TRAVA: A Temporal RAG Validity Audit of Knowledge Cutoffs and Evidence Support

Xuan-Bach Le[1,2,*†] **Phat T. Tran-Truong**[1,2,*] **Minh-Hieu Le**[1,2]
Duy Nguyen[3] **Son Ha Xuan**[4]
[1]Faculty of Computer Science and Engineering,
Ho Chi Minh City University of Technology (HCMUT), Vietnam
[2]Vietnam National University Ho Chi Minh City, Vietnam
[3]University of Economics Ho Chi Minh City, Vietnam
[4]The Business School, RMIT University, Vietnam
{lexuanbach,phatttt,hieu.lehieulhks}@hcmut.edu.vn
duynguyen.726202321440@st.ueh.edu.vn ha.son@rmit.edu.vn

Abstract

A correct answer from a retrieval-augmented generation (RAG) system does not show how much retrieval helped. The model may answer without retrieval, or its evidence may have been published too late. This makes retrievers and index updates difficult to compare. We introduce TRAVA[1], which jointly checks question dates, evidence publication dates, and answer support in the text shown to the model. On 878 dated questions, we test different calendar splits and compare retrieval conditions on the same questions. Performance differences without retrieval depend on the calendar split and remain uncertain. Although 11.7% of retrieved passages fall outside a three-day publication grace window, excluding them changes scores averaged over weekly quizzes by at most 1.9 token-F1 points for the tested models. In contrast, restricting the index to pre-2025 documents costs about 21 F1 on questions from 2025 onward across three generators. Removing each question's linked evidence lowers scores on both sides of the boundary. Automatic labels and legacy human annotations place retrieval gains mainly where the shown text supports the answer. TRAVA connects temporal checks to measured performance effects, while keeping answer support distinct from evidence use.

1 Introduction

The problem. Retrieval-augmented generation (RAG) gives models access to information beyond their training data. Yet a correct answer need not come from retrieval. TriviaQA (Joshi et al., 2017), Natural Questions (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018) may contain answers already present in pretraining. BEIR (Thakur et al., 2021) tests retrieval on fixed collections, while temporal benchmarks test changing knowledge (Kasai et al., 2023; Vu et al., 2024). Even on recent questions, a high score can be hard to interpret. Could the model answer without retrieval? Was the evidence published in time? Did the model see the part containing the answer? These distinctions guide decisions about retrievers, index updates and context length. Without them, we can credit retrieval for an answer the model already produces or mistake missing evidence for a generator failure. Figure 1 illustrates our three checks:

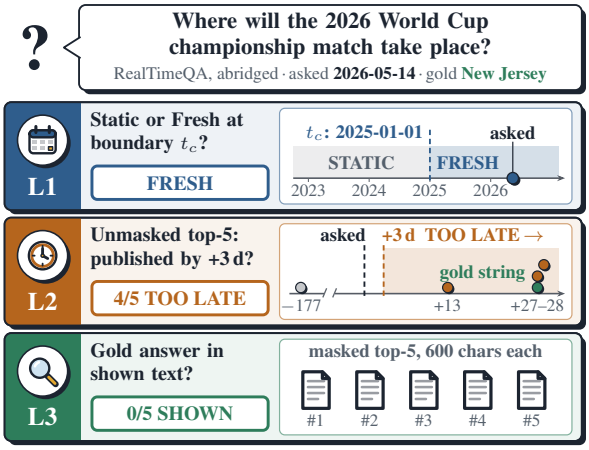


Figure 1: One dated question illustrates the three checks. L2 shows unmasked top-5 dates and a full-text gold-string match. L3 checks masked top-5 prefixes (600 characters each).

- **L1, question time.** We call questions before 2025-01-01 *Static* and those on or after that date *Fresh*. These date labels do not establish what appeared in training. Training cutoffs and differences in question difficulty also affect the comparison.
- **L2, evidence time.** We admit passages published by three days after each question. This

*Equal contribution.
†Corresponding author: Xuan-Bach Le.
[1]Artifact: https://github.com/lexuanbach/TRAVA.

grace window allows some later evidence. We call this passage filter a *temporal mask*.

- **L3, evidence content.** Using the known reference answer, we check whether the text shown to the model contains or entails it. We then compare retrieval gains with and without this support.

The World Cup question in Figure 1, asked on 2026-05-14, is Fresh. Four original passages fall outside the window, including an answer-bearing article from +27 days after the question. After masking, passages 3 and 5 mention "New Jersey" as a group-stage venue or host city, not the final's location. These mentions also lie beyond the 600 characters shown. The late article identifies the final's location, again beyond its shown prefix. Thus a string match need not answer the question, and a retrieved answer need not reach the model.

What TRAVA adds. ARES and RAGChecker diagnose retrieval and generation quality (Saad-Falcon et al., 2024; Ru et al., 2024). TRAVA brings time and shown-text support into a common experimental design. We vary the calendar boundary, exclude late passages, and change the index while keeping retrieval comparisons paired on the same questions. This links temporal violations to their measured effects on answers. On 878 dated RealTimeQA questions (Kasai et al., 2023), we also compare the original index with a pre-2025 *stale* index and a *leave-one-out* index that removes each question's linked evidence. Support labels locate where retrieval helps, with legacy human passage annotations providing a separate comparison.

Research questions. RQ1: how stable are Static/Fresh performance differences without retrieval across boundaries, question mixes, samples, metrics and documented cutoffs (§5.1)? **RQ2:** how often is evidence too late, and how much does excluding it change scores (§5.2)? **RQ3:** how do retrieval gains vary with the index, shown-text support and model size (§5.3)?

Our contributions.

1. **A joint audit of time and support.** We bring model cutoff records, passage dates and shown-text support into the same evaluation. Resampling whole quizzes captures shared question variation, and a versioned claim ledger tracks the evidence behind each result (Appendix A).
2. **A paired temporal-mask experiment.** With 11.7% of contexts outside the grace window, masking changes 38.0% of ordered top-5 lists but moves every tested score, averaged over weekly quizzes, by at most 1.9 F1. Frequent exposure to late evidence need not imply a large score effect.
3. **Index changes linked to answer support.** A stale index costs about 21 Fresh F1 at 3B, 32B and 123B, with Static within 1 F1. Retrieval gains concentrate where the shown text supports the answer across automatic labels, temporal masking and human annotation. We test this pattern's sensitivity to metrics, labels and context length.

Scope and organization. Support does not establish evidence use. Its scale analysis uses six parameter-known Hybrid models, while L1 uses seven. Date-based policies have model-specific trade-offs. Sections 3–5 present the method, setup and findings. Appendix W gives examples.

2 Related Work

RAG on fixed benchmarks. Combining a language-model generator with an external retriever improves knowledge-intensive question answering (QA) (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022) using dense (Karpukhin et al., 2020), late-interaction (Khattab and Zaharia, 2020), and learned-sparse (Formal et al., 2021) retrievers. Their relative accuracy varies across settings (Luan et al., 2021). Later work adds reflection (Asai et al., 2024), re-ranking (Sun et al., 2023), and systematic pipelines (Rau et al., 2024). These methods, together with BEIR (Thakur et al., 2021) and KILT (Petroni et al., 2021), evaluate against fixed corpora. Citation generation (Gao et al., 2023), factuality evaluation (Min et al., 2023), and grounding checks (Es et al., 2024; Tang et al., 2024) assess complementary aspects of evidence and responses without requiring a temporally fixed corpus. We use answer coverage as a retrieval statistic and reserve faithfulness for these dedicated methods.

Fine-grained RAG evaluation. Decomposing RAG quality into component judgements is established. ARES (Saad-Falcon et al., 2024) scores context relevance, answer faithfulness, and answer relevance with fine-tuned judges calibrated on a small human-annotated set. RAGChecker (Ru et al., 2024) checks claim-level entailment

among responses, gold answers, and retrieved chunks to separate retriever from generator errors. eRAG (Salemi and Zamani, 2024) labels each retrieved document by the answer the generator produces from it. Layer 3 uses a narrower, gold-conditioned label inside a dated design: support is judged on the passage text the generator saw and paired with interventions on the index (a publication mask, a stale index, and leave-one-out removal) rather than a judge alone. These frameworks supply alternative labels that could be evaluated within the same temporal design.

Fresh QA and cutoff measurement. RealTimeQA (Kasai et al., 2023) showed that search helps closed-book GPT-3 on recent news. FreshQA (Vu et al., 2024) reports the same pattern for fast-changing knowledge, while GrowOVER (Ko et al., 2024) maintains a continuously updated benchmark. Pretraining corpora may overlap evaluation sets (Magar and Schwartz, 2022; Golchin and Surdeanu, 2024; OpenAI, 2023; Liang et al., 2023), which Dynabench (Kiela et al., 2021), LiveBench (White et al., 2025), and AntiLeak-Bench (Wu et al., 2025) counter with adversarial, freshly published, or post-cutoff questions. Our audit also checks the dates of the retrieved documents. Cheng et al. (2024) trace a model's *effective* cutoff from dated corpora. We test the analogous signal in QA and find that accuracy also changes across content within the training period (Longpre et al., 2024). This within-period variation can exceed the boundary contrast and reinforces Zhang et al.'s (2026) caution that post-cutoff decay is a fragile contamination signal. Concurrently, Yan et al. (2026) train same-size models at successive annual cutoffs. Their intervention isolates horizon from size, while our shared-boundary audit measures how existing evaluations combine them.

Evidence support, conflict, and retrieval harm. Mallen et al. (2023) provide the closest result on the *popularity* axis: retrieval helps most when parametric recall is weak and can harm answers the model already produces. CRAG (Yang et al., 2024) finds that naive RAG's value tracks question dynamism. We extend this line from popularity and dynamism to three temporal validity conditions. HoH (Ouyang et al., 2025) measures harm from outdated facts, which connects directly to our stale-index intervention. DYNAMICQA (Marjanovic et al., 2024) and Tug-of-War (Jin et al., 2024) study arbitration between parametric and contextual evidence under conflict, with Tug-of-War evaluating eight models. We compare generators of different sizes within a common temporal evaluation, while retaining the resulting family and training confounds. Ren et al. (2025) evaluate whether models can judge what they know and recognize unknowns. Our question concerns how closed-book performance varies across question dates. Astute RAG (Wang et al., 2025) addresses imperfect evidence from the generator side, and concurrent work measures compliance with conflicting context (Chen et al., 2026). SituatedQA (Zhang and Choi, 2021) and PAT-Questions (Meem et al., 2024) make the *answer* time-dependent. Our audit makes the *split* time-dependent, which partitions every model at the same calendar boundary. Temporal robustness also covers language modelling and classification (Loureiro et al., 2022), time-aware knowledge probing (Dhingra et al., 2022), time-scoped QA (Chen et al., 2021), train and evaluation misalignment (Luu et al., 2022), and distraction by irrelevant context (Shi et al., 2023).

Temporal RAG resources and robustness suites. Zeng et al. (2025) stress-test retrieval-augmented systems under query and document perturbations on time-sensitive corpora. We examine how the definition of temporal quantities affects their interpretation. Chen et al. (2025) release a time-sensitive RAG benchmark over Chinese news, and Uddin et al. (2025) construct contamination-free synthetic temporal questions. Both resources could host the same three-layer audit. Gabburo et al. (2024) study retrieval complexity and QA difficulty. Our shown-text support label is a narrower observable, not a validated approximation of that construct. Yu et al. (2026) monitor faithfulness through model internals during generation. Such monitoring complements our offline labels, although it cannot supply a pre-answer routing signal.

Our position. TRAVA unifies boundary sensitivity, per-query evidence time, and shown-text support across explicit model subsets with a replayable claim ledger. We apply the same calendar split to each model, then use provider cutoff records and controlled slices to assess its interpretation. Index interventions and temporal masking complement the shown-text support diagnostic. RetrievalQA (Zhang et al., 2024) and SKR (Wang et al., 2023) provide the closest routing formulations. Our date-policy study asks whether a date-conditioned policy changes the closed-book-versus-

retrieval choice on held-out quizzes. Appendix P provides a claim-level comparison with the nearest work.

