The retained base revision is `2cfc18c9415c912f9d8155881c133215df768a70`. MuSiQue supplies 4,096 training queries, selected with proportional two-, three-, and four-hop quotas using largest remainders. Within each stratum, source IDs are sorted by SHA-256 of `20260831:train:hop:<hop>:<source-id>`. Eligibility requires unique nonempty normalized questions, 20 nonempty indexed paragraphs, at least one support paragraph, and two to four decomposition steps. Labels use `is_supporting`. This stratification applies to ranker training, not central calibration.

Hash partitioning with seed 20260902 assigns 3,277 queries to fitting and 819 to epoch selection; the first 819 hashes of `<partition-seed>:partition:<qid>` form validation. All paragraphs are used without negative subsampling or class weighting. Tokenizer pairs are the question and `<title> [SEP] <paragraph text>`, with special tokens, longest-sequence truncation to 256 combined tokens, and dynamic batch padding. Training uses binary cross-entropy, AdamW with learning rate $2 \times 10^{-5}$, batch size 64, weight decay .01 except biases and layer norms, betas $(.9, .999)$, epsilon $10^{-8}$, gradient clipping 1, and bf16 autocasting. All parameters are updated over three planned epochs, with 10% linear warmup and linear decay. Python, NumPy, Torch and batch ordering use seed 20260903.

Epoch selection first maximizes the fraction of queries with complete support in the top 4 passages. Ties are resolved by complete support at 3 and 2 passages, then any support at 1; remaining ties favor earlier epochs. The selected three-epoch run is refitted from base weights on all 4,096 queries (81,920 pairs per epoch). All 4,096, including the 819 selection queries, are excluded from calibration and test. Inference uses the same pair format, 256-token limit, batches of 128, and raw logits, with smaller source index breaking ties. Evaluation labels are absent during scoring.

The 2Wiki main study reuses a dataset-specific ranker. It was trained on 4,096 training IDs sorted by `20260822:train:<source-id>`, title-based supporting-fact labels, partition/training seeds 20260823/20260824, and epoch objectives complete recall@4, @2, then any@1. The remaining recipe is the same; the selected three epochs are refitted on 40,960 paragraph pairs per epoch.

Training manifests record A100-SXM4-80GB hardware. The later execution environment records A800-SXM4-80GB, Python 3.10.20, PyTorch 2.4.1/CUDA 12.1, Transformers 4.40.2, Accelerate 1.0.1, FlashAttention 2.6.3, NumPy 1.26.4, and SciPy 1.15.3. This capture is not a contemporaneous training environment. Weight and input hashes are retained in the artifact manifests. Trained ranker weights and all historical inputs are not redistributed; replay using recorded BGE orders is therefore distinct from end-to-end retraining.

### B.3 Prompt, serialization, and output comparison

The literal two-example prompt prefix is:

```
Answer the question using the context paragraphs.  Give only
the short answer.

Context:
[1] Nile .  The Nile is a major north-flowing river in
northeastern Africa.  It is the longest river in Africa.
[2] Africa .  Africa is the world's second-largest
continent.
Question:  On which continent is the longest river located?
Answer:  Africa

Context:
[1] Mercury (planet) .  Mercury is the smallest planet in
the Solar System.
[2] Mars .  Mars is the second-smallest planet in the Solar
System.
Question:  Are Mercury and Mars the same size?
Answer:  no
```

After a blank line, append `Context:`, selected paragraph lines, `Question: <question>`, and `Answer:`, each on its own line. Each paragraph is `[j] <title> . <text>`; wrapped paragraph lines above are joined. Subsets are sorted by original source index and renumbered from one. The empty context is `(none)` and retains demonstrations. All reported search studies wrap this text in the model's native user-message chat template with an assistant generation prefix.

Central generation is greedy, one beam, at most 64 new tokens, with no silent prompt truncation. Per-model EOS and scoring details appear in Appendix B.1. After removing special tokens, comparison retains and strips the first line, lowercases it, removes ASCII punctuation and whole-word English articles (a, an, the), and collapses whitespace. This defines behavioral recovery, not exact-string equality or semantic correctness. The central likelihood score is the mean log probability of the reference's first line; actual generation, not its likelihood, decides release.

### B.4 Evidence criterion and reproducibility scope

Both main benchmarks require retaining every annotated supporting paragraph; additional passages are allowed. This criterion does not test whether the released context semantically entails the model's reference output.

Query provenance, eligibility, role counts, policies, and confidence allocation for the main experiments appear in Appendix B.1. The accompanying CPU supplement checks all 72 calibration bounds, 36 policy summaries, admission decisions, and 1,440 time-budget cells. It uses exported counts and reduced per-query checkpoints, not model inference. Historical protocols needed for the retained diagnostics appear in Appendix E. Model weights and paragraph text are not included in the CPU package. A reconstruction utility uses public benchmark files and an exported source-ID/role manifest to verify all 7,680 statistical query-role inputs against canonical content hashes. Model-file hashes identify checkpoint dependencies. Released-set prompt-token counts come from saved generation checks whose output-file hashes were verified before export. The package separately recomputes release burden and common-release comparisons. These checks do not rerun ranking or generation. Independent end-to-end reproduction remains unestablished.

### B.5 Document-isolated finite-population experiment

We constructed the document-isolated population from 2Wiki training data. Exclusions covered supervised-training candidate titles, duplicate normalized questions, and query IDs in the enumerated local input history (12,076 historical queries excluded). Remaining questions were processed in an outcome-blind hash order. A question was retained only if its ten candidate titles and paragraph-text hashes did not overlap the training reserve or an already retained question. Normalization lowercases text, replaces non-alphanumeric ASCII characters with spaces, and collapses whitespace. It may merge distinct same-title documents and miss semantic paraphrases. Historical evaluation documents were not globally excluded; historical exclusion concerned query IDs and normalized questions. The resulting finite population contained $N = 6{,}495$ queries; eight further queries were held out for engineering smoke checks. This construction reduces overlap within the retained corpus but does not establish independence from model pretraining or guarantee future open-domain behavior.

After freezing the population, a random seed permuted it and assigned disjoint roles of 1,000 recovery calibration, 560 evidence calibration, and 1,000 test queries. Qwen2.5-7B and Phi-3 were each evaluated with BM25, base BGE, fixed-weight reciprocal-rank fusion ($w = .3$), and the dataset-supervised BGE. Each ranker used ordered addition, protected forward search, and recursive protected deletion with ten candidates, four protected initial passages, and an eight-passage cap. The fusion weight was fixed from an earlier independent selection study and was not tuned on these roles. Evidence labels were excluded from ranking and search decisions; the full-context reference output was available to the recovery procedure.

Because sampling was without replacement from the frozen finite population, calibration used exact hypergeometric inversion rather than an IID binomial bound. The family tail was $0.05/48$, covering 24 combinations and two risks. Recovery failure was bounded over the finite population directly. For evidence, the sample contains a random number of releases. Because the population release count is unknown, the bound takes the largest error fraction compatible with any feasible value of that count. A combination passed only when the upper recovery failure was at most .10, the upper conditional evidence error at most .20, and at least 40 evidence releases were observed. Small-population exhaustive checks of the implementation passed; these statements remain limited to the frozen population.

For a precise construction, let $H(x; N, K, n)$ be the probability of at most $x$ marked items in $n$ draws without replacement from $N$ items, of which $K$ are marked. Set $\eta = .05/48$. Given $x$ recovery failures in $n = 1{,}000$ draws, set

$$U_{\mathrm{rec}} = N^{-1} \max\{K \in [x, N - n + x] : H(x; N, K, n) \geq \eta\}.$$

For the evidence bound, let $R$ be the population release count and $E$ the number of those releases with missing support. Observing $e$ errors among $m$ releases in $n_e = 560$ draws gives

$$U_{\mathrm{ev}} = \max_{m \leq R \leq N - n_e + m} \frac{\max\{E \in [e, R - m + e] : H(e; R, E, m) \geq \eta\}}{R}.$$

Zero observed releases return the conservative bound one and fail admission. For a fixed policy, the random role is a simple random sample without replacement. Conditional on $m$ and the true population release count $R$, errors follow $\mathrm{Hypergeom}(R, E, m)$. Inverting its lower tail gives one-sided coverage; maximizing over feasible $R$ preserves it. A union bound over all 48 events then gives

simultaneous coverage at least .95, without requiring independence between roles or decoders. This assumes fixed, repeatable query-level policy outcomes. The target includes sampled and unsampled members of the frozen population; it is neither a bound specifically on the held-out test subset nor on an unrestricted future-query distribution. The .95 family is separate from earlier experimental families. Resource scheduling changed from sequential four-GPU execution to concurrent Qwen/Phi execution on GPUs 0–3/4–7; inputs, policies, and statistical allocation were unchanged.