3 Method

TRAVA is a four-stage evaluation protocol for time-stamped question answering. It constructs date-based question slices, measures closed-book and retrieval-augmented performance, audits retrieved evidence, and tests the resulting interactions. The three layers define what we audit, while the four stages specify how: stages 1–2 establish dates and slices, stage 3 measures the L1 contrast, and stage 4 evaluates L2/L3 conditions. Appendix C records the invariants and glossary (§C.2).

Stage 1: Build dated slices. We use RealTimeQA (Kasai et al., 2023), a public CC BY-SA 4.0 collection of weekly news-quiz questions with gold answers, publication dates, and dated evidence documents. We ingest every weekly slice from 2022-06 to 2026-07 with both a question and an evidence file, retain questions with one gold span, and record the answer a_i, question date t_i, and linked evidence, which serves as the relevance judgement. For a document collected with question i, we enforce $t_{\text{publish}}(d) \leq t_i + 3$ days. This *three-day grace convention* allows a bounded publication lag rather than guaranteeing availability at question time. §5.2 tests windows of 0, 1 and 7 days. The invariant is relative to the source question, so pooled documents can become future evidence for other questions, which Layer 2 measures. We construct a *main* corpus from sampled questions and a *full* corpus that adds same-period distractors (Appendix K).

Stage 2: Choose a calendar split. Let $t_c = 2025\text{-}01\text{-}01$ denote the shared calendar boundary. We define *Static* questions by $t_i < t_c$ and *Fresh* questions by $t_i \geq t_c$. Static answers may have been available during pretraining, whereas Fresh denotes only post-boundary status. Seven of nine generators have undisclosed cutoffs, and the Claude training-data dates extend through July 2025 in the provider's model specifications, covering the dates of 95 Fresh items for each model (Appendix C.1). We therefore use the 2025 boundary as a shared descriptive split, 2023 against 2026 as a conservative sensitivity contrast that spans the documented cutoff range (some 2023 items follow GPT-4o-mini's October cutoff), and evaluate model-specific effective cutoffs in Section 5.1.

Stage 3: Compare closed-book performance. Closed-book accuracy measures whether a generator produces the answer without retrieval. A correct answer may reflect memorization, inference from related knowledge, question cues, or chance, and the closed-book contrast, Δ_{cut}, combines these sources. Its estimand is the *closed-book performance contrast* across the boundary. Identifying knowledge or memorization requires evidence that separates these sources.

Layer-3 terms. Layer 3 partitions the retrieval effect by the content shown, the first 600 characters of each retrieved passage. We keep four notions apart: *answer mention* (the gold string occurs in the shown text), *answer entailment* (a natural language inference (NLI) model judges that the shown text entails "the answer to the question is a_i" with an undated hypothesis), *answer correctness* (the output matches a_i) and *evidence use* (the answer was produced from the passage). Our support labels measure mention and undated entailment in the shown text. Date-resolved support requires temporal validation. We evaluate answer quality separately (Section 4). This offline, gold-conditioned split locates retrieval gains. Evidence use and faithfulness require separate tests.

Four comparisons. We express each comparison as a difference in token-F1 points. The first compares question groups. The others compare retrieval conditions or support groups. The unpaired closed-book contrast across L1 slices is

$$\Delta_{\text{cut}} = \text{F1}_{\text{Static}} - \text{F1}_{\text{Fresh}}.$$

The paired retrieval effect within a slice is

$$\Delta_{\text{RAG}} = \text{F1}_{\text{RAG}} - \text{F1}_{\text{closed-book}}.$$

The paired L2 effect after re-ranking under the grace convention is

$$\Delta_{\text{mask}} = \text{F1}_{\text{masked}} - \text{F1}_{\text{unmasked}}.$$

For L3, let G_1 and G_0 denote the retrieval gains on supported and unsupported questions, respectively. Their descriptive difference-in-differences is

$$\Delta_{\text{sup}} = G_1 - G_0,$$

with both arms recomputed inside each bootstrap draw. A positive interaction means retrieval helps

more (or harms less) on supported questions, even if neither arm gains overall. Question-pooled means weight questions equally, whereas quiz-weighted means give each weekly quiz equal weight. They differ by up to 6 F1, so each table names its weighting and the result files retain both.

Stage 4: Evaluate and test. Each system answers every question under one fixed prompt, with top-k=5 passages from one of six retrievers or from gold evidence. Static and Fresh contrasts are unpaired because their questions are disjoint. Within a slice, we pair retriever and closed-book answers to the same question. We resample whole weekly quizzes for question-level inference, exhaustively permute outcomes across parameter-known rungs for model-level inference, and correct for multiplicity within each analysis family (Section 4).

4 Experimental Setup

Benchmark. The 209 RealTimeQA weekly file pairs from 2022-06 to 2026-07 contain 4,916 distinct questions, of which 4,176 retain evidence after the timestamp filter (the uncapped population, Appendix N). A seeded sample takes 160 questions per year in 2022–2025 before filtering for surviving evidence, leaving 150/131/123/128 by year, plus all 346 retained calendar-2026 questions: 878 questions, 404 Static and 474 Fresh at the 2025 boundary (Table 11).

Replication sample. A second sample takes 800 unused questions, 200 per year in 2022–2025. All 2026 questions were already in the working sample. We answer the new questions over the frozen 26,611-document corpus, with the original questions re-answered under the same configuration. It spans 182 weekly quizzes, 172 shared with the original sample and resampled jointly, and 43 questions in ten new quizzes, so it is a within-source, largely same-week replication, not an independent temporal or cross-domain test (Appendix X.5).

Indexes. The **main** corpus pools 6,692 deduplicated evidence documents from the sampled questions (only 320 serve several questions, a restricted retrieval setting). The **full** corpus holds 26,611 documents from all raw questions, including 19,919 same-period distractors, with a per-question temporal mask (Appendix K).

Retrievers. We evaluate six CPU retrievers with public checkpoints and top-k=5: **BM25** (Robertson and Zaragoza, 2009), **DPR** (Karpukhin et al., 2020), **SPLADE++** (Formal et al., 2022), **ColBERTv2** (Santhanam et al., 2022), **BGE** (Xiao et al., 2024), and a **Hybrid** that fuses BM25 with BGE through Reciprocal Rank Fusion (Cormack et al., 2009). Table 12 reports retrieval quality. The six-retriever comparison uses Qwen2.5-32B, while cross-generator analyses use Hybrid. *Closed-book* and *Gold-Evidence* (source provenance, §Q) are the references. Each shown passage is its first 600 characters, and every support label is computed on that text.

Generators. Panels contain the generators with outputs for each analysis (Appendix U). The nine-model cutoff panel comprises Qwen2.5 at 0.5, 1.5, 3, 7, 32 and 72B, Mistral-Large-2 (Mistral 123B), and Claude Haiku 4.5 and Sonnet 4.5. The ten-model Hybrid panel replaces 0.5B with Llama-3-8B and adds GPT-4o-mini. The decision panel excludes Haiku. Qwen and Mistral disclose no cutoff. Llama and the API models have published cutoffs (Appendix C.1). Open weights run locally, with 32B, 72B and 123B quantised to 4 bits. Claude and GPT use APIs. Five additional Llama configurations test cutoff and size (Appendix H). All use temperature 0 and the fixed short-answer prompt (Appendix B). B denotes billions of parameters, our scale proxy. The six Qwen sizes and Mistral form the seven-model scale ladder.

Mistral prompt format. The Mistral 123B407 GGUF lacks a chat template, so the cached runs use the llama-2 fallback format. A native-format sensitivity run leaves the L1 change unresolved but lowers its quiz-weighted Static support interaction. The magnitudes thus depend on the prompt configuration (Appendix X.7).

Answer metrics. Our primary metric is **token-F1**, the harmonic mean of token precision and recall after SQuAD normalization, reported on a 0–100 scale. **Exact Match** is a lexical sensitivity check, and **choice-grounded accuracy** is a complementary endpoint: RealTimeQA is natively multiple-choice, so each free-form answer is mapped to its closest option by token-F1 (a distractor is credited on 0.9% of answers). **Answer coverage**, the fraction of retrieved documents whose full stored text contains the gold span, is a retrieval statistic rather than a support label. **Recall@5** mea-

sures the fraction of linked relevant documents retrieved in the top five, while **hit@5** records whether that set contains at least one linked document.

Token-F1 rewards surface overlap and penalises correct paraphrases (Bulian et al., 2022), so three annotators and a blind adjudicator audited the thresholded decision $\mathbf{1}[\mathrm{F1} \geq 0.5]$ on the unscaled 0–1 score: among answers judged correct it identifies 76% and misses 24% ([9.6, 38.3], Appendix T.1). On the same adjudicated labels and weights, human-correct answers scored below threshold comprise 14.4% of all Static outputs and 3.6% of all Fresh outputs. These are prevalences rather than conditional miss rates, and their difference remains unresolved. We report continuous token-F1 alongside choice-grounded accuracy.

Support labels. A question is *supported* when at least one shown passage contains (string label) or entails (entailment label) the gold answer. These labels use the answer and are computed offline. They are distinct from temporal *admissibility*, which depends only on passage and question dates. The NLI checkpoint is `MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli`. The premise is the raw first 600 passage characters, and the hypothesis is "The answer to [question] is [gold answer]." Input pairs are capped at 384 tokens. A passage is supported when entailment is the highest-probability class, rather than exceeding a 0.5 threshold. Any of the five passages can establish question support.

Statistics. Unless specified otherwise, bootstrap intervals are 95%, with 90% intervals used for equivalence checks. The 878 questions occur in 199 weekly quizzes, each entirely on one side of the boundary, so we block-bootstrap over quizzes, test interactions directly as **differences-in-differences** (Gelman and Stern, 2006), and correct for multiplicity within each family. A permutation null, cross-fitting and per-model minimum detectable effects provide additional checks. Model-level claims use $n{=}7$ inference over parameter-known rungs, separate from the quiz bootstrap. We documented the primary window, replication, context-budget and policy-selection protocols with their decision rules before their designated analyses. Later corrections and additional analyses are identified separately, with deviations recorded (Appendix X.1). We report the current offline replay outcomes separately (Appendix A).

Answer-bearing evidence. At least one own linked gold document contains the answer string for 52.3% of questions. This question-level incidence is neither passage-level coverage nor a ceiling for retrieval from the pooled corpus (Appendix J).

5 Results

5.1 L1: question-time validity

The boundary mixes question cohorts. The observed closed-book curve motivates the 2025 boundary: **eight of nine** generators dip in 2025 and recover in 2026, including Qwen2.5-0.5B, and the pattern survives removing each question's gold evidence (an observed swing of about 10 F1, Figure 5, Appendix R). Thus Δ_{cut} mixes closed-book answering with cohort difficulty. The small Qwen rungs score 5 to 18 F1 per year against an answer-permutation null below 1 F1, so their performance is inconsistent with that null. This test does not distinguish memorization from inference or question cues.

A conservative split changes estimates. The 2023/2026 pair spans the documented cutoff range, but some 2023 items follow GPT-4o-mini's October cutoff and seven generators disclose none, so it is a sensitivity contrast, not a controlled comparison. The pair reduces the Qwen2.5-32B contrast from 4.9 to **0.2** and the 72B contrast from 6.2 to 2.9, both intervals including zero (Table 22). The original 4.9 estimate was uncertain ($[-2.5, 11.7]$) and the change is unresolved: resampled jointly, the two contrasts differ by $+4.8$ $[-2.2, 12.1]$. In the sample, unadjusted intervals exclude zero only for Sonnet, Haiku, Mistral 123B and GPT-4o-mini.

Scale, abstention and robustness. The prefix detector flags at most 7% of open-weight closed-book outputs, but misses about half of human-labelled abstentions. Detected abstention rates rise for Claude on Fresh questions (34% → 75% for Sonnet, Appendix M), so we anchor the capability analysis on Mistral 123B. Conditional on the fixed ladder, joint quiz resampling gives **5.9 F1 per decade of parameters** $[3.1, 8.7]$ at the 2025 boundary. On the 2023/2026 contrast with equal quiz weighting, the slope is 3.46 under token-F1 and 2.44 under choice-grounded accuracy ($p{=}0.27$). The association depends on the endpoint and contrast (Appendix O). Cluster-robust inference with Benjamini–Hochberg (BH) correction identifies Mistral 123B and the two frontier mod-

Table 1: Paired effects of the three-day temporal mask. Unmasked retrieval gains and masking effects use quiz-weighted F1. † marks intervals excluding zero.

		Static		Fresh	
Generator	Par.	$\Delta^{\text{unm}}_{\text{RAG}}$	Δ_{mask}	$\Delta^{\text{unm}}_{\text{RAG}}$	Δ_{mask}
Qwen2.5	1.5B	$+14.1^{\dagger}$	$+1.0$	$+16.2^{\dagger}$	$+0.5$
Qwen2.5	3B	$+20.1^{\dagger}$	-0.2	$+16.1^{\dagger}$	$+0.7$
Qwen2.5	7B	$+9.1^{\dagger}$	$+0.3$	$+14.4^{\dagger}$	$+0.8^{\dagger}$
Llama-3	8B	$+3.7$	-0.5	$+11.8^{\dagger}$	-0.3
Qwen2.5	32B	$+5.2^{\dagger}$	-0.4	$+14.2^{\dagger}$	-0.7
Qwen2.5	72B	$+3.1$	-0.7	$+10.9^{\dagger}$	-0.4
Mistral 123B	123B	$-5.0^{\dagger}$	-0.1	$+9.3^{\dagger}$	$+0.2$
Haiku 4.5	n/a	$+0.0$	-0.1	$+17.7^{\dagger}$	$+0.1$
Sonnet 4.5	n/a	$-7.9^{\dagger}$	$+0.2$	$+16.3^{\dagger}$	$+1.9^{\dagger}$
GPT-4o-mini	n/a	$-7.4^{\dagger}$	-0.4	$+7.8^{\dagger}$	$+0.3$

els, and cross-fitting retains that set (Table 10). Effective cutoffs follow the data-age gradient (Cheng et al., 2024; Zhang et al., 2026) rather than a boundary step, which is why the 2025 boundary is descriptive (Appendices L and H).

Key finding. Closed-book contrasts depend on the calendar split and remain uncertain. They compare question cohorts without establishing knowledge cutoffs.

5.2 L2: evidence-time validity

Pooling exposes later evidence. The three-day grace convention requires $t_{\text{publish}}(d) \leq t_q + 3$ days. The index builder filtered documents against their source question before pooling, exposing other questions to later evidence. In the shipped Hybrid rankings $\mathbf{11.7}\%$ of contexts post-date the grace window, and 34.3% of questions receive at least one such passage (19.8% of Static, 4.8% of Fresh contexts).

Masking separates exposure from effect. Table 1 pairs each generator's masked and unmasked answers within question. Unmasked gains are relative to closed-book. The mask changes the ordered top-5 for 38.0% of questions. Masking before rank fusion can reorder even admissible lists. Yet 18 of 20 quiz-weighted F1 intervals include zero. Restricting to originally admissible contexts instead changes the question set (179/404 Static, 398/474 Fresh). With questions fixed, Static retrieval harm persists: 3 of 10 masked intervals stay below zero. Future evidence therefore cannot account for those generators' Static retrieval cost.

The grace window is a convention. The window study uses historical untruncated reruns for 1.5B, 3B and 7B, and frozen outputs where available for 32B and Mistral. Across 0, 1, 3 and 7 days, all 20 masked-minus-unmasked differences in quiz-weighted overall F1 lie inside a prespecified ± 2 F1 margin (90% intervals), though within Static or Fresh alone several do not fit (Table 31, Appendix X.4). A zero-day rule costs the two smallest generators 0.28 F1 (Holm $p \leq 0.04$) and up to 9.6 F1 on the 41 questions whose context set changes. This equivalence does not extend to the frozen 1.5B panel (90% interval $[-0.61, 2.35]$). The seven-day arm covers only the retained corpus.

Key finding. Temporal exposure is common, yet masking changes covered scores by at most 1.9 quiz-weighted F1. Overall equivalence need not hold within temporal subgroups.

5.3 L3: evidence-content validity

Table 2 reports question-pooled token-F1 losses on 474 Fresh and 404 Static questions, with 95% intervals resampling 199 weekly quizzes. The Fresh$-$Static difference averages $+21.3$ $[16.8, 25.7]$ across three generators.

A stale index costs about 21 Fresh F1. Three generators (3B–123B) answered the same 878 questions over the unmasked, *stale* pre-2025 and *leave-one-out* indices. The last removes each question's gold evidence. The stale index lowers Fresh F1 by 21.6, 21.7 and 21.3 points at 3B, 32B and 123B. Equality across scale was not tested. For the two larger models, stale-index retrieval scores *below* closed-book on Fresh questions by 7.8 $[4.0, 11.3]$ and 11.4 $[8.1, 14.9]$ F1 (-2.1 $[-5.8, 1.3]$ at 3B). The intervention changes the document pool, ranking and topical coverage together. Attributing the resulting harm to outdated facts requires a separate failure audit (Appendix Q.1). Static is a partly treated comparison: the stale index changes 184 of 404 Static context sets, yet Static F1 moves by less than a point. Leave-one-out lowers both regimes, by 18 to 22 F1.

Shown-text support locates retrieval gains. We partition each slice by whether the 600 characters shown per passage *mention* the gold answer (string label) or *entail* it (NLI label, undated hypothesis). Both arms are recomputed in each draw (Table 3, Section 3). Under the masked rankings the Static interaction excludes zero for *all* 10 generators under both labels (Table 3). Both labels match the shown text and retrieval condition (correction

Table 2: Stale-index intervention: Fresh and Static token-F1 losses and their difference, with quiz-bootstrap 95% intervals.

Generator	Fresh drop	Static drop	Fresh−Static
Qwen2.5-3B	$+21.6$ $[16.9, 26.2]$	$+0.7$ $[-0.8, 2.3]$	$+20.8$ $[15.9, 25.7]$
Qwen2.5-32B	$+21.7$ $[17.1, 26.2]$	-0.03 $[-1.3, 1.2]$	$+21.7$ $[16.9, 26.6]$
Mistral 123B	$+21.3$ $[16.6, 25.7]$	$+0.0$ $[-1.3, 1.4]$	$+21.2$ $[16.4, 25.8]$

Table 3: Static support interactions: masked automatic labels (600 characters) and unmasked human labels (700 characters).

		Masked (auto, 600 chars)		Unmasked
Generator	Par.	String	Entail.	Human
Qwen2.5	1.5B	$+48.9$	$+49.7$	$+55.0$
Qwen2.5	3B	$+49.0$	$+43.5$	$+55.2$
Qwen2.5	7B	$+48.9$	$+38.1$	$+41.8$
Llama-3	8B	$+38.0$	$+36.2$	$+41.8$
Qwen2.5	32B	$+44.7$	$+37.8$	$+54.7$
Qwen2.5	72B	$+46.6$	$+34.2$	$+44.1$
Mistral 123B	123B	$+37.7$	$+33.5$	$+30.6$
Haiku 4.5	n/a	$+35.8$	$+28.4$	$+38.7$
Sonnet 4.5	n/a	$+31.8$	$+34.4$	$+34.5$
GPT-4o-mini	n/a	$+38.6$	$+33.9$	$+45.0$

history in Appendix X.2). A separate open-weight judge (Qwen2.5-14B-Instruct) applying the annotators' rubric to the exact 600 characters yields the same direction: the interaction excludes zero for all ten generators in both conditions and is larger than under our labels. It receives the question and passage dates, although temporal relation validity was not evaluated. Shared model families may correlate judge and generator errors. Calibration uses legacy human labels (Appendix X.3). The support interaction also excludes zero under the choice-grounded endpoint (10 of 10) and in the joint mask-and-human estimand (630 passages, Krippendorff's α=0.933, 10 of 10 sample-level interactions, Appendix T.8).

Table 3 reports the support interaction Δ_{sup}: retrieval minus closed-book token-F1 (0–100), paired within question, on supported minus unsupported questions over the 404 Static questions. Arm means are question-pooled except in the design-weighted, adjudicated human column. Table 25 gives the 95% quiz-bootstrap intervals. Automatic columns use *masked* contexts with labels computed on the 600-character passage text the generator saw. The human column uses the *unmasked* rankings and 700-character annotator excerpts. It covers only 127 Static questions and up to three annotated passages per question–context instance, versus five passages for automatic labels. Thus "human-unsupported" means unsupported in that annotated subset. These columns are separate robustness checks, not a controlled comparison of labelers. Every interval excludes zero.

Human agreement uses longer excerpts. Annotators judged 700-character excerpts while generators saw 600, so the human statistics are agreement with legacy labels, not validation on the exact input (Table 30, Appendix X.3). On the shown text, Cohen's κ against human passage-support labels (Task B) is 0.65 $[0.55, 0.73]$ (string), 0.80 $[0.74, 0.85]$ (entailment) and 0.79 $[0.74, 0.84]$ (judge). The masked-context annotation values and intervals are in Appendix X.3. A paired comparison is needed to resolve the label ranking. Dropping the 11 rows whose answer appears only in characters 600–700 changes each κ by at most 0.03. On matched human-labelled questions, automatic estimates match or exceed the human one (32B Static: human 54.7, string 56.8). The larger human column of Table 3 therefore does not establish label noise. A dated hypothesis keeps every interaction positive but agrees less with the undated human labels (κ=0.52), so we name the label *answer entailment*. Both packs retain valid design weights from stratification on the original full-passage label, although 23 of 254 and 30 of 200 questions would change stratum at 600 characters (Appendix T.3).

Retrieved but hidden answers. Figure 2 separates gold answers shown in the 600-character prefix, retrieved but hidden, and absent from retrieved passages. For all ten generators, the 103 hidden-answer questions have negative retrieval gains, 45 to 57 F1 below the shown group. These groups differ in difficulty and retrieval rank, so the contrast is observational (Appendix X.8).

Context budget and replication. With untruncated inputs, the *prefix* arm keeps the five k=5 passages and adds five. Its Static interaction is 40.4 to 44.5 F1 against 47.5 to 48.9 at k=5, every interval excluding zero. Some questions change support membership (33 of 878 enter the supported arm, including 17 Static, and 0 leave). Yet holding the k=5 partition fixed yields no resolved interaction

Table 4: Date-policy selection: held-out audit-minus-naive F1 and retrieval call rates for the five generators whose actions change.