### B.6 Same-environment inference rerun

We selected 24 historical test queries per main study by the smallest SHA-256 values of `rerun-v1:<source-id>`, without filtering outcomes. This gives 72 query instances, not 72 necessarily distinct questions across decoders. Public benchmark files reconstructed their inputs. Using the fixed checkpoints, we recomputed supervised-BGE rankings, full-context references, candidate likelihoods, and every generation check for addition, forward search, and recursive deletion. Historical rankings and outputs were not supplied to the inference procedure; the comparison ran after collection. Interrupted executions retained complete records and resumed unfinished queries without changing the methods.

All 72 rankings and reference token sequences matched the saved originals. Across all 216 policy runs, candidate-set/hit trajectories, generated token sequences, final released sets, and annotation-completeness outcomes matched. Table 5 reports all three studies. This author-run check uses the same software environment and checkpoints, with execution across several GPUs. It supports implementation repeatability, not independent training, third-party reproduction, or calibration independence. No new population certificate is inferred from these previously evaluated queries.

Table 5: Same-environment rerun. Entries are matching/total query instances or policy runs. Trajectory equality compares candidate sets and match decisions; token equality compares every generated sequence along the search.

| study | ranking/reference | trajectories | step tokens | released sets |
|---|---|---|---|---|
| MuSiQue/Qwen | 24/24 | 72/72 | 72/72 | 72/72 |
| 2Wiki/Qwen | 24/24 | 72/72 | 72/72 | 72/72 |
| MuSiQue/Phi-3 | 24/24 | 72/72 | 72/72 | 72/72 |

### B.7 Descriptive fixed-context controls

The fixed-context controls in Appendix C.5 were frozen before control outputs were read. They are descriptive test comparisons only: no recovery or evidence calibration was allocated to them, and no original certificate was transferred. All controls protect the ranker top four. Fixed-four and fixed-eight check one set each; protected-random forward checks the protected four and then adds remaining passages in a fixed hash order through size eight.

### B.8 Reproducibility materials and their scope

The reproducibility materials separate numerical checks from model execution. Appendix B.1 specifies query roles and provenance; Appendix B gives training, prompts, model interfaces, and evidence criteria. Historical inputs used for ablations and inspection cases are documented in Appendix E.

The CPU supplement recomputes 72 main-study calibration bounds, 36 policy summaries, and three admission rules from exported counts and per-query metrics. It reconstructs time-budget curves from reduced checkpoints, including scoring time, and checks unlimited-budget counts. Separate modules verify 18 bounds, nine policies, and six paired comparisons for intermediate retrieval; 48 bounds and 24 summaries for document isolation; and 18 summaries with 72 bounds across three admission rules for MuSiQue adaptation. The supplements also include fixed-context controls, initial-evidence strata, and returned-context size comparisons (Appendices C–D).

These checks use recorded outcomes and do not rerun inference or establish IID sampling, historical freezing, or calibration independence. Original scripts retain local data and model dependencies.

Benchmark inputs, model checkpoints, trained ranker weights, and historical training data remain necessary for end-to-end reproduction; trained weights are not distributed here.

An author-run inference check separately reconstructed 24 historical queries per main study in the same environment. Recomputed rankings, all 72 reference token sequences, and all 216 search trajectories and returned sets matched (Appendix B.6). This supports implementation repeatability, without supplying new calibration evidence or third-party replication.

# C Main-study costs and operating range

This section reports the numerical details behind the main comparisons: search costs, retrieval conditions, new evaluation populations, and policy admission. Each study retains its own calibration family and target population.

## C.1 Descriptive time-budget replay

To compare searches under a common deadline, we replayed their saved call timings and returned sets. We replayed all 36,000 test policy schedules and checked that unlimited budgets reproduced the original counts. The post-hoc grid was 0.25, 0.5, 1, 2, 4, 8, 16, 32, and 64 seconds per query. Elapsed search timestamps include all preceding scoring and generation. Table 6 also charges ranking and full-context reference generation. At a deadline, the replay returns the last completed, matching set within the size cap, if any. Rejected recursive removals leave that set unchanged; unfinished calls cannot contribute an output. This is an anytime variant of the stopping rule, not a newly calibrated policy. A separate replay requiring original-policy completion and search-only accounting is retained in the analysis artifacts for all rankers and policies. Equal time caps do not imply equal average compute. Historical GPU contention affects the measurements, and replay does not simulate actual preemption, model loading, or scheduling. No test-derived budget was selected for deployment.

Table 6: Complete matching test coverage (%) under equal recorded time caps, supervised BGE. Each study has 1,000 queries. A/F/R denote addition, forward, and recursive deletion. All budget points are shown; these are descriptive outputs, including uncertified policies.

| seconds | MuSiQue/Qwen A/F/R | 2Wiki/Qwen A/F/R | MuSiQue/Phi-3 A/F/R |
|---|---|---|---|
| 0.25 | 0.0/0.0/0.0 | 32.8/32.6/31.3 | 0.3/0.4/0.0 |
| 0.5 | 27.4/26.5/16.4 | 61.6/57.4/65.3 | 18.3/17.8/21.0 |
| 1 | 53.6/41.1/60.9 | 77.8/76.2/81.6 | 41.7/31.3/52.3 |
| 2 | 65.9/55.6/67.2 | 83.0/88.4/89.5 | 56.2/47.6/59.4 |
| 4 | 75.9/72.9/79.1 | 84.7/92.7/92.5 | 65.9/64.0/70.7 |
| 8 | 77.6/84.2/85.9 | 84.9/94.0/93.1 | 68.7/75.9/79.9 |
| 16 | 77.8/90.4/87.6 | 84.9/94.1/93.3 | 69.8/83.5/83.7 |
| 32 | 77.8/91.5/87.9 | 84.9/94.1/93.3 | 69.8/85.9/84.8 |
| 64 | 77.8/91.6/88.0 | 84.9/94.1/93.3 | 69.8/86.0/84.8 |
| unlimited | 77.8/91.6/88.0 | 84.9/94.1/93.3 | 69.8/86.0/84.8 |

## C.2 Intermediate retrieval-quality range validation

This study tests certification when initial retrieval is less complete than in the main supervised-BGE setting. It combines base and supervised BGE rankings, selecting the combination before calibration. The separately frozen 2Wiki/Qwen study used 256 selection queries, eight engineering checks, 1,000 recovery-calibration queries, 560 evidence-calibration queries, and 1,000 test queries. Selection evaluated fusion weights $w = 0, .1, \ldots, 1$, requiring an initial complete-support rate in [.80,.90]. Among eligible weights, it chose the rate closest to .85, breaking ties toward smaller weights; this selected $w = .3$. The selected rate was 216/256; this is selection performance, not test performance. Let $r_{\text{base}}$ and $r_{\text{sup}}$ denote passage ranks under base and supervised BGE, starting at one. Fusion uses $(1 - w)/(60 + r_{\text{base}}) + w/(60 + r_{\text{sup}})$, with source-index tie breaking. The calibration family contained nine policies (three starting rankers crossed with addition, protected forward, and recursive protected deletion); each risk bound used tail $.05/18$. Selection preceded

calibration, calibration preceded test inference, and failed policies remain in the table as diagnostic results. The 2,824 queries were hash-selected after excluding 9,252 historical query IDs and duplicate normalized questions from the 2Wiki training pool. No hop quotas or outcome filtering were used. All policies use native-chat Qwen2.5-7B, the frozen prompt, greedy 64-token generation, four protected passages and an eight-of-ten release cap. Both BGE models and their training recipe are unchanged. The query split and model files were content-hashed; query-level exclusion does not establish document independence or IID beyond the stated sampling model.