Generator	20-rep., cond.	5-rep., refit	Calls
Qwen2.5-72B	-0.8 $[-1.5, -0.2]$	-0.5 $[-2.8, 1.6]$	$1.00 \to 0.94$
Mistral 123B	$+4.3$ $[2.2, 6.6]$	$+3.4$ $[-1.3, 6.7]$	$0.60 \to 0.55$
Llama-3-8B	-0.23 $[-0.49, -0.04]$	0.00 $[-3.15, 0.65]$	$1.00 \to 0.99$
GPT-4o-mini	$+3.2$ $[0.6, 5.9]$	$+3.1$ $[-0.7, 7.8]$	$0.05 \to 0.53$
Sonnet 4.5	$+5.4$ $[1.9, 9.0]$	$+5.3$ $[-0.3, 10.3]$	$0.95 \to 0.54$

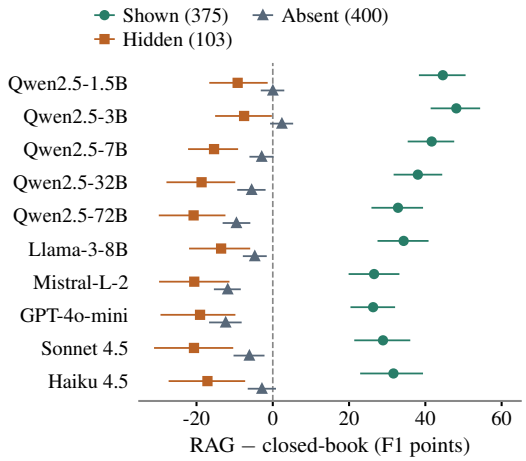


Figure 2: Retrieval gains by gold-string visibility in unmasked Hybrid contexts. Points show question-pooled means with quiz-bootstrap 95% intervals. The legend gives question counts.

difference (-0.1 to $+2.3$). The intervals also permit nonzero effects, so partition movement alone need not explain the difference (Appendix S).

On the 800 replication questions, quiz-weighted Δ_{sup} is 42.7 to 48.9 for four generators against 39.9 to 48.9 on the 532 common-year originals (Table 32). The pattern recurs, but stability within the prespecified margins holds only for Mistral's Δ_{mask}. Under question pooling, the Qwen2.5-3B interaction differs between samples ($+10.2$ $[0.2, 20.1]$, Appendix X.5).

Key finding. Index interventions cause large losses. Gains concentrate where shown text supports the answer. Gold-conditioned diagnostics locate gains without identifying evidence-use mechanisms.

5.4 Date-policy decisions

The exploratory *naive* policy selects closed-book or retrieval by development-quiz F1. The *audit* policy also selects a date-routing boundary. Four of nine decision-panel models retain identical held-out actions (Table 4). Conditional inference resolves gains for Mistral and Sonnet and losses for 72B and Llama after Holm correction. GPT-4o-mini's interval excludes zero, but its adjusted $p = 0.068$ does not pass. Refitting selection within each bootstrap leaves all adjusted $p \geq 0.54$. These Layer 1 policies use inspected data, including the full-data 2025 boundary. Cross-fitting retains this dependence. Masking changes at most 3.6% of actions, unresolved (Appendix X.6). On matched terms (same Fresh questions, three generators, question-pooled) the stale index costs 21.3 to 21.7 F1 while the mask moves scores by -0.2 to $+0.6$. This benchmark-specific contrast does not make temporal validity checks dispensable (Appendix Q).

Near-zero estimates in Tables 2 and 4 retain two decimals. Table 4 reports quiz-weighted F1 over both slices with masked Hybrid retrieval and 95% intervals over 199 quizzes. Conditional and selection-refit analyses use 20×5 and 5×5 folds, respectively, with 1,000 outer resamples for refitting. Calls give naive-to-audit retrieval rates. Table 33 lists all nine models.

Key finding. Date-policy gains remain unresolved after selection-refit inference and Holm correction. Model-specific trade-offs do not establish a deployment improvement.

6 Conclusion

TRAVA examines closed-book performance across a calendar split, retrieval within a publication window, and answer support in the shown text. In this corpus, stale-index and leave-one-out interventions have larger effects than temporal masking, while shown-text support consistently locates retrieval gains. The index interventions jointly change the document pool, ranking and coverage. Isolating recency or evidence use requires separate interventions. Temporal RAG reports should therefore name their date convention, show boundary sensitivity, distinguish retrieved from visible support, and report paired condition effects with uncertainty. Policy utility gains remain unresolved under selection-refit inference.

Limitations

Identification and generality. Layer 1 measures uncertain closed-book performance contrasts, not knowledge or memorization. We selected the shared boundary from the observed curve. Most cutoffs are undisclosed, and question composition and training differences remain confounded. Stale-index and leave-one-out interventions change document availability, ranking and topical coverage together. Static questions are partly treated. Their total effects do not isolate recency or semantic evidence use. The main three-layer audit and replication use RealTimeQA: the 800-question replication is disjoint by question but shares 172 of 182 quiz weeks with the working sample. It is not an independent temporal or cross-domain test. The supplementary FreshQA study examines only the closed-book contrast.

Labels and measurement. Layer 3 is an offline, gold-conditioned diagnostic. Answer mention and undated entailment do not establish date-resolved support, correctness, or evidence use. The existing human labels concern 700-character excerpts, although generators saw only 600 characters per passage. The exact-input packs (900 Task B and 630 masked rows) remain uncollected. The separate judge receives dates, but temporal relation validity was not evaluated. It shares a model family with much of the panel, and overlapping agreement intervals do not demonstrate superiority. Token-F1 also misses some human-correct answers. We report complementary endpoints and legacy-label sensitivity, not exact-input human validation.

Configuration and decisions. Window equivalence within ± 2 F1 applies to overall quiz-weighted estimates for the five-generator window study, which includes historical reruns. The frozen 1.5B panel and several subgroups do not meet that criterion. The seven-day arm uses the retained corpus. The original truncated $k{=}10$ result is withdrawn. Untruncated prefix and re-fused arms distinguish added context from order changes, but fixed-partition intervals do not exclude nonzero effects. Mistral's fallback prompt format affects its Static support interaction under quiz weighting. API prompt lengths use a proxy tokenizer. The exploratory date-policy study is internal validation on an already inspected dataset: selection-refit inference resolves no utility gain after multiplicity correction. The artifact is public, while exact-input human validation remains outstanding. Appendix Y gives the numerical details and Appendix A reports local validation outcomes.

Ethics Statement

RealTimeQA's English-language, Western-news focus limits representation. Its datasheet specifies CC BY-SA 4.0 for questions and annotations. Third-party article excerpts require separate redistribution consideration. The public artifact documents these separate terms in `DATA_LICENSE.md`. We retain source URLs and document dates for attribution and traceability.

The existing annotation study used blinded items, majority voting and separate adjudication. We report the unblinded 90-case author audit separately. All three annotators and the adjudicator volunteered without monetary compensation. The prepared exact-input packs remain uncollected. Open-weight models ran locally, and API outputs are cached. The released replay uses cached answers and model labels. Authoring tools assisted code and manuscript work. Detailed annotation, model-usage and provenance disclosures appear in Appendix Z.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, pages 9112–9141.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. Beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Yihang Chen, Pin Qian, Su Wang, Sipeng Zhang, Huan Xu, Shuhuai Lin, and Xinpeng Wei. 2026. Does RAG know when retrieval is wrong? Diagnosing context compliance under knowledge conflict. In *Proceedings of the 1st Workshop on Grounding Language Models: Learning Faithfully and Efficiently (GroundLM 2026)*, pages 263–276, Budapest, Hungary. Association for Computational Linguistics.

Ziyang Chen, Erxue Min, Xiang Zhao, Yunxin Li, Xin Jia, Jinzhi Liao, Jichao Li, Shuaiqiang Wang, Baotian Hu, and Dawei Yin. 2025. A question answering dataset for temporal-sensitive retrieval-augmented generation. *Scientific Data*, 12(1):1855.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated data: Tracing knowledge cutoffs in large language models. In *Conference on Language Modeling (COLM)*. ArXiv:2403.12958.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. 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*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural IR models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2353–2359. Association for Computing Machinery.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Matteo Gabburo, Nicolaas Paul Jedema, Siddhant Garg, Leonardo F. R. Ribeiro, and Alessandro Moschitti. 2024. Measuring retrieval complexity in question answering systems. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14636–14650, Bangkok, Thailand. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Andrew Gelman and Hal Stern. 2006. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4):328–331.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*, pages 43008–43029.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024.

Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 16867–16878. ELRA and ICCL.

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

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

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. RealTime QA: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pages 49025–49043.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Dayoon Ko, Jinyoung Kim, Hahyeon Choi, and Gunhee Kim. 2024. GrowOVER: How can LLMs adapt to growing real-world knowledge? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3282–3308, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research (TMLR)*.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! Analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Linguistics.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

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

Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. DYNAMICQA: Tracing internal knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14346–14360, Miami, Florida, USA. Association for Computational Linguistics.

Jannat Ara Meem, Muhammad Shihab Rashid, Yue Dong, and Vagelis Hristidis. 2024. PAT-Questions: A self-updating benchmark for present-anchored temporal question-answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13129–13148, Bangkok, Thailand. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, and Qi Liu. 2025. HoH: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6036–6063, Vienna, Austria. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024. BERGEN: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7640–7663, Miami, Florida, USA. Association for Computational Linguistics.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2025. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3697–3715, Abu Dhabi, UAE. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

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. 2024. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems (NeurIPS 2024), Datasets and Benchmarks Track*, volume 37, pages 21999–22027. Curran Associates, Inc.

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

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2395–2400, Washington, DC, USA. Association for Computing Machinery.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

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

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. Will LLMs replace knowledge graphs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 311–325.

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

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Md Nayem Uddin, Amir Saeidi, Divij Handa, Agastya Seth, Tran Cao Son, Eduardo Blanco, Steven Corman, and Chitta Baral. 2025. UnSeenTimeQA: Time-sensitive question-answering beyond LLMs' memorization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1873–1913, Vienna, Austria. Association for Computational Linguistics.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arık. 2025. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571, Vienna, Austria. Association for Computational Linguistics.

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

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. 2025. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations (ICLR)*, pages 91595–91631.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. 2025. AntiLeakBench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18403–18419, Vienna, Austria. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649. Association for Computing Machinery.

Yutong Yan, Raphael Tang, Zhenyu Gao, Wenxi Jiang, and Yao Lu. 2026. DatedGPT: Preventing lookahead bias in large language models with time-aware pretraining. *Preprint*, arXiv:2603.11838.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024. CRAG – comprehensive RAG benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 10470–10490.

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

Zhe Yu, Wenpeng Xing, and Meng Han. 2026. LatentAudit: Real-time white-box faithfulness monitoring for retrieval-augmented generation with verifiable deployment. *Preprint*, arXiv:2604.05358.

Yixiao Zeng, Tianyu Cao, Danqing Wang, Xinran Zhao, Zimeng Qiu, Morteza Ziyadi, Tongshuang Wu, and Lei Li. 2025. RARE: Retrieval-aware robustness evaluation for retrieval-augmented generation systems. *Preprint*, arXiv:2506.00789.

Michael Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Terry Jingchen Zhang, Gopal Dev, Ning Wang, Max Obreiter, Punya Syon Pandey, Keenan Samway, Wenyuan Jiang, Yinya Huang, Bernhard Schölkopf, Mrinmaya Sachan, and Zhijing Jin. 2026. Test of time: Rethinking temporal signal of benchmark contamination. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36538–36555, San Diego, California, United States. Association for Computational Linguistics.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6963–6975, Bangkok, Thailand. Association for Computational Linguistics.

A Reproducibility

The public artifact at `https://github.com/lexuanbach/TRAVA` contains code, cached outputs, annotations and reference results. It includes the following components:

- `configs/`: declarative specifications for the dataset sample (`datasets.yaml`), retrievers (`retrieval.yaml`), generators (`models.yaml`), and evaluation (`eval.yaml`).
- `pipeline/`: data acquisition, corpus construction, retrieval, generation, metric computation, aggregation, temporal controls, clustered inference, multiplicity correction, and power analysis. The main entry points are `download_data.py`, `build_dataset.py`, `build_corpus_full.py`, `retrieval.py`, `retrieval_full.py`, `generate.py`, `local_generate.py`, `gen_gguf_ranked.py`, `metrics.py`, and `aggregate.py`. The temporal analyses use `heldout_index.py`, `loo_control.py`, `robustness.py`, `honest_stats.py`, `cluster_robust.py`, `why_null.py`, and `rcb_all.py`.
- `results/`: one JSON file per analysis and `claim_ledger.csv`. Each ledger row names a result path and covers the main tables, temporal mask, support interactions, robustness analyses, stale and leave-one-out controls, and routing contrasts. Values outside the ledger's emitter set reproduce by running the corresponding analysis script over the shipped caches. The command `python pipeline/rebuild_ledger.py --check` compares the stored ledger with a fresh byte-for-byte rebuild.
- `data/processed/generations*/*.jsonl`: cached responses for every generator and condition, which support offline metric recomputation.