Table 7: Intermediate retrieval-quality range validation on 1,000 2Wiki/Qwen test queries. Release and complete matching coverage (CM) use all queries; evidence error conditions on release. All three are percentages. Bounds use tail $.05/18$ and targets $.10/.20$. The fusion ranker was selected before calibration and testing.

| ranker | policy | release | CM | error | $U_{\mathrm{rec}}$ | $U_{\mathrm{ev}}$ | pass |
|---|---|---|---|---|---|---|---|
| base BGE | addition | 89.2 | 72.1 | 19.2 | .162 | .238 | no |
| | forward | 95.8 | 77.8 | 18.8 | .070 | .241 | no |
| | recursive | 94.9 | 77.5 | 18.3 | .082 | .235 | no |
| fusion ($w = .3$) | addition | 88.5 | 75.0 | 15.3 | .155 | .197 | no |
| | forward | 95.5 | 80.7 | 15.5 | .063 | .198 | yes |
| | recursive | 95.0 | 80.9 | 14.8 | .080 | .192 | yes |
| supervised BGE | addition | 87.7 | 87.3 | 0.46 | .171 | .016 | no |
| | forward | 95.4 | 95.0 | 0.42 | .071 | .015 | yes |
| | recursive | 94.3 | 93.9 | 0.42 | .078 | .015 | yes |

The fusion condition retains joint certification but loses 13–15 percentage points of complete matching relative to supervised BGE. Its evidence upper bounds are close to the .20 target. Certification therefore holds for this configuration with little margin; it does not establish a minimum retrieval quality sufficient for certification. The experiment uses the fixed ten-passage 2Wiki candidate pool and does not test open-domain retrieval.

**Counts, costs, and paired comparisons.** Test initial completeness was 791/1,000, 836/1,000 and 996/1,000 for base BGE, fusion and supervised rankers. Fusion uses both component rankers; recorded total costs below include both rankings, reference generation and all search scoring. They reflect shared-hardware execution, not controlled online latency. Sizes condition on release. The six pre-specified two-sided discordant-pair tests use one Holm family; intervals are descriptive query-level percentile bootstrap intervals (10,000 resamples, analysis seed 20260923). The analysis seed was set during this post-execution numerical recount.

Table 8: Calibration numerators/denominators and recorded test costs for the nine range-validation policies. Recovery is failures/queries; evidence is errors/releases. Mean size counts passages among test releases.

| ranker | policy | recovery | evidence | s/query | mean size |
|---|---|---|---|---|---|
| base BGE | addition | 130/1000 | 90/488 | 1.245 | 4.42 |
| base BGE | forward | 48/1000 | 101/532 | 1.507 | 4.42 |
| base BGE | recursive | 58/1000 | 97/526 | 2.110 | 4.55 |
| fusion | addition | 124/1000 | 71/483 | 1.224 | 4.39 |
| fusion | forward | 42/1000 | 80/531 | 1.508 | 4.39 |
| fusion | recursive | 57/1000 | 77/530 | 2.095 | 4.55 |
| supervised | addition | 138/1000 | 1/488 | 1.150 | 4.34 |
| supervised | forward | 49/1000 | 1/534 | 1.437 | 4.38 |
| supervised | recursive | 55/1000 | 1/527 | 2.013 | 4.53 |

Table 9: Forward minus comparator in complete matching coverage on the same 1,000 queries. Differences and intervals are percentage points.

| ranker | comparator | wins/losses | difference | 95% interval | adjusted $p$ |
|---|---|---|---|---|---|
| base BGE | addition | 68/11 | 5.7 | [4.0, 7.4] | 1.77e-10 |
| base BGE | recursive | 11/8 | 0.3 | [-0.5, 1.2] | 1 |
| fusion | addition | 68/11 | 5.7 | [4.0, 7.4] | 1.77e-10 |
| fusion | recursive | 10/12 | -0.2 | [-1.1, 0.7] | 1 |
| supervised | addition | 80/3 | 7.7 | [6.0, 9.4] | 1.18e-19 |
| supervised | recursive | 18/7 | 1.1 | [0.1, 2.1] | 0.13 |

For fusion, forward exceeds addition by 5.7 points [4.0,7.4] but differs from recursive deletion by $-0.2$ points $[-1.1,0.7]$, adjusted $p = 1$. This does not establish equivalence. The observed supervised-versus-fusion contrast is a paired retrieval-condition comparison within one cohort, not an estimate of robustness to arbitrary retrieval degradation.

**Artifact and scope.** A separate CPU supplement recomputes 18 bounds, nine policy summaries, 6 paired tests and their bootstrap intervals from exported per-query metrics. The export was recounted from saved outputs and support indices; checks verified role disjointness and the recorded execution settings, excluding model-weight checks. Local file timestamps place selection before calibration and the certificate before test outputs, but cannot independently establish historical freezing or IID. This numerical verification did not rerun inference, rehash model weights or reconstruct all historical exclusions. The guarantee covers this nine-policy family only; it is not simultaneous with the three main-study certificates.

### C.3 Initial evidence completeness in the three main studies

We next return to the original three studies and group queries by whether the protected starting set already contains all annotated support.

Table 10: Descriptive operating-range analysis for supervised-BGE protected forward search. Initial support groups queries by completeness before search. Release and complete matching coverage (CM) use all queries in each stratum; evidence error conditions on release. Entries other than $n$ are proportions. These strata are not separately calibrated.

| dataset / decoder | initial support | $n$ | release | CM | evidence error |
|---|---|---|---|---|---|
| MuSiQue / Qwen | complete | 930 | .965 | .965 | .000 |
| | incomplete | 70 | .914 | .271 | .703 |
| 2Wiki / Qwen | complete | 993 | .946 | .946 | .000 |
| | incomplete | 7 | 1.000 | .286 | .714 |
| MuSiQue / Phi-3 | complete | 930 | .906 | .906 | .000 |
| | incomplete | 70 | .900 | .243 | .730 |

The table is descriptive: because the initial set is protected, initially complete cases cannot lose annotated support. The high error among initially incomplete releases shows why output recovery should not be interpreted as evidence repair.

### C.4 Document-isolated validation results

The population and sampling procedure are defined in Appendix B.5. All 24 combinations passed both finite-population bounds. Table 11 reports every combination, including addition and base BGE. Both decoders share the same question roles; this is not an independent-question replication.

Table 11: Document-isolated 2Wiki: 1,000 test queries per decoder. Release (rel.), complete matching coverage (CM), and conditional evidence error (err.) are test percentages. Calibration bounds target the frozen population, with joint tails $.05/48$ across both decoders. All rows meet .10/.20 and the 40-release minimum. Cost includes recorded ranking, reference generation, and search, not service latency.

| model | ranker | policy | rel.% | CM% | err.% | $U_{\mathrm{rec}}$ | $U_{\mathrm{ev}}$ | s/query |
|---|---|---|---|---|---|---|---|---|
| Qwen | BM25 | addition | 96.7 | 83.0 | 14.17 | 0.0560 | 0.1797 | 0.328 |
| Qwen | BM25 | forward | 99.3 | 85.2 | 14.20 | 0.0154 | 0.1773 | 0.381 |
| Qwen | BM25 | recursive | 98.8 | 84.8 | 14.17 | 0.0269 | 0.1762 | 0.522 |
| Qwen | Base BGE | addition | 96.1 | 90.8 | 5.52 | 0.0560 | 0.0980 | 0.338 |
| Qwen | Base BGE | forward | 99.2 | 94.0 | 5.24 | 0.0103 | 0.0970 | 0.393 |
| Qwen | Base BGE | recursive | 98.7 | 93.5 | 5.27 | 0.0242 | 0.0928 | 0.531 |
| Qwen | Fusion | addition | 95.9 | 92.1 | 3.96 | 0.0610 | 0.0789 | 0.343 |
| Qwen | Fusion | forward | 99.2 | 95.2 | 4.03 | 0.0120 | 0.0814 | 0.400 |
| Qwen | Fusion | recursive | 98.8 | 94.7 | 4.15 | 0.0256 | 0.0750 | 0.532 |
| Qwen | Sup. BGE | addition | 96.4 | 96.1 | 0.31 | 0.0573 | 0.0162 | 0.321 |
| Qwen | Sup. BGE | forward | 99.5 | 99.2 | 0.30 | 0.0137 | 0.0157 | 0.370 |
| Qwen | Sup. BGE | recursive | 99.2 | 98.9 | 0.30 | 0.0228 | 0.0157 | 0.497 |
| Phi-3 | BM25 | addition | 94.6 | 81.4 | 13.95 | 0.0868 | 0.1680 | 0.347 |
| Phi-3 | BM25 | forward | 99.1 | 85.0 | 14.23 | 0.0199 | 0.1795 | 0.409 |
| Phi-3 | BM25 | recursive | 98.5 | 84.6 | 14.11 | 0.0348 | 0.1676 | 0.575 |
| Phi-3 | Base BGE | addition | 94.1 | 89.2 | 5.21 | 0.0845 | 0.0839 | 0.359 |
| Phi-3 | Base BGE | forward | 99.0 | 93.5 | 5.56 | 0.0199 | 0.0953 | 0.419 |
| Phi-3 | Base BGE | recursive | 98.5 | 93.3 | 5.28 | 0.0282 | 0.0936 | 0.567 |
| Phi-3 | Fusion | addition | 94.5 | 90.5 | 4.23 | 0.0787 | 0.0784 | 0.372 |
| Phi-3 | Fusion | forward | 98.9 | 94.4 | 4.55 | 0.0212 | 0.0799 | 0.428 |
| Phi-3 | Fusion | recursive | 98.6 | 94.2 | 4.46 | 0.0269 | 0.0753 | 0.579 |
| Phi-3 | Sup. BGE | addition | 94.4 | 94.2 | 0.21 | 0.0868 | 0.0126 | 0.355 |
| Phi-3 | Sup. BGE | forward | 99.4 | 99.2 | 0.20 | 0.0282 | 0.0158 | 0.411 |
| Phi-3 | Sup. BGE | recursive | 99.1 | 98.9 | 0.20 | 0.0336 | 0.0118 | 0.559 |

Initial test support completeness was 82.9%, 92.6%, 94.4%, 99.7% for BM25, base BGE, fusion, and supervised BGE, respectively. Protected forward exceeds addition descriptively by 2.2–3.3 percentage points for Qwen and 3.6–5.0 points for Phi-3. Its advantages over recursive deletion are only 0.2–0.5 points; no paired superiority test was prespecified or performed for this experiment. All policies certify here, so this cohort does not demonstrate an additional exclusion benefit of dual-risk admission. Selection of a disjoint subset changes the population and may change difficulty; higher rates than in earlier cohorts are not evidence of improved model capability. These annotation endpoints do not measure answer accuracy, semantic entailment, or human utility.

**Question-type composition.** Document exclusions also change the question mix. Table 12 compares the two test cohorts and reports supervised-BGE forward within each document-isolated stratum. These descriptive strata have no separate certificates. The 10 inference questions are too few to support a stable subgroup conclusion.

Table 12: Question-type audit. Counts refer to the intermediate-retrieval and document-isolated test sets (1,000 each). CM is complete matching coverage for supervised-BGE forward within the document-isolated stratum; initial completeness is shared across decoders.

| Type | Intermediate $n$ | Isolated $n$ | Initial complete | Qwen CM | Phi-3 CM |
|---|---|---|---|---|---|
| comparison | 325 | 880 | 880/880 | 878/880 | 876/880 |
| compositional | 452 | 66 | 66/66 | 64/66 | 65/66 |
| bridge comparison | 203 | 44 | 41/44 | 41/44 | 41/44 |
| inference | 20 | 10 | 10/10 | 9/10 | 10/10 |

Comparison questions increase from 32.5% to 88.0%, and all 880 have complete initial support. The selected population thus combines a different question mix with highly complete initial evi-

dence. The 99.2% aggregate therefore describes this selected population; it is not an estimate of performance under the earlier question mixture.

### C.5 SIMPLE FIXED-CONTEXT CONTROLS

These controls compare fixed-size checking and random addition with likelihood-guided search on the same 1,000 document-isolated 2Wiki test queries. Each decoder and each of base, fusion, and supervised BGE used fixed top-four checking, fixed top-eight checking, and protected random forward addition. The random policy kept the initial four passages and added others in a per-query hash order. All controls were prespecified and locally frozen before their outputs were read. They received no new calibration samples and inherit no certification.

Table 13: Descriptive controls on the document-isolated test role. Release and complete matching coverage (CM) use all queries; evidence error conditions on release. Rates are percentages; size is mean passages among releases. The seconds column is control generation time only and excludes shared ranking and full-context reference generation.

| model | ranker | policy | release | CM | error | size | gen. s/q |
|---|---|---|---|---|---|---|---|
| Qwen | Base BGE | fixed4 | 86.2 | 81.5 | 5.45 | 4.00 | .119 |
| | | fixed8 | 92.4 | 91.1 | 1.41 | 8.00 | .136 |
| | | random | 96.3 | 91.4 | 5.09 | 4.21 | .180 |
| | Fusion | fixed4 | 86.4 | 82.8 | 4.17 | 4.00 | .115 |
| | | fixed8 | 92.6 | 92.3 | .32 | 8.00 | .135 |
| | | random | 96.4 | 92.2 | 4.36 | 4.23 | .175 |
| | Sup. BGE | fixed4 | 86.6 | 86.3 | .35 | 4.00 | .112 |
| | | fixed8 | 91.9 | 91.9 | .00 | 8.00 | .134 |
| | | random | 96.6 | 96.3 | .31 | 4.19 | .162 |
| Phi-3 | Base BGE | fixed4 | 81.2 | 77.1 | 5.05 | 4.00 | .123 |
| | | fixed8 | 89.4 | 88.1 | 1.45 | 8.00 | .138 |
| | | random | 95.9 | 90.8 | 5.32 | 4.33 | .196 |
| | Fusion | fixed4 | 81.7 | 78.4 | 4.04 | 4.00 | .119 |
| | | fixed8 | 89.9 | 89.5 | .44 | 8.00 | .136 |
| | | random | 96.3 | 92.1 | 4.36 | 4.33 | .195 |
| | Sup. BGE | fixed4 | 80.4 | 80.2 | .25 | 4.00 | .119 |
| | | fixed8 | 88.9 | 88.9 | .00 | 8.00 | .133 |
| | | random | 95.5 | 95.3 | .21 | 4.33 | .188 |

On these test queries, fixed top-four recovery is below the 90% target for every decoder–ranker combination. Fixed top-eight is closer but remains below 90% for Phi-3. Random addition improves recovery over fixed top-four. For complete matching coverage, supervised-BGE random addition remains below protected forward search (96.3% versus 99.2% for Qwen and 95.3% versus 99.2% for Phi-3). These controls did not include BM25. This supports adaptive selection as a useful mechanism in this candidate pool; it does not establish a universal superiority claim, and the controls have no new risk certificates.

### C.6 MUSIQUE TARGET-DOMAIN ADAPTATION

**Question and evaluation population.** This follow-up asks whether adapting the evidence ranker to MuSiQue improves complete matching and permits joint admission. It is separate from the historical MuSiQue studies in Appendix E and the document-isolated 2Wiki study. Enumeration of recorded local candidate, training, source, and inference files left 1,346 MuSiQue training-split queries. This is a bounded history audit, not proof of no prior access. After constructing this population, a single seeded random permutation (seed 202609251017), without hop quotas, allocated 500 recovery-calibration, 346 evidence-calibration, and 500 test queries. Both decoders share these roles; their results are not independent query replications. The statistical scope is this selected finite population, not arbitrary future MuSiQue queries.

**Training and isolation.** Target adaptation starts from BGE-reranker-base and trains on 137,738 historical question–paragraph pairs from 18,003 queries, including 16,863 positive pairs. Training excludes every pair whose normalized title or paragraph text occurs in any evaluation, development, or smoke context. Development uses 256 historical queries and smoke checks use eight; these query IDs are disjoint from training and evaluation. This filtering does not establish document isolation for all historical models: base and 2Wiki-transfer checkpoints retain their historical exposure. No new human annotations were collected. Training uses binary cross entropy, AdamW (learning rate $2 \times 10^{-5}$, weight decay .01), 10% linear warmup, gradient norm cap 1, bf16, maximum length 256, and effective batch size 64 for three epochs. The training seed is 202609251019 (reduced modulo $2^{32}$ for NumPy). Epoch selection maximizes the number of development queries whose top-four passages contain all support, breaking ties toward earlier epochs. Epoch 3 obtains 179/256 (69.9%) and is selected. Inputs and dependencies were hashed before training, and the selected checkpoint was hashed before evaluation. An initial NumPy seed-range failure occurred before any training step and was recorded; inference scheduling subsequently expanded from four to eight GPUs without changing the policies.