Exact configurations. The 32B, 72B, and 123B open-weight generators use 4-bit (`Q4_K_M` GGUF) inference on an Apple M-series Metal GPU through `llama-cpp-python` with full offload (`n_gpu_layers`$=-1$), `n_ctx`$=2048$, and `max_tokens`$=32$. The model identifiers are `Qwen/Qwen2.5-{32B,72B}-Instruct-GGUF` and `lmstudio-community/Mistral-Large-Instruct-2407-GGUF`. The Mistral GGUF file carries no `tokenizer.chat_template`, so `llama-cpp-python` rendered every Mistral prompt in its `llama-2` fallback format (`[INST] <<SYS>> ... [/INST]`), identically across conditions. A full rerun of the closed-book and masked-retrieval conditions in the model's own instruction format gives identical answers on 80% and 76% of questions, does not resolve a change in the Layer 1 contrast ($+0.9$ $[-1.3, 3.2]$), and lowers the Static support interaction under quiz weighting ($38.2 \rightarrow 30.7$). The change is -7.5 $[-14.0, -1.7]$, compared with the unresolved question-pooled change of -2.9 $[-7.9, 2.0]$. Qwen2.5 at 0.5B, 1.5B, 3B, and 7B runs on CPU through `transformers`. The Anthropic reference uses `claude-haiku-4-5-20251001` and `claude-sonnet-4-5-20250929` with `max_tokens`$=64$. Every generator uses temperature 0. GPT-4o-mini uses the OpenAI chat-completions API with the alias `gpt-4o-mini` and 32 output tokens. Its resolved serving snapshot was not retained. Llama-3-8B and Llama-3.1-8B use the NousResearch instruction checkpoints in bfloat16. Llama-3.2 at 1B and 3B uses the unsloth instruction checkpoints in bfloat16. The additional Llama-3.1 8B and 70B GGUF configurations use LM Studio community checkpoints at Q4_K_M precision. The inventory distinguishes these configurations and records their cache coverage.

Retriever checkpoints are `facebook/dpr-question_encoder-multiset-base` and

facebook/dpr-ctx_encoder-multiset-base, BAAI/bge-base-en-v1.5, naver/splade-cocondenser-ensembledistil, and colbert-ir/colbertv2.0 through pylate. BM25 uses k_1=1.5 and b=0.75. Hybrid uses RRF with k=60 over BM25 and BGE. All retrievers run on CPU with top-k=5. Resampling settings are analysis-specific. The paired mask uses 5,000 draws and seed 20260803, choice-grounded contrasts use 5,000 and 20260801, human metric errors use 5,000 and 20260802, and the primary window, replication and context-budget analyses use 5,000 and 20260917. Selection-refit inference uses 1,000 outer resamples. Legacy analyses retain their recorded settings, including 2,000 draws for honest_stats.py. Result records and source defaults specify each seed.

Cost. Main-corpus retrieval and the two API references complete in under one hour. The models up to 7B are CPU-bound. The 32B, 72B, and 123B conditions take about 0.5 to 4 hours each on one GPU. Building and searching the 26,611-document index adds about 20 minutes. The open-weight headline results, including the capability trend, the 123B gap, and the support interaction, reproduce offline from local caches. The Claude responses are also cached for reference and abstention analyses.

README.md gives installation and cached-analysis commands. To check distributed file hashes, run scripts/verify_snapshot.py. The scripts/reproduce.py entry point runs core or full analyses in an isolated copy. scripts/compare_results.py compares numerical outputs with the frozen reference results. These CPU workflows require no model downloads or API calls.

Pinned external dependencies. Entailment relabelling uses MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli. results/environment_manifest.json records the NLI revision and replay environment, not the original generation environment. The public docs/model_inventory.json records available generator identifiers and settings. Missing checkpoint revisions and API-served snapshot IDs are explicitly marked unavailable. Cached responses keep replay independent of live endpoints.

Prompt-length audit. The historical prompt-length audit re-tokenised the cached prompts after generation. The 36 locally run conditions have exact counts from the generator's own tokenizer or GGUF vocabulary. For the 14 API conditions, a Qwen2.5 tokenizer serves as a proxy. The provider-side model versions served at generation time were not recorded beyond the identifiers in configs/models.yaml. No local GGUF prompt exceeds 1,182 tokens against n_ctx=2048, and llama-cpp-python raises rather than truncates on overflow. The only truncated condition was the original k=10 run, which is withdrawn (Appendix S).

Correction manifest and replay. The support labels used at submission read more text than the generators saw. The retained revision record includes a per-analysis correction manifest covering all 36 support analyses: 22 recomputed on the 600-character shown text, 7 that remain agreement with legacy 700-character human labels, 6 unaffected, and 1 withdrawn. The public docs/corrections.md describes the measurement corrections, and results/analysis/CORRECTION_MANIFEST.json maps them to results. The cached-analysis runner records successful, failed and skipped stages separately. It replays analysis of stored answers and labels. It does not regenerate model responses, entailment labels or judge labels. The metric-error summaries share adjudicated labels, population weights, threshold and resamples. The population support-rate denominator excludes consistency-only cases.

Validation scope. The public release includes focused computational tests and a full cached-analysis suite. Its run summary gives the current test and stage counts, avoiding conflation with manuscript-dependent workspace tests. Regression checks cover visible-prefix support, masked passage alignment, context-budget prompt retention, adjudicated denominators, annotation units, capability centering, interval markers and reporting consistency. Fresh tokenization, model generation and human collection remain outside cached replay. Historical prompt-length results are retained, but a fresh tokenizer audit remains outstanding.

Release status and snapshot version. The corrected artifact is public at the page-one link, with an MIT license for code and separate data terms. It contains frozen inputs, reference results, reproduction commands and SHA-256 hashes. Manuscript sources, reviewer files and internal audits remain outside the public package. The expanded ledger

contains 36,311 rows: historical support-dependent claims are marked superseded, and corrected numeric entries resolve through exact JSON pointers. Public availability does not establish additional rights to third-party article excerpts, whose redistribution basis remains unconfirmed. Exact-input human collection and fresh tokenizer validation also remain outstanding.

TRAVA names the protocol, and `v2026-07` names the benchmark snapshot used here. The snapshot includes RealTimeQA quizzes through 2026-07-03, the last quiz in the question set, and fixes the question set, corpus, and qrels. The 40 questions of Appendix N.2 have question dates 2026-07-09, 2026-07-10, and 2026-07-17 across two weekly quiz IDs and are held in a separate file. Later snapshots can append weekly slices while preserving earlier ones. The command `python pipeline/rebuild_ledger.py --check` validates the claim ledger against this snapshot and returns a nonzero status when a source value changes.

Snapshot updates. Each release carries a snapshot tag such as `v2026-07`. Updating the benchmark follows five steps:

1. ingest new weekly RealTimeQA quizzes and linked documents
2. append dated documents while preserving existing `doc_ids`
3. retrieve evidence for new questions under the per-question mask $t_{\text{pub}} \leq t_q + 3$ days
4. generate closed-book and RAG answers for the new questions
5. assign the new snapshot tag

Only the new slice incurs generation and retrieval cost. The fixed analysis boundary and snapshot tag preserve reproducibility for earlier releases. This specified update procedure defines the protocol for future public snapshots.

Sampling rationale and balance. The 878-question working set contains all 346 questions from 2026 and about 123 to 150 questions from each earlier year. This allocation keeps the per-year cells informative while bounding generation cost across the analysis panels. The sampling cap also makes the 2024 cell influential in the shared-boundary contrast. We therefore report the uncapped set of all 4,176 questions that pass the same filters for Qwen2.5-32B, Qwen2.5-72B, and GPT-4o-mini, using cached API outputs for the last model. No API calls are needed for this replay. The current full-question file has 4,176 rows and 4,176 unique IDs. Recounting from the file corrects an older metadata summary that reported 4,182 after overcounting 2023 by six. Both manifests use the same snapshot tag. We use the uncapped analysis to interpret the 32B and 72B results in Section 5.1.

Independent check of the conservative contrast. Two implementations compute the 2023 against 2026 gap: `pipeline/cluster_robust.py` uses cluster-robust standard errors, and `pipeline/verified_slice.py` uses the quiz block bootstrap. Their point estimates agree within 0.01 F1 for all nine generators. Two-sided p-values differ by at most 0.045, with the largest difference near p=0.99. We report the bootstrap estimate and use the independent implementation as a consistency check.

B Prompts

We use a single fixed prompt for all conditions (verbatim from `configs/models.yaml`), varying only whether evidence documents are included.

System instruction. *"You are a precise question-answering system. Answer the question with the shortest possible span (a name, number, date, or short phrase). Do not explain. If evidence documents are provided, answer using them."*

Closed-book user message. *"Question (t_i): {question} Answer with the shortest exact span."*

Retrieval-augmented user message. The top-k=5 retrieved passages, each prefixed with its index and publish date *"[j] ($t_{publish}$) {passage}"*, followed by *"Question (t_i): {question} Answer with the shortest exact span, using the documents."*

Generation is deterministic (temperature 0), capped at 64 new tokens for the Claude API models and 32 for GPT-4o-mini and local generators. The recorded settings and their sources appear in `docs/model_inventory.json`. The short-span instruction does not ensure reliable abstention detection. Appendix T.5 reports detector recall. We treat answer coverage as a retrieval statistic. Appendix T.3 reports its passage-level validation by comparing both automatic support labels with human labels.

C Protocol assumptions

This appendix records the two procedural guarantees the evaluation relies on. They define how we conduct and replay the run.

Fixed replay trace. For every example i and system s, the run stores a deterministic record (x_i, y_{is}, z_{is}) containing the input, the model output, and the retrieved document ids. Every metric in the paper is recomputed offline from these records at temperature 0 with all responses cached. The `results/claim_ledger.csv` file maps each reported number to the record set it is computed from. This makes every aggregate auditable to example-level evidence and reproducible without re-querying any model.

Timestamp invariant and scope. For each item i we record the question date t_i and, for every evidence document d *collected for* q_i, its publish date $t_{\text{publish}}(d)$. A document enters the pool only when $t_{\text{publish}}(d) \leq t_i + 3$ days.

This invariant governs admission to the pool. It constrains each document against its source question. A shared index can then rank that document for another question q_j with an earlier date. Section 5.2 measures this exposure in the shipped rankings at 11.7% of Hybrid contexts and 34.3% of questions. We then re-run retrieval under the per-query mask $t_{\text{publish}}(d) \leq t_q + 3$ days. The full index (Appendix K) applies this mask at ranking time for every retrieved passage. The held-out condition in Appendix N.2 applies a stronger restriction to documents published before the boundary.

C.1 Per-generator cutoff evidence

We store cutoff evidence per generator to calibrate the shared calendar boundary. A reported "cutoff" may denote the end of *training data*, the end of a vendor's *reliable knowledge* range, or the model's *release*. These dates can differ by six months. The reported training-data date is a boundary on documented data inclusion, not an observed boundary on model knowledge.

Provider evidence is recorded in `results/verified_slice.json` and `docs/provider_cutoffs.md`. The original nine-model cutoff panel has seven undisclosed cutoffs. Llama and GPT-4o-mini are additional controls. Meta's cards identify the Llama dates as pretraining-data cutoffs, whereas OpenAI labels its date a knowledge cutoff.