**Policies and calibration.** Three rankers (base BGE, 2Wiki-transfer BGE, and target-adapted BGE), three searches (ordered addition, protected forward, and protected recursive deletion), and two decoders define 18 policies. Each starts with protected top-four passages and permits releases of at most 16 of 20 passages. Prompts, deterministic generation with a 64-token cap, answer normalization, and search semantics are held fixed. No policy is retuned using calibration or test results. Certification is stored and hashed before test inference. Recovery failure must be at most .10 and conditional evidence error at most .20, with at least 40 evidence-calibration releases. Bounds use exact hypergeometric inversion, maximizing over the unknown number of population releases for conditional evidence risk, as in the finite-population extension. Unlike the main-study reanalysis, this study protects all three admission rules jointly. The total error budget .05 is divided equally among recovery-only, evidence-only, and joint admission. Each single-risk family has 18 bounds and uses tail $.05/54$; the joint family has 36 bounds and uses $.05/108$. A union bound covers all three rules and both decoders; it does not require their independence. Disjoint roles sampled without replacement are not asserted independent.

**Test results and limits.** Tables 14 and 15 report all policies. Complete matching counts output recovery together with retention of every annotated support passage; it is not answer accuracy. All test differences are descriptive, with no post-hoc significance claims. Compared with base BGE on the same queries, target-adapted forward improves complete matching by 23.8 and 19.8 percentage points for Qwen and Phi-3. It exceeds target-adapted addition by 7.6/7.0 points, but trails recursive deletion by 0.6/0.8 points. Forward uses 5.25/5.95 passages on average among releases, versus 6.80/7.29 for recursive deletion; these means condition on each policy's own releases. The consistent gains across decoders identify ranker quality as an important factor in evidence retention. The comparison with recursive deletion remains a tradeoff between recovery, complete matching, and context size.

Table 14: MuSiQue adaptation: descriptive results on the same 500 test queries per decoder. Recovery and complete matching use all 500 queries; evidence error and mean passages condition on successful recovery. All rates are percentages.

| Decoder | Ranker | Search | Recovery | Complete | Ev. error | Passages |
|---|---|---|---|---|---|---|
| Qwen | Base | Addition | 87.6 | 58.8 | 32.9 | 6.54 |
| Qwen | Base | Forward | 97.6 | 58.0 | 40.6 | 5.25 |
| Qwen | Base | Recursive | 96.6 | 64.4 | 33.3 | 6.92 |
| Qwen | Transfer | Addition | 87.4 | 58.4 | 33.2 | 7.41 |
| Qwen | Transfer | Forward | 96.8 | 53.8 | 44.4 | 5.43 |
| Qwen | Transfer | Recursive | 94.8 | 58.6 | 38.2 | 6.84 |
| Qwen | Target | Addition | 84.6 | 74.2 | 12.3 | 6.46 |
| Qwen | Target | Forward | 98.0 | 81.8 | 16.5 | 5.25 |
| Qwen | Target | Recursive | 95.2 | 82.4 | 13.4 | 6.80 |
| Phi-3 | Base | Addition | 79.2 | 53.2 | 32.8 | 6.92 |
| Phi-3 | Base | Forward | 92.8 | 55.8 | 39.9 | 5.98 |
| Phi-3 | Base | Recursive | 91.8 | 61.8 | 32.7 | 7.44 |
| Phi-3 | Transfer | Addition | 77.4 | 48.2 | 37.7 | 7.22 |
| Phi-3 | Transfer | Forward | 92.4 | 51.4 | 44.4 | 6.06 |
| Phi-3 | Transfer | Recursive | 91.2 | 56.2 | 38.4 | 7.29 |
| Phi-3 | Target | Addition | 78.0 | 68.6 | 12.1 | 7.02 |
| Phi-3 | Target | Forward | 91.8 | 75.6 | 17.6 | 5.95 |
| Phi-3 | Target | Recursive | 90.6 | 76.4 | 15.7 | 7.29 |

Table 15: MuSiQue adaptation: simultaneous calibration upper bounds and admission decisions. The displayed bounds use the joint-admission tail. R/E/J denote recovery-only, evidence-only, and joint admission; R/E use separately computed bounds with less conservative tails. No policy passes joint admission.

| Decoder | Ranker | Search | $U_{\mathrm{rec}}$ | $U_{\mathrm{ev}}$ | R | E | J |
|---|---|---|---|---|---|---|---|
| Qwen | Base | Addition | 0.2229 | 0.4279 | no | no | no |
| Qwen | Base | Forward | 0.0773 | 0.5185 | yes | no | no |
| Qwen | Base | Recursive | 0.1085 | 0.4573 | no | no | no |
| Qwen | Transfer | Addition | 0.2363 | 0.4371 | no | no | no |
| Qwen | Transfer | Forward | 0.0669 | 0.5140 | yes | no | no |
| Qwen | Transfer | Recursive | 0.1062 | 0.4423 | no | no | no |
| Qwen | Target | Addition | 0.2578 | 0.1596 | no | yes | no |
| Qwen | Target | Forward | 0.0721 | 0.2451 | yes | no | no |
| Qwen | Target | Recursive | 0.1129 | 0.2118 | no | no | no |
| Phi-3 | Base | Addition | 0.3046 | 0.4628 | no | no | no |
| Phi-3 | Base | Forward | 0.1315 | 0.5369 | no | no | no |
| Phi-3 | Base | Recursive | 0.1545 | 0.4354 | no | no | no |
| Phi-3 | Transfer | Addition | 0.3314 | 0.4628 | no | no | no |
| Phi-3 | Transfer | Forward | 0.1204 | 0.5416 | no | no | no |
| Phi-3 | Transfer | Recursive | 0.1337 | 0.4632 | no | no | no |
| Phi-3 | Target | Addition | 0.3358 | 0.2011 | no | yes | no |
| Phi-3 | Target | Forward | 0.1248 | 0.2430 | no | no | no |
| Phi-3 | Target | Recursive | 0.1478 | 0.2080 | no | no | no |

Although target-forward test point estimates meet both thresholds, its calibration evidence upper bounds are .2451/.2430; Phi-3 also misses the recovery bound (.1248). These failures prevent joint certification of target-adapted forward search; Table 15 shows that every other policy also fails at least one joint-admission bound. Recovery-only admits all three Qwen forward configurations, whereas evidence-only admits target addition for both decoders, whose test recovery rates are only 84.6%/78.0%. The two single-risk rules thus admit different policies, while the joint rule admits none.

After all roles have been evaluated, the 1,346-query census gives target-forward recovery rates of 97.0%/91.6% and conditional evidence errors of 16.3%/16.8% for Qwen/Phi-3. These are exact de-

scriptive summaries of the selected population, including calibration queries. They do not retroactively change the frozen admission decisions or certify performance on future queries. Comparisons to older MuSiQue cohorts cannot isolate adaptation because their query populations differ; the within-cohort ranker comparisons above provide the relevant evidence.

**Where the adaptation gains occur.** To locate the source of the gain, a post-hoc paired analysis holds the 500 test queries fixed and groups them by initial support completeness under base and target BGE (Table 16). Initial completeness rises from 241/500 (48.2%) to 361/500 (72.2%). Among the 156 queries changing from incomplete to complete initial support, complete matching increases by 118 queries for Qwen and 106 for Phi-3, compared with overall net gains of 119 and 99. Other strata contain both gains and losses. This descriptive decomposition locates most of the net improvement in queries whose starting evidence becomes complete; it is not a causal mediation analysis.

Table 16: Paired test-query strata for forward search, grouped by initial evidence completeness under base and target BGE. Entries give complete-match counts before and after adaptation, using the same queries within each row. No subgroup is separately calibrated.

| Initial support: base → target | $n$ | Qwen CM | Phi-3 CM |
|---|---|---|---|
| Incomplete → incomplete | 103 | 18 → 34 | 18 → 32 |
| Incomplete → complete | 156 | 36 → 154 | 36 → 142 |
| Complete → incomplete | 36 | 35 → 19 | 31 → 13 |
| Complete → complete | 205 | 201 → 202 | 194 → 191 |

Under target BGE, 139 queries still begin with incomplete support; only 53 (Qwen) and 45 (Phi-3) end in complete matching releases. Better retrieval therefore does not remove the limitation of search on missing evidence. These counts cannot be compared as repair rates against the base-incomplete stratum without accounting for its different membership.

### C.7 Release burden and reference correctness

The following tables return to the three main studies. They separate returned-context size, performance on common releases, reference correctness, and paired method comparisons, which use different denominators.