Among 474 Fresh questions, 128 are dated in 2025 and 95 fall through July. Anthropic's model specifications give July 2025 as the training-data cutoff for both Claude models, with reliable knowledge through January for Sonnet and February for Haiku. Haiku's October 2025 system card separately describes public-Internet data through February. We retain both records and use the later overall date for the conservative boundary analysis. These dates do not establish that any particular question appeared in training. The February alternative covers 23 Fresh question dates and remains a sensitivity analysis in `results/cutoff_relative.json`.

Table 5: Provider dates. Training-data and reliable-knowledge dates are distinct. A dash means no separate date is disclosed.

Generator	Data through	Reliable knowledge
Qwen2.5 (all 6)	undisclosed	—
Mistral 123B	undisclosed	—
GPT-4o-mini	—	2023-10-01
Llama-3-8B	2023-03	—
Llama-3.1/3.2 controls	2023-12	—
Sonnet 4.5	2025-07	2025-01
Haiku 4.5	2025-07	2025-02

We use the 2025 boundary descriptively and the **2023 against 2026** contrast as sensitivity to a split spanning the documented dates (Section 5.1). Some 2023 questions follow GPT-4o-mini's October cutoff, so its earlier slice is not purely pre-cutoff. Undisclosed cutoffs remain unknown.

C.2 Glossary

Boundary: the analyst-chosen date partitioning evaluation questions. *Provider cutoff*: a provider-reported upper date for some training data. *Static*/*Fresh*: before and on or after the chosen boundary. *Admissible*: the passage's publish date is no later than its question's date plus three days, a chosen publication-lag convention. Admissibility under this window is neither guaranteed availability at question time nor content currency. A passage can be admissible and later become outdated. *Supported*: at least one shown passage contains or entails the gold answer. This label is gold-conditioned. *Future* means published after the query, and *superseded* means once-current and later replaced. The stale index is a separate pre-2025 intervention. *Rung* is reserved for models with a published parameter count, *panel* for any analysis subset (Appendix U).

D Item-level model

We fit an item-level model as a directional check on the capability association in Section 5.1. This appendix gives its specification and inferential scope.

Model. One row per (question, generator) closed-book answer, over the six open-weight rungs in the fit: Qwen2.5 at 1.5B, 3B, 7B, 32B, and 72B plus Mistral 123B. Qwen2.5-0.5B sits at the closed-book floor on both slices, so the capability fit begins at 1.5B:

$$\mathrm{F1}_{ij} = \beta_0 + \beta_1 f_i + \beta_2 c_j + \beta_3 f_i c_j + u_i + \varepsilon_{ij}.$$

Here f_i indicates a Fresh question, c_j is centered capability, and u_i is a random intercept per question. We fit the model by restricted maximum likelihood (REML). At capability c, the fitted Static-minus-Fresh contrast is

$$D(c) = -(\beta_1 + \beta_3 c).$$

At $c = 0$ this reduces to $-\beta_1$. A negative β_3 means that the contrast widens as capability increases. The random intercept accounts for repeated measurement of the same question across generators. Static/Fresh difficulty remains a between-question quantity because f_i is a deterministic function of i. Estimating β_1 therefore uses the random-effects independence assumption. The 2023 against 2026 contrast tests boundary sensitivity without controlling question difficulty.

Capability proxy. Let h_j be generator j's mean token-F1 on the *Fresh* slice under Hybrid retrieval. The implementation uses

$$c_j = h_j - 36.4664.$$

The centering constant is the recorded mean across the eight models below, including GPT-4o-mini and Sonnet, although the primary fit uses six open-weight models. Thus $-\beta_1$ refers to that eight-model mean, not the six-model mean. The proxy is measured under retrieval and combines closed-book performance with the availability and use of evidence. We use it as a descriptive covariate. Measuring evidence-use ability directly requires separating those components. For Sonnet, 95 of 474 Fresh items fall through its July 2025 training-data date, and abstention lowers the proxy because it scores answered and declined questions together. Sonnet 4.5 therefore scores 30.4, below Qwen2.5-3B at 35.2. We fit the reported model on open-weight rungs, where detected abstention remains below 7%, and use the all-model fit to show the effect of this proxy choice.

Results. The interaction is -7.86 F1 per decade of parameters and -2.24 per point of the capability proxy. These negative Fresh interactions imply a widening Static-minus-Fresh gap, the same direction as the positive 5.9 F1 scale slope. We report these as directional point estimates because model-level information comes from six rungs.

Inference level. Capability varies between generators, which makes six rungs the effective sample size for β_3. Additional question rows improve within-model precision while leaving the model-level sample unchanged. We therefore use the item-level fit to check the direction of the aggregate slope after accounting for repeated questions. Model-level inference remains the exact rung permutation in the main text. The leave-one-model-out refits below are likewise interpreted by sign.

Leave-one-model-out. The direct test of whether one rung carries the association is to drop each in turn:

Refit (drop)	β_3
Qwen2.5-1.5B	-2.81
Qwen2.5-3B	-2.08
Qwen2.5-7B	-2.24
Qwen2.5-32B	-2.57
Qwen2.5-72B	-2.41
Mistral 123B	-1.46

The interaction stays negative in every refit, including the one that drops Mistral 123B. The item-level direction therefore persists across every leave-one-model-out panel. Table 22 reports separate per-generator boundary contrasts with uncertainty. The two summaries address different inferential levels.

Capability proxy values. The following means define the eight-model centering set. They summarize Fresh retrieval performance, which combines answer quality and refusals. The lower Sonnet value illustrates why these means cannot rank latent capability.

Generator	Fresh RAG-F1
Qwen2.5-1.5B	33.49
Qwen2.5-3B	35.21
Qwen2.5-7B	35.92
Qwen2.5-32B	39.11
Qwen2.5-72B	38.96
Mistral 123B	39.77
GPT-4o-mini	38.84
Sonnet 4.5	30.44

Table 6: Question-pooled policies with unmasked Hybrid retrieval (token-F1). Bold: best deployable policy. "head" and Δt: oracle margins.

	Deployable			Oracle		Margin	
Generator	never	alw.	abst.	temp.	per-q	head	Δt
Qwen2.5-1.5B	16.1	**34.1**	16.1	25.1	40.1	6.0	−9.0
Qwen2.5-3B	14.8	**35.5**	16.9	25.2	41.4	5.9	−10.3
Qwen2.5-7B	21.6	**36.3**	21.8	29.7	44.4	8.1	−6.6
Qwen2.5-32B	27.5	**39.0**	27.9	35.0	49.9	10.9	−4.0
Qwen2.5-72B	32.3	**39.5**	32.3	37.4	51.5	12.0	−2.1
Mistral 123B	37.2	**40.8**	37.5	42.6	54.0	13.2	+1.8
GPT-4o-mini	36.0	**39.4**	36.6	40.9	53.3	13.9	+1.5
Sonnet 4.5	25.1	32.2	**39.5**	34.5	43.5	4.0	−5.0

Table 7: **Automatic stale-index failure taxonomy.** Table 8 reports the manual audit.

Automatic label (% of failures)	3B	32B	123B
Traceable to context*	56.6	53.6	37.2
Answer present, model missed	32.2	24.3	21.9
Declined to answer	1.4	15.7	30.7
Untraceable to context	9.8	6.4	10.2
*audited precision (%)	24	30	47
corrected: faithful stale read	13.7	15.9	17.4

E Routing and stale-index failures

Tables 7 and 8 distinguish automatic failure categories from the manual check of one category. The first table's percentages cover all stale-index failures for each model, whereas the second counts only sampled failures labeled traceable to context. Its precision is the fraction judged once correct, not agreement on all failure types. Multiplying the automatic traceable fraction by that precision yields the corrected faithful-stale-read row. Echoes outnumber superseded facts in the pooled audit, so lexical traceability alone overstates faithful use of outdated facts (Appendix Q.1).

Table 6 compares question-pooled routing policies. Appendix K defines its policies and oracle margins, then examines the larger-index control. Appendix S.2 evaluates cross-fitted pre-answer gates.

Table 8: **Audited stale-evidence labels.** The lexical label identifies superseded evidence in 30 of 90 sampled cases.

Adjudicated verdict	3B	32B	123B	all
superseded (once correct)	8	8	14	**30**
echo (never correct)	16	15	10	41
type error	7	2	1	10
non-answer	2	2	5	9
sampled	33	27	30	90
precision (%)	24	30	47	33

F Pre-answer routing

Repeated two-fold cross-fits select the routing threshold. Four unadjusted intervals exclude zero, and three remain significant after Holm adjustment across nine generators. The per-question ceiling selects the better cached answer for each question.

Table 9 is generated directly from `results/routing_baselines.json` by `pipeline/routing_baselines.py`. Three design properties determine its interpretation.

The table's rate is the fraction of questions sent to retrieval, and thresh. is the range of selected lexical-overlap thresholds across cross-fits. Ceiling selects the better cached answer using gold outcomes, which are unavailable to a routing policy. The gap between overlap and ceiling shows that most of this retrospective opportunity remains unrecovered, even where overlap improves over the better fixed policy.

The reference is the better fixed policy, chosen per generator. We compare each gate with the stronger fixed policy. For GPT-4o-mini, never-retrieve scores 36.2, compared with 34.5 for always-retrieve. Relative to the stronger reference, the overlap gate gains +2.20 with an interval that includes zero.

The contrast is paired. Both policies are scored on the same question from the same cached generation, so the paired interval directly estimates the policy difference. Under the paired quiz-clustered bootstrap, the gate's advantage is +0.14 at 7B. Its tuned threshold retrieves 94% of questions, which closely approximates always-retrieve.

Thresholds are cross-fitted, not chosen once. Over 20 repeated two-fold splits the selected overlap threshold varies (column *thresh.*), and with it the retrieve rate. Reporting a single split's threshold would present one draw from that spread as the policy.

Thresholds are fitted and evaluated within one snapshot. The two post-snapshot RealTimeQA weekly clusters in Appendix N.2 provide a small temporal check. The routing estimates therefore describe this snapshot.

Table 6 pools questions, whereas Table 9 weights quizzes equally. Both use unmasked Hybrid retrieval. Changing weights reverses GPT-4o-mini's best fixed policy: always-retrieve wins under question pooling, while never-retrieve wins under equal quiz weighting.

Table 9: Answer-independent routing with quiz-weighted F1 on unmasked rankings. Paired 95% intervals compare overlap routing with the better fixed policy. Stars mark Holm-adjusted $p < 0.05$.

	Policy mean (F1)			Overlap − reference (paired)			Cross-fit		
Generator	never	always	overlap	Δ	95% CI	p_{Holm}	thresh.	rate	ceiling
Qwen2.5-1.5B	16.2	31.0	30.9	−0.10	$[-0.28, 0.06]$	0.965	0.00–0.65	0.98	37.8
Qwen2.5-3B	13.6	32.2	32.1	−0.07	$[-0.21, 0.07]$	0.965	0.00–0.65	0.97	37.3
Qwen2.5-7B	20.9	32.0	32.1	+0.14	$[-0.33, 0.62]$	0.965	0.50–0.70	0.94	40.2
Qwen2.5-32B	26.2	34.8	35.8	+1.02	$[-0.21, 2.33]$	0.530	0.50–0.80	0.81	45.9
Qwen2.5-72B	29.7	35.7	38.3	+2.66*	$[1.13, 4.35]$	0.007	0.65–0.80	0.80	46.7
Llama-3-8B	24.8	31.5	33.1	+1.59	$[0.18, 3.19]$	0.170	0.50–0.75	0.84	42.3
Mistral 123B	36.2	36.5	40.0	+3.45*	$[1.22, 5.76]$	0.034	0.65–0.80	0.71	51.1
GPT-4o-mini	36.2	34.5	38.4	+2.20	$[-1.42, 5.79]$	0.965	0.65–0.80	0.75	50.5
Claude Sonnet 4.5	27.1	28.2	31.2	+3.05*	$[1.23, 5.12]$	0.007	0.60–0.80	0.76	40.8

G Retrieval across date slices

We define the slice-by-condition interaction as the Fresh retrieval gain minus the Static retrieval gain, following Gelman and Stern (2006). The question-pooled estimates are 13.8 for Mistral 123B, 14.4 for Haiku and 22.4 for Sonnet. Their 95% quiz-cluster bootstrap intervals are $[6.8, 20.5]$, $[4.4, 23.8]$ and $[13.4, 30.6]$, respectively. The corresponding two-sided tail probabilities are below 0.001, 0.006 and below 0.001, all retained by BH and Bonferroni adjustment over these five models. The 32B and 72B interactions are 5.2 $[-2.6, 13.1]$ and 5.0 $[-2.8, 12.5]$. These estimates use `results/cluster_robust.json`.