Table 17: Supervised-BGE release burden in the three main tests. Size and prompt-token summaries condition on release. Size counts passages; tokens are means. P90 is the 90th percentile of size. Reduction is the median per-query reduction versus the full-context prompt. Tokens include the question, demonstrations, and chat formatting, so this is input burden, not evidence-only length or measured human effort.

| study | policy | rel. | mean size | median/P90 | tokens | reduction% |
|---|---|---|---|---|---|---|
| MuSiQue/Qwen | top4 | 438 | 4.00 | 4.0/4.0 | 709 | 73.2 |
| MuSiQue/Qwen | addition | 806 | 6.27 | 4.0/12.0 | 1017 | 67.9 |
| MuSiQue/Qwen | forward | 961 | 5.41 | 5.0/8.0 | 901 | 68.2 |
| MuSiQue/Qwen | recursive | 919 | 6.78 | 6.0/12.0 | 1070 | 63.9 |
| 2Wiki/Qwen | top4 | 673 | 4.00 | 4.0/4.0 | 643 | 40.1 |
| 2Wiki/Qwen | addition | 853 | 4.49 | 4.0/6.0 | 699 | 36.7 |
| 2Wiki/Qwen | forward | 946 | 4.49 | 4.0/6.0 | 704 | 36.3 |
| 2Wiki/Qwen | recursive | 938 | 4.66 | 4.0/7.0 | 718 | 34.9 |
| MuSiQue/Phi-3 | top4 | 339 | 4.00 | 4.0/4.0 | 781 | 73.4 |
| MuSiQue/Phi-3 | addition | 723 | 6.82 | 5.0/13.0 | 1185 | 66.2 |
| MuSiQue/Phi-3 | forward | 906 | 6.11 | 5.0/10.5 | 1094 | 66.4 |
| MuSiQue/Phi-3 | recursive | 892 | 7.64 | 7.0/13.0 | 1299 | 59.2 |

Table 18: Forward minus comparator on the same queries released by both supervised-BGE policies. Entries are mean differences; negative values favor forward. These descriptive contrasts condition on common release and do not establish population or human-audit benefits.

| study | comparator | common releases | passages | prompt tokens |
|---|---|---|---|---|
| MuSiQue/Qwen | addition | 805 | -1.38 | -181 |
| MuSiQue/Qwen | recursive | 916 | -1.52 | -185 |
| 2Wiki/Qwen | addition | 851 | -0.20 | -17 |
| 2Wiki/Qwen | recursive | 928 | -0.18 | -15 |
| MuSiQue/Phi-3 | addition | 717 | -1.52 | -208 |
| MuSiQue/Phi-3 | recursive | 867 | -1.55 | -206 |

Table 19: Current main-test reference accuracy strata for supervised-BGE forward search. Correctness uses normalized benchmark answer/alias matching. Complete counts releases retaining all annotated support. These descriptive strata are not separately certified.

| study | reference | queries | releases | complete |
|---|---|---|---|---|
| MuSiQue/Qwen | correct | 299 | 299 | 286 |
| MuSiQue/Qwen | incorrect | 701 | 662 | 630 |
| 2Wiki/Qwen | correct | 535 | 533 | 531 |
| 2Wiki/Qwen | incorrect | 465 | 413 | 410 |
| MuSiQue/Phi-3 | correct | 173 | 173 | 164 |
| MuSiQue/Phi-3 | incorrect | 827 | 733 | 696 |

Table 20: Prespecified supervised-BGE forward comparisons. Wins/losses count discordant complete-match outcomes. Two-sided exact tests use Holm correction over all six comparisons within each study, including the other rankers.

| study | comparator | wins/losses | adjusted $p$ |
|---|---|---|---|
| MuSiQue/Qwen | addition | 149/11 | 2.75e-31 |
| MuSiQue/Qwen | recursive | 46/10 | 6.23e-06 |
| 2Wiki/Qwen | addition | 95/3 | 5.94e-24 |
| 2Wiki/Qwen | recursive | 18/10 | 0.555 |
| MuSiQue/Phi-3 | addition | 180/18 | 5.11e-34 |
| MuSiQue/Phi-3 | recursive | 42/30 | 0.389 |

### C.8 ALL MAIN-STUDY CONFIGURATIONS

These tables include every ranker–search combination, including policies that failed calibration. Each row separates calibration counts and bounds from test release counts and evidence error. In the tables, base and sup. denote base and supervised BGE; top4 checks only the initial four passages.

Table 21: MuSiQue/Qwen. All 12 frozen configurations. Recovery and evidence fractions are calibration failures/evaluations and errors/releases. Test releases and complete matches use 1,000 queries; error conditions on release. Bounds use tail .05/24. MuSiQue numerical passes do not establish independence from historical development.

| policy | rec. | $U_r$ | ev. | $U_e$ | rel. | complete | err.% |
|---|---|---|---|---|---|---|---|
| BM25/top4 | 685/1000 | 0.7264 | 90/160 | 0.6740 | 289 | 114 | 60.6 |
| BM25/addition | 197/1000 | 0.2353 | 207/432 | 0.5491 | 767 | 365 | 52.4 |
| BM25/forward | 29/1000 | 0.0476 | 297/542 | 0.6094 | 954 | 409 | 57.1 |
| BM25/recursive | 52/1000 | 0.0753 | 254/519 | 0.5531 | 925 | 450 | 51.4 |
| base/top4 | 601/1000 | 0.6452 | 93/223 | 0.5151 | 396 | 218 | 44.9 |
| base/addition | 157/1000 | 0.1924 | 151/461 | 0.3932 | 825 | 522 | 36.7 |
| base/forward | 37/1000 | 0.0575 | 213/537 | 0.4589 | 954 | 541 | 43.3 |
| base/recursive | 48/1000 | 0.0706 | 178/515 | 0.4082 | 924 | 585 | 36.7 |
| sup./top4 | 572/1000 | 0.6168 | 15/235 | 0.1231 | 438 | 418 | 4.6 |
| sup./addition | 173/1000 | 0.2097 | 17/456 | 0.0702 | 806 | 778 | 3.5 |
| sup./forward | 29/1000 | 0.0476 | 29/544 | 0.0869 | 961 | 916 | 4.7 |
| sup./recursive | 48/1000 | 0.0706 | 21/516 | 0.0722 | 919 | 880 | 4.2 |

Table 22: 2Wiki/Qwen. All 12 frozen configurations. Recovery and evidence fractions are calibration failures/evaluations and errors/releases. Test releases and complete matches use 1,000 queries; error conditions on release. Bounds use tail .05/24. Certification assumes IID queries and independently fixed policies on the newly reserved cohort.

| policy | rec. | $U_r$ | ev. | $U_e$ | rel. | complete | err.% |
|---|---|---|---|---|---|---|---|
| BM25/top4 | 409/1000 | 0.4545 | 128/329 | 0.4690 | 595 | 369 | 38.0 |
| BM25/addition | 185/1000 | 0.2225 | 156/455 | 0.4094 | 827 | 553 | 33.1 |
| BM25/forward | 48/1000 | 0.0706 | 179/538 | 0.3934 | 953 | 635 | 33.4 |
| BM25/recursive | 62/1000 | 0.0870 | 171/531 | 0.3828 | 934 | 634 | 32.1 |
| base/top4 | 329/1000 | 0.3730 | 95/382 | 0.3169 | 683 | 516 | 24.5 |
| base/addition | 149/1000 | 0.1838 | 104/480 | 0.2749 | 847 | 663 | 21.7 |
| base/forward | 52/1000 | 0.0753 | 112/536 | 0.2632 | 955 | 753 | 21.2 |
| base/recursive | 66/1000 | 0.0916 | 105/531 | 0.2514 | 932 | 744 | 20.2 |
| sup./top4 | 314/1000 | 0.3576 | 2/389 | 0.0263 | 673 | 669 | 0.6 |
| sup./addition | 154/1000 | 0.1892 | 2/471 | 0.0218 | 853 | 849 | 0.5 |
| sup./forward | 47/1000 | 0.0694 | 2/540 | 0.0190 | 946 | 941 | 0.5 |
| sup./recursive | 60/1000 | 0.0847 | 2/530 | 0.0194 | 938 | 933 | 0.5 |

Table 23: MuSiQue/Phi-3. All 12 frozen configurations. Recovery and evidence fractions are calibration failures/evaluations and errors/releases. Test releases and complete matches use 1,000 queries; error conditions on release. Bounds use tail .05/24. MuSiQue numerical passes do not establish independence from historical development.

| policy | rec. | $U_r$ | ev. | $U_e$ | rel. | complete | err.% |
|---|---|---|---|---|---|---|---|
| BM25/top4 | 742/1000 | 0.7806 | 103/161 | 0.7445 | 271 | 85 | 68.6 |
| BM25/addition | 240/1000 | 0.2807 | 217/422 | 0.5845 | 718 | 323 | 55.0 |
| BM25/forward | 72/1000 | 0.0985 | 293/517 | 0.6291 | 908 | 365 | 59.8 |
| BM25/recursive | 102/1000 | 0.1323 | 252/505 | 0.5635 | 893 | 419 | 53.1 |
| base/top4 | 681/1000 | 0.7226 | 61/172 | 0.4657 | 326 | 180 | 44.8 |
| base/addition | 217/1000 | 0.2564 | 123/420 | 0.3603 | 752 | 492 | 34.6 |
| base/forward | 73/1000 | 0.0996 | 198/504 | 0.4571 | 908 | 509 | 43.9 |
| base/recursive | 95/1000 | 0.1245 | 164/509 | 0.3843 | 896 | 550 | 38.6 |
| sup./top4 | 659/1000 | 0.7014 | 10/185 | 0.1196 | 339 | 321 | 5.3 |
| sup./addition | 276/1000 | 0.3182 | 14/406 | 0.0690 | 723 | 698 | 3.5 |
| sup./forward | 80/1000 | 0.1076 | 25/509 | 0.0830 | 906 | 860 | 5.1 |
| sup./recursive | 103/1000 | 0.1334 | 20/499 | 0.0720 | 892 | 848 | 4.9 |

### C.9 SINGLE-RISK ADMISSION AT MATCHED FAMILYWISE CONFIDENCE

This analysis asks whether joint admission rejects policies because of the second risk or merely because it divides confidence across more bounds. It recomputes bounds post hoc from the saved calibration counts. Each single-risk rule allocates $.05/12$ to each of its 12 bounds; joint admission retains the original $.05/24$ allocation for 24 bounds. Each rule separately targets 95% familywise confidence within a study. This does not give a joint 95% guarantee across rules or studies. All admission sets match the common-tail comparison, including the Phi-3 failure. No policy, threshold, or ranking was selected from test performance.

Table 24: Policy admission at matched familywise confidence. Counts refer to fixed policies, not independent query samples. "Excluded" counts single-risk passes rejected by the joint rule. "Test exceeds" counts those excluded policies whose omitted risk exceeds its target empirically on test data; these are not population-risk tests. MuSiQue retains its historical-independence limitation.

| study | single-risk rule | admitted | excluded | test exceeds | joint admitted |
|---|---|---|---|---|---|
| 2Wiki / Qwen | recovery | 6 | 4 | 4 | 2 |
| | evidence | 4 | 2 | 2 | 2 |
| MuSiQue / Qwen | recovery | 6 | 4 | 4 | 2 |
| | evidence | 4 | 2 | 2 | 2 |
| MuSiQue / Phi-3 | recovery | 2 | 2 | 2 | 0 |
| | evidence | 4 | 4 | 3 | 0 |

Every excluded policy fails the omitted risk's joint-calibration bound. None is excluded only because the checked risk receives a stricter tail allocation. The comparison concerns policy admission, not a new evidence-only release filter. Recovery failure remains one minus coverage for the unchanged online policies. On 2Wiki, the evidence-only rule also admits top-four checking and ordered addition, costing 0.384 and 0.661 seconds per query, respectively. Jointly admitted forward and recursive search cost 0.868 and 1.143 seconds per query. These recorded costs illustrate the tradeoff; they do not measure certification overhead or compare policies at equal compute. Nonadmitted policies were still run diagnostically, and their results are not treated as deployment releases. The numerical supplement reports all 36 policy rows, both confidence allocations, test denominators, and costs; no cross-policy mixture is assigned a certificate.

## D ABLATIONS AND FAILURE ANALYSIS

### D.1 STARTING-SET ABLATION

To test whether output recovery persists when initial evidence is degraded, we selected 256 of the historical 1,000 Phi-3 test queries by a fixed source-ID hash, without outcome filtering. A second hash ordered passages to define one random four-passage starting set per query. Both starting sets used the same downstream likelihood-guided search and were rerun in hash-randomized order. This is a diagnostic on previously evaluated queries, not an untouched confirmatory cohort or an average over random seeds. Supervised-BGE and random starting sets initially contained all support on 235 and 5 queries, respectively. Final releases were 241 and 237, of which 227 and 42 were complete. Evidence errors were 14/241 and 195/237. The complete-release difference was 72.27 points, with descriptive paired 95% interval [66.41, 77.73]. Initial evidence selection, rather than recovery alone, is therefore important in this tested configuration. No certificate is assigned to the random-starting-set rule.

### D.2 ANSWER CORRECTNESS AND INSPECTION CASES

The ablation concerns missing evidence. The next analysis asks a different question: whether a complete returned context can accompany an incorrect reference output. The following historical cohorts are defined in Appendix E.

Table 25: Protected forward search stratified by full-context reference accuracy. Complete + match counts releases retaining all annotated support. Evidence error is a proportion among releases. These subgroups have no separate certificate.

| cohort | reference | queries | releases | complete + match | evidence error |
|---|---|---|---|---|---|
| Initial Phi-3 | correct | 222 | 222 | 212 | .0450 |
| | incorrect | 778 | 714 | 671 | .0602 |
| Phi-3 comparison | correct | 153 | 153 | 144 | .0588 |
| | incorrect | 615 | 558 | 529 | .0520 |
| Qwen comparison | correct | 226 | 226 | 212 | .0619 |
| | incorrect | 542 | 510 | 485 | .0490 |

The comparison references matched a normalized gold answer or alias on 153/768 Phi-3 queries and 226/768 Qwen queries. Correctness-stratified release, completeness, size, and cost summaries are retained separately for each cohort; these are not subgroup certificates. An audit retaining annotated support around a wrong reference preserves material for inspecting that output, but does not validate the answer. Table 25 does not show consistently higher evidence error for incorrect references across cohorts. No human error-diagnosis benefit or semantic entailment improvement was measured. Cases were selected by a fixed hash within six categories defined by reference correctness and complete release, incomplete release, or abstention. The local casebook contains up to two cases per category, with source passages and annotation indices; it is not an independent human evaluation.

**Concrete inspection cases.** The first hash-selected Qwen complete-release case with a correct reference asks who Alfred D'Orsay Tennyson Dickens's mother was (query `2hop__125595_849321`). The model returns "Catherine Dickens." The released set contains passages 6, 8, 10, and 15, including both annotated support passages, 8 and 15. This is a four-passage reproduction of the output, not evidence that all four passages were causally necessary. The corresponding first case with an incorrect reference asks who was ordered to repel an assault in the location associated with Gonchen Monastery (query `2hop__553369_872`). The model returns "not applicable." Its four-passage release, 3, 4, 7, and 15, retains both support passages, 4 and 15. Passage 4 explicitly names "the Ming general Qu Neng," the benchmark answer. Thus, an inspector can see the model's unchanged output alongside the retained support that conflicts with it. This illustrates the audit's output and its limit: evidence completeness does not repair or certify the answer.

## E HISTORICAL COHORTS SUPPORTING RETAINED DIAGNOSTICS

These studies are separate from the three main tests. They document the historical certificate and the sources of the starting-set ablation and real cases. Their sample sizes and confidence families must not be combined with the main results.

### E.1 INITIAL PROTECTED-FORWARD VALIDATION

Table 26: Initial frozen validation: MuSiQue, Phi-3 native chat, completeness endpoint, at most 16 of 20 passages. Panel A certifies only the prespecified protected forward policy using one-sided 97.5% bounds. Panel B evaluates all three fixed policies on the same 1,000 held-out queries; risks are empirical proportions and released size counts passages. "Complete + match" counts queries whose released set both reproduces the reference and contains all annotated support. Search time is summed across queries, excluding shared ranking and reference generation; it is not wall time or an equal-compute comparison.