H Controlled-cutoff comparison

Table 19 is generated from `results/controlled_cutoff.json`. In Arm A, Δ is Llama-3.1-8B minus Llama-3-8B token-F1 on the same questions, with n counting questions in each date group. All three intervals include zero, and the between-cutoff point does not exceed both comparison groups. Arm B instead reports each model's pre-minus-post contrast around its disclosed cutoff, with a guard band. These are different estimands. Four points on the design follow.

The guard band matches the rest of the paper. A provider cutoff is a claim, not a measurement, so items within three months of it are dropped as ambiguous. Both `pipeline/cutoff_relative.py` and `pipeline/controlled_cutoff.py` use this rule. Using a hard split here instead would have reported Llama-3-8B's own-cutoff gap as 16.7 against the 9.5 the cutoff-relative analysis reports for the same estimand. A regression test confirms that the two scripts agree.

Arm A has defined resolution. With 93 questions in 37 quizzes, the paired contrast detects 7.7 F1 at 80% power. This limits the analysis to relatively large performance differences at the disclosed date. The pre-both and post-both date groups provide controls for the between-cutoff group, allowing release differences to appear on either side.

Arm B uses quiz-clustered inference. The three bf16 rungs (1B, 3B and 8B) give the exact permutation test a resolution floor of $2/3! = 0.33$, above conventional significance thresholds. We use the quiz-clustered bootstrap as the primary analysis and report the permutation result with its resolution floor.

What each arm still confounds. Arm A's two models are different releases, so every cell includes a quality difference. The estimand therefore uses a difference-in-differences against both controls. Arm B holds cutoff, family, and numeric precision fixed for those three rungs. The 4-bit 70B row is reported separately and excluded from the slope. Its 1B and 3B rungs are distilled from larger Llama-3.1 models, while the 8B rung uses a different training procedure. Both arms are observational, and the disclosed cutoff remains a provider claim.

I Additional tables

Table 10: Closed-book boundary contrasts under permutation, cross-fitted and cluster-robust inference for five generators.

Generator	Δ	Perm. p	X-fit p	Clust. p
Qwen2.5-32B	+4.9	.195	.077	.198
Qwen2.5-72B	+6.2	.093	.063	.077
Mistral 123B	+16.0	.005	<.001	<.001†
Haiku 4.5	+17.3	.001	<.001	<.001†
Sonnet 4.5	+26.2	<.001	<.001	<.001†

Appendix E defines the routing reference policies, and Appendix G reports retrieval interactions

across date slices.

Table 10 checks whether the reported closed-book contrasts depend on the inference procedure. Here Δ is Static minus Fresh token-F1, and Perm., X-fit and Clust. denote permutation, cross-fitted and cluster-robust p values. Daggers mark cluster-robust results retained after BH correction. The 32B and 72B estimates remain unresolved across these procedures, whereas Mistral and the two Claude references retain positive contrasts. This is robustness of the performance contrast, not control of question difficulty.

Table 11 separates question counts by year from corpus sizes. The larger 2026 cell reflects inclusion of all available questions that year, whereas earlier years are sampled. Corpus totals describe the two indexes, not documents assigned separately to each year. Consequently, the shared boundary compares differently composed year mixtures.

Table 11: TRAVA composition from RealTimeQA, June 2022 to July 2026. Results use the shared 2025 boundary (404/474). The calendar split describes composition.

Question year	Questions
2022	150
2023 / 2024 / 2025	131 / 123 / 128
2026	346
Total	**878**
Index	**Documents**
Main	6,692
Full	26,611
Added distractors	19,919

Table 12: Question-pooled retrieval quality at the 2025 boundary. S / F gives Static / Fresh hit@5. Oracle definitions appear in Appendix J.

	Main (6.7k)			Full (26.6k)	
Retriever	R@5	hit@5	S / F	hit@5	S / F
DPR	0.204	0.662	.66 / .66	n/a	n/a
BGE	0.285	0.788	.79 / .79	0.711	.73 / .69
ColBERT	0.292	0.801	.81 / .80	n/a	n/a
SPLADE	0.300	0.798	.79 / .81	n/a	n/a
Hybrid	0.303	**0.821**	.82 / .82	**0.763**	.76 / .77
BM25	**0.305**	0.817	.81 / .82	0.746	.75 / .75
Gold-Ev.	0.623	1.000	1.00 / 1.00	1.000	1.00 / 1.00

Table 12 distinguishes retrieving any linked document (hit@5) from retrieving the full linked set (R@5). Hybrid has the highest main-index hit@5, whereas BM25 has slightly higher R@5. Their differing rankings reflect the distinction between finding one linked document and the full set. Adding distractors lowers hit@5 for the re-run retrievers. By construction, Gold-Evidence selects linked documents. The separate own-linked answer incidence is 52.3% overall (49% Static, 55% Fresh). It records whether any linked document contains the answer. Neither measures answer visibility in the shown prefix.

Table 13: Qwen2.5-32B token-F1 by year. Hybrid-LOO removes each question's own linked evidence and scores highest in 2026.

System	2022	2023	2024	2025	2026
Closed-book	34.4	27.3	27.9	19.9	27.1
BM25	43.0	35.8	28.7	28.6	38.4
ColBERT v2	46.5	37.6	30.0	29.0	41.6
Hybrid	46.4	38.5	29.9	30.8	42.2
Gold-Ev.	43.3	35.5	28.3	32.2	37.3
Hybrid-LOO	20.0	16.1	16.4	13.9	23.8

Table 13 holds the generator fixed and breaks token-F1 down by question year. The 2025 trough and 2026 rebound persist in Hybrid-LOO, which removes each question's linked evidence. Year variation therefore survives that intervention. The residual combines closed-book answering and remaining evidence rather than isolating intrinsic question difficulty.

Multiple-choice scoring. Table 14 applies the same nearest-option rule to every cached free-form answer. Exact and unique containment matches precede token-F1 matching. Ties and answers matching no option score zero. Thus 25% is the uniform-choice reference, not a floor for these free-form responses. Small-model unmatched answers and frontier refusals can both yield sub-chance accuracy. This scoring rule does not simulate an option-conditioned prompt. Closed-book accuracy falls from 33.7 to 27.0 for 32B and 50.7 to 33.8 for Mistral. The public `results/reporting_summary.json` records counts, detected abstentions and unmatched answers separately.

J Gold-evidence reference

"Gold-Evidence" contains a question's own evidence, unranked and truncated to k=5. It is a provenance reference rather than an upper bound, and Hybrid scores higher at 38.8 versus 36.2. Restricting each gold set to documents that contain the gold span explains this result. The filter leaves 47.7% **of questions without answer-bearing evidence**, because RealTimeQA snippets often provide background rather than the answer. This source-data property bounds our recall and difficulty-control

Table 14: Question-pooled nearest-option accuracy at the 2025 boundary. Each available condition covers 404 Static and 474 Fresh questions.

	Closed-book		Unmasked Hybrid	
Generator	Static	Fresh	Static	Fresh
Qwen2.5-0.5B	10.6	9.9	n/a	n/a
Qwen2.5-1.5B	14.6	17.3	39.6	38.0
Qwen2.5-3B	13.9	17.1	40.8	39.7
Qwen2.5-7B	24.0	23.2	41.8	42.2
Qwen2.5-32B	33.7	27.0	45.3	45.8
Qwen2.5-72B	40.3	32.5	46.5	44.1
Mistral 123B	50.7	33.8	48.8	46.8
Haiku 4.5	37.9	19.6	44.3	42.0
Sonnet 4.5	51.7	18.8	45.5	42.4
Llama-3-8B	30.7	24.1	39.1	42.0
GPT-4o-mini	47.8	32.9	48.3	47.5

Table 15: **Full** 26,611**-document index.** Both re-run retrievers still beat closed-book, so the main index's small size is not what produces the retrieval gains.

	Index	Static		Fresh	
		F1	hit@5	F1	hit@5
Hybrid	Main (6.7k)	38.8	0.82	39.1	0.82
	Full (26.6k)	41.2	0.76	38.8	0.77
	Δ	$+2.4^{ns}$		-0.3^{ns}	
BM25	Main (6.7k)	36.3	0.81	35.8	0.82
	Full (26.6k)	39.0	0.75	34.2	0.75
	Δ	$+2.7^{ns}$		-1.6^{ns}	

interpretations, as reflected by answer coverage of 0.22 to 0.28. On the 52.3% of questions with answer-bearing evidence, the filtered oracle scores 64.0 F1 versus 34.3 closed-book on that same subset. These subset scores cannot be subtracted from the all-Static Gold-Evidence score. Gold-Evidence is a provenance baseline, not a ceiling.

In Table 15, Δ is full minus main index F1 and "ns" denotes a two-sided paired item-bootstrap $p \geq 0.05$ (2,000 draws). Hit@5 falls while F1 changes remain unresolved for both retrievers and slices. This dissociation cautions against reading a change in linked-document retrieval as an equal change in answer quality, without establishing equivalence between the indexes.

Table 16 fixes Qwen2.5-32B and compares six retrievers. S and F denote Static and Fresh, with R−CB subtracting closed-book F1 within each slice. All six point estimates improve over closed-book, with larger gains on Fresh questions. Hybrid leads Static and BGE narrowly leads Fresh, although paired between-retriever intervals would be needed to establish superiority.

Table 17 compares Hybrid retrieval for Qwen2.5-32B under three indexes. A stale (pre-2025) index changes Static by less than one point and drops

Table 16: Token-F1 by retriever and temporal slice. R−CB gives gains over closed-book. Gold-Evidence is a provenance reference.

System	Static	Fresh	R−CB (S)	R−CB (F)
Closed-book	30.1	25.2	n/a	n/a
BM25	36.3	35.8	+6.2	+10.6
DPR	34.7	34.0	+4.6	+8.8
SPLADE	38.0	38.0	+7.9	+12.8
ColBERT v2	38.6	38.2	+8.5	+13.0
BGE	36.5	**39.2**	+6.4	**+14.0**
Hybrid (RRF)	**38.8**	39.1	**+8.7**	+13.9
Gold-Evidence	36.2	35.9	+6.1	+10.7

Table 17: Qwen2.5-32B under main, stale and leave-one-out indexes. Token-F1 and retrieval hit@5 by temporal slice.

Index	Static	Fresh	Fresh hit@5
Closed-book	30.1	25.2	n/a
Main (6,692 docs)	38.8	39.1	0.82
Stale (pre-2025 docs)	38.9	17.4	0.61
Leave-one-out (own gold)	17.6	21.1	0.00

Fresh below closed-book (§5.3). Leave-one-out removal of gold evidence lowers both slices. We assess uncertainty by paired bootstrap.