*A. Calibration of protected forward search*

| requirement | failures / evaluated | upper bound | target |
|---|---|---|---|
| Recovery | 97/1,200 | .0977 | .10 |
| Released evidence | 25/555 | .0658 | .20 |

*B. Held-out test performance and cost*

| policy | releases | complete + match | recovery risk | evidence risk | mean released size | search (s) |
|---|---|---|---|---|---|---|
| Direct cropping | 712 | 687 | .288 | .0351 | 7.12 | 1,298.15 |
| Static protected deletion | 819 | 793 | .181 | .0317 | 9.02 | 1,257.09 |
| Protected forward search | 936 | 883 | .064 | .0566 | 6.05 | 3,895.81 |

**Development and role isolation.** The policy was fixed after development on 64 previously inspected MuSiQue queries. These supplied the choice of policy, not confirmation evidence. The new validation selected from the MuSiQue training split after excluding source IDs and normalized questions in 28 enumerated historical files, including all 4,096 BGE training queries. Of 19,938 source queries, 16,341 were excluded for historical overlap or duplicate normalized questions, and 21 failed structural eligibility; 3,576 remained. Eligible pools had 20 indexed, nonempty passages and two to four annotated supporting paragraphs, consistent with their hop count. Content-blind hash ordering assigned 8 smoke, 1,200 recovery-calibration, 600 evidence-calibration, and 1,000 test queries, leaving 768 unused. There were no hop quotas and no additional selection role: model, policy, budget, and release rule had already been fixed. Shared documents and model pretraining exposure were not eliminated by this query-level exclusion.

**Search and comparison methods.** The decoder is Phi-3-mini-128k-instruct with its native chat template, bf16, greedy decoding, a 64-new-token limit, and native EOS IDs $\{32000, 32001, 32007\}$. All subsets are displayed in source-index order. Protected forward search generates from supervised BGE top-four first. After a failure, it scores every remaining single-passage addition by mean log probability of the frozen reference's first line, selects the highest score (ties by source index), and generates again. It stops at the first normalized output match or after 16 passages. Candidate scoring uses batches of two with explicit attention-mask-derived position IDs for left padding. This is a standard greedy search with a protected initial set, not a new optimization algorithm. At most 126 candidate sequences, 66 scoring batches, and 13 generations are used per query.

Direct cropping generates at BGE depths $4, 6, 8, 10, 12, 14, 15, 16$, stopping at the first match. Static protected deletion starts with all 20 passages, protects BGE top-four, and tries successive four-passage blocks in increasing BGE relevance order. It accepts a removal only when the output matches, does not subdivide rejected blocks, and returns the last accepted set if its size is between 4 and 16. Returning the initial full context is a failure. It makes at most four block attempts here. All methods use the same normalization and completeness labels; no official label enters inference. The comparison matches queries, decoder, reference, and maximum context size, not model computation or actual token length.

**Certification and execution.** Only protected forward search has a prespecified joint certification claim. The two calibration bounds each use tail probability $.025$; at least 40 evidence-calibration releases are required. There is no threshold tuning or additional rejection rule after recovery. A calibration failure would prevent use as a certified policy, but would not cancel the diagnostic test evaluation. The saved calibration decision precedes test preparation. The 1,200-query recovery sample permits at most 99 failures; 97 were observed. Under hypothetical IID true recovery risks

.07, .08, .09, exact passing probabilities are .958, .650, .197, respectively. These are design sensitivities, not predicted success probabilities for this experiment.

The three methods ran independently on every query in a fixed hash-randomized order. Cross-stage timings are not compared as evidence of speedup. Eight repeated full-context smoke outputs matched their references. All planned rows completed; test candidate generations hit the 64-token limit 14, 14, and 20 times for cropping, deletion, and forward search, respectively, and none were empty. Capped outputs were retained.

**Answer correctness and scope.** The full-context reference matched a normalized gold answer or alias on 222/1,000 test queries. Completeness concerns benchmark support annotations, not whether those passages entail an incorrect reference. The study does not establish factual correctness, causal dependence, or superiority across models and datasets. The protected-forward policy has higher conditional evidence error than both simple controls on this test cohort.

### E.2 Stronger search controls and decoder transfer

**Design and query provenance.** The Phi-3 comparison used all 768 remaining queries from the initial exclusion-audited pool, after the 8/1,200/600/1,000 role allocation. Its policies and analyses were frozen before collecting these outputs. Each query received all four strategies in a hash-randomized order on the recorded execution hardware. The output-size cap remained 16 of 20 passages. BGE-ordered addition generated at every depth from 4 to 16. Recursive protected deletion tried the same low-relevance four-passage blocks as static deletion, then recursively halved rejected blocks in depth-first order. An unsuccessful deletion left the current set unchanged. The four binary subdivision trees permit at most 28 generation checks. Neither method used official evidence labels during inference.

Qwen2.5-7B-Instruct reused the original 1,200 recovery and 600 evidence queries, and the 768-query comparison set. These roles were query-disjoint within the decoder, but their Phi-3 evaluations do not constitute independent replications. The Qwen plan was frozen while the Phi-3 comparison was running, without selecting policies from its outcomes. Previous project work had used Qwen as a relation verifier; we do not claim that the model was new to the project. This study changes the reference decoder, not the dataset or trained ranker. The query-level historical exclusions do not establish document independence or absence of pretraining exposure.

**Decoder and search implementation.** Qwen used native chat formatting, bf16, FlashAttention-2, greedy decoding, one beam, a 64-new-token limit, and repetition penalty 1.0. Sampling parameters were disabled explicitly; EOS IDs were 151645 and 151643. The user message contained the same few-shot context/question prompt as the initial experiment. Passages were serialized in source-index order. The native template then added the assistant generation prefix. No prompt was silently truncated; exceeding the 32,768-token context limit would stop execution. The supervised BGE ordering was reused unchanged. Full-context references were regenerated with Qwen for each role. Candidate scoring used batches of two, explicit left-padding position IDs, and mean log probability of the reference's first line. The generation check, not likelihood alone, determined whether to release. Both model interfaces passed eight repeated-reference checks before their studies.

Table 27: Qwen calibration with simultaneous bounds for three prespecified policies. Each bound uses tail probability $.05/6$; recovery and evidence targets are .10 and .20. Evidence denominators count releases among 600 calibration queries. At least 40 releases are required.

| policy | recovery failures | upper | evidence errors | upper | status |
|---|---|---|---|---|---|
| Protected forward | 48/1,200 | .0556 | 29/577 | .0764 | certified |
| BGE addition | 225/1,200 | .2160 | 17/485 | .0604 | not certified |
| Recursive deletion | 74/1,200 | .0803 | 22/562 | .0633 | certified |

**Certification and comparisons.** Certification was saved before Qwen test inference began. Testing proceeded regardless of certification. An uncertified policy was ineligible for certified deployment, while its empirical test coverage was still reported. These are separate decisions and mea-

surements. The main comparisons were forward search minus BGE-ordered addition and forward search minus recursive deletion in complete, matching releases. Each experiment resampled query IDs jointly 10,000 times for descriptive paired bootstrap intervals. Two prespecified one-sided exact tests on discordant pairs used Holm correction within each experiment. Phi-3 wins/losses were 143/17 against BGE addition and 36/27 against recursive deletion; Qwen counts were 124/9 and 27/12. Adjusted $p$ values were $5.33 \times 10^{-26}$ and .1568 for Phi-3, and $5.38 \times 10^{-27}$ and .0119 for Qwen. Absence of significance in the Phi-3 deletion comparison does not establish equivalence. Certificates, descriptive intervals, and hypothesis-test families have distinct scopes.

### E.3 Earlier 768-query comparison

The following table reports the earlier 768-query studies, not the three 1,000-query main tests. Their calibration family and query roles are specified in Appendix E.2.

Table 28: MuSiQue comparison on 768 queries per decoder. Counts and percentages for releases and complete matches use all 768 queries; complete + match requires both output recovery and all annotated support. Bold marks the best observed value within each decoder, not statistical significance. Evidence error is a proportion and size is mean passages, both among releases. Search time includes scoring but excludes shared ranking and reference generation; it is cumulative search time, not wall time or an equal-compute budget.

| decoder | policy | releases (%) ↑ | complete + match (%) ↑ | error ↓ | size ↓ | search (s) ↓ |
|---|---|---|---|---|---|---|
| Phi-3 | Static deletion | 641 (83.46) | 618 (80.47) | **.0359** | 9.24 | **942.86** |
| | BGE addition | 569 (74.09) | 547 (71.22) | .0387 | 6.85 | 1,453.12 |
| | Recursive deletion | 694 (90.36) | 664 (86.46) | .0432 | 7.62 | 4,035.84 |
| | Protected forward | **711 (92.58)** | **673 (87.63)** | .0534 | **6.17** | 3,234.71 |
| Qwen2.5-7B | BGE addition | 607 (79.04) | 582 (75.78) | **.0412** | 6.13 | **1,142.80** |
| | Recursive deletion | 715 (93.10) | 682 (88.80) | .0462 | 6.83 | 3,107.70 |
| | Protected forward | **736 (95.83)** | **697 (90.76)** | .0530 | **5.35** | 2,345.42 |