Abstention detail

Whether a model abstains or guesses when it lacks knowledge changes the meaning of a cutoff gap. The detector flags abstentions in open-weight outputs, including Mistral 123B, on at most 7% of closed-book questions. The corresponding rates for the 123B model are 0.0% and 0.4%. The human audit finds incomplete detector recall (Appendix T.5), so they do not bound true abstention. Claude models instead show higher abstention. The closed-book rate rises from 50% to 76% for Haiku and from 34% to 75% for Sonnet. Among answered questions, Haiku remains similar across the boundary at 64 Static versus 61 Fresh (Appendix V).

The frontier aggregate gap therefore includes a strong abstention component, whereas Mistral 123B provides a contrast with few detected refusals. We anchor the capability claim on this 123B open model. The 0.5B model has no complete matching retrieval run and is excluded from the Hybrid panel.

Table 18 shows why pooled retrieval gains can conceal opposing slice effects. Mistral, GPT-4o-mini and Sonnet have negative Static and positive Fresh gains, with pooled intervals including zero. Pooled here combines questions through quiz weighting rather than averaging the two displayed slice means. Bold denotes interval exclusion, not

Table 18: **Retrieval minus closed-book by temporal regime**, paired, quiz-weighted, *unmasked* ranking, ten generators. Bold: 95% quiz-bootstrap intervals excluding zero.

Generator	Static	Fresh	Pooled
Qwen2.5-1.5B	**+14.1** [8.5, 19.5]	**+16.2** [9.2, 23.3]	**+14.9** [10.4, 19.2]
Qwen2.5-3B	**+20.1** [15.2, 25.4]	**+16.1** [9.9, 22.6]	**+18.6** [14.5, 22.6]
Qwen2.5-7B	**+9.1** [4.5, 13.9]	**+14.4** [7.3, 21.9]	**+11.1** [6.9, 15.0]
Llama-3-8B	+3.7 [−1.3, 8.7]	**+11.8** [5.2, 18.7]	**+6.7** [2.5, 10.8]
Qwen2.5-32B	**+5.2** [0.1, 10.2]	**+14.2** [7.2, 21.6]	**+8.6** [4.2, 12.9]
Qwen2.5-72B	+3.1 [−2.1, 8.2]	**+10.9** [4.5, 17.8]	**+6.0** [1.8, 10.2]
Mistral 123B	**−5.0** [−9.9, −0.3]	**+9.3** [1.6, 17.5]	+0.3 [−4.3, 4.6]
GPT-4o-mini	**−7.4** [−12.6, −2.2]	**+7.8** [0.2, 15.7]	−1.7 [−6.3, 2.7]
Sonnet 4.5	**−7.9** [−13.8, −2.1]	**+16.3** [10.9, 22.0]	+1.1 [−3.5, 5.4]
Haiku 4.5	+0.0 [−6.1, 5.9]	**+17.7** [12.0, 23.6]	**+6.6** [2.1, 10.9]

multiplicity-adjusted significance. The intervals use the same result record as Table 1. The earlier interval record used seed 20260801 and sorted quiz ordering, while the paired mask analysis used seed 20260803 and its own quiz order, both with 5,000 draws. Moving the Qwen2.5-32B Static lower endpoint from -0.2 to $+0.06$ shows how bootstrap Monte Carlo variation affects this near-zero result.

Alternative axes

Scale, family, training recipe, and cutoff co-vary in the shared-boundary panel. Table 19 separates two of these axes within the Llama family. Arm A holds size at 8B while varying the documented cutoff. Arm B fixes the December 2023 cutoff while varying size (excluding the 4-bit row from the precision-matched slope), which roughly doubles the slope to 11.3 F1 per decade over three rungs. Its exact permutation resolution floor is p=0.33. In the complementary cutoff-relative analysis, an 8B model shows a 9.5 F1 contrast with interval $[0.2, 18.1]$, confirming that the shared boundary also captures cutoff offset. FreshQA (Vu et al., 2024) supplies the second-source comparison (Table 20), where the contrast appears at every scale while volatility and split membership remain nearly collinear. Together these calibrate the panel result across family-controlled, cutoff-relative, and second-source settings (Appendix H).

Table 19: Llama comparisons: Arm A varies cutoff at 8B. Arm B varies size at a fixed December 2023 cutoff. ‡: 4-bit, excluded from the slope.

Arm A: cutoff varies, size fixed at 8B Questions	n	Δ	95% CI	p
before both cutoffs	178	+2.65	$[-2.5, 8.3]$	0.331
between the cutoffs	93	+1.88	$[-2.8, 7.0]$	0.440
after both cutoffs	607	+0.17	$[-1.6, 1.9]$	0.847

Arm B: size varies, cutoff fixed Model	Par.	Cutoff	Gap	95% CI
Llama-3.2-1B	1B	2023-12	+1.16	$[-4.4, 6.9]$
Llama-3.2-3B	3B	2023-12	−0.50	$[-8.0, 7.0]$
Llama-3.1-8B	8B	2023-12	+11.61	$[3.5, 19.6]$
Llama-3.1-70B‡	70B	2023-12	+29.46	$[20.9, 38.2]$

K Routing and distractor controls

Routing. Cached per-question conditions let us evaluate three deployable policies: never retrieve, always retrieve, and retrieve after a model declines, following SKR (Wang et al., 2023). We also evaluate a date oracle and a per-question oracle (Table 6). In that table, "never", "alw." and "abst." denote these three policies. The "temp." column uses the date rule and "per-q" selects the higher-scoring cached answer for each question. "head" subtracts the best deployable score from per-q, while Δt subtracts it from temp. These oracle margins describe opportunities within the two cached answer conditions, not achievable gains for an observed gate. Detected open-weight abstention rates are 0.0 to 6.9% of closed-book questions, so the self-knowledge router closely matches never retrieving. It performs best among our tested policies only for Claude Sonnet 4.5, whose detected closed-book abstention rate is 75% on Fresh questions. The date oracle improves over the best deployable policy for Mistral 123B (+1.8) and GPT-4o-mini (+1.5), while trailing it by up to 10.3 elsewhere. A served query lacks a publication date, so our deployable baseline excludes the date oracle. Appendix F extends this analysis with lexical-overlap and retriever-consensus gates, paired quiz-clustered contrasts, and a bounded recovery fraction.

Always-retrieve gains 18 F1 at 1.5B and 3.6 at 123B over never-retrieve.

Distractors. Rebuilding the index from all raw RealTimeQA questions with evidence yields 26,611 documents, including 19,919 same-period distractors, with a per-question time mask. BGE, Hybrid and BM25 were re-run for retrieval quality. Hybrid and BM25 also have generation runs (Table 15). Their hit@5 drops by six to seven points

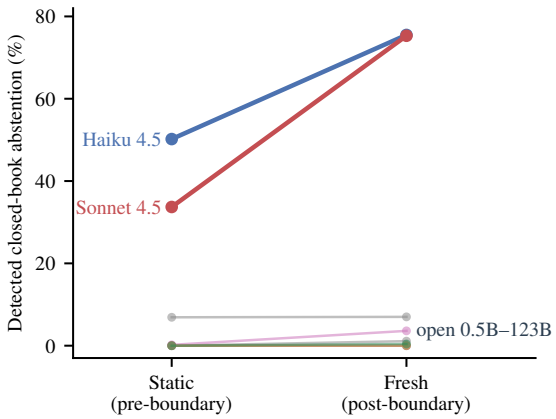


Figure 3: **Closed-book abstention across temporal regimes.** Static and Fresh rates by generator.

while their F1 differences remain unresolved. Only 320 main-index documents serve more than one question. This is a larger in-domain control, not a test of production-scale mixed-domain retrieval.

L The cutoff estimator

We estimate each generator's *effective* cutoff from its accuracy-by-year curve. **GPT-4o-mini** is the only model with a day-precise disclosed cutoff, October 1, 2023, so it supplies the only scored estimator output. Given this cutoff, closed-book F1 falls from 51.3 to 27.4 across the boundary , while the same-width contrast one year earlier is nearly flat at -0.3. For cutoff discovery, the largest quarter-over-quarter drop occurs in 2025-Q2 for GPT-4o-mini, Qwen2.5-32B and Qwen2.5-72B. This shared low-accuracy quarter shows that data-age gradients can dominate cutoff estimation (Cheng et al., 2024; Zhang et al., 2026). We therefore treat the curve-selected boundary as an exploratory estimate (Appendix C.1).

M Abstention behaviour

Figure 3 connects each generator's detected abstention percentage on Static and Fresh questions. The blue Haiku and red Sonnet lines rise sharply, while the overlapping model-colored lines represent the seven open-weight rungs. These are descriptive rates, with no intervals plotted. Whether a model abstains or guesses changes what a gap means. The open models, including Mistral 123B, anchor the capability claim because detected closed-book abstentions range from 0.0 to 6.9%. Incomplete detector recall prevents treating those rates as bounds on true abstention. The Claude models abstain heavily instead (50% → 76% Haiku, 34% → 75% Sonnet), so part of their gap is refusal, not forgetting (Appendix I).

Table 20: **FreshQA replication.** A different source, answer convention and notion of freshness. Intervals exclude zero at all six scales. Never-changing facts form the control.

Generator	Par. (B)	Never-chg.	Static	Fresh	Gap
Qwen2.5-0.5B	0.5	17.7	14.3	5.5	**+8.9** [5, 13]
Qwen2.5-1.5B	1.5	31.8	23.3	9.2	**+14.1** [8, 20]
Qwen2.5-7B	7.0	48.0	37.4	9.9	**+27.5** [21, 34]
Llama-3-8B	8.0	52.2	39.6	9.0	**+30.6** [24, 37]
Qwen2.5-32B	32.0	58.5	45.7	11.5	**+34.2** [27, 41]
Mistral 123B	123.0	73.9	60.5	15.7	**+44.8** [38, 52]

Table 20 uses 451 non-false-premise FreshQA questions and the maximum token-F1 over accepted answers. Static/Fresh follows answer effective year around 2025 rather than question publication date. B denotes billions of parameters. The never-chg. column contains the never-changing-fact subset. Gap is Static minus Fresh, with 95% bootstrap intervals. Positive gaps recur across the six rungs, but answer volatility and slice membership are strongly associated, limiting comparison with the main benchmark.

N Full-sample sensitivity

The same association at the item level. A per-answer refit with a random intercept for each question and Fresh-slice RAG F1 as capability gives the same direction. Every leave-one-model-out refit preserves the sign. We interpret this result directionally because capability varies between six model rungs, so the effective sample size remains six (Appendix D).

Removing the sampling cap changes the estimates. We scored *all* 4,176 questions from cached closed-book outputs, including stored API responses. Two-sided tail probabilities use a bootstrap over weekly quizzes. This analysis requires neither retrieval nor new API calls. The 32B's boundary contrast falls from 4.93 to +2.13 (p=0.32) and its conservative contrast to -1.81. Removing the cap changes both sample composition and n. Under joint resampling at full scale, the shared minus conservative difference remains unresolved at +3.9 $[-2.0, 10.2]$. The 2024 cohort illustrates how composition can affect the estimate. Our 123 sampled 2024 questions scored 27.9, while all 1,218 score 20.0, showing substantial variation within the Static sample. The 72B sharpens the

warning: at full scale its naive split becomes *significant* (+4.87, p=0.027) while its conservative contrast is +0.33 (p=0.91). Appendix N.1 varies the year pair directly. The full-scale run exposes sensitivity to sample composition without resolving the 32B estimate.

N.1 Year-pair sensitivity

How much the year choice matters. Refitting the seven-model parameter slope under every boundary-straddling year pair gives a range of $[-0.8, 12.7]$, a median of 5.9, and a positive value in five of six pairs. Same-side pairs serve as placebo contrasts (Appendix I). The direction is stable across most eligible pairs, while its magnitude depends on the selected years.

N.2 Post-snapshot questions

The 40 questions span two weekly quiz IDs and three question dates in July. The six available generators are Qwen2.5 at 0.5B, 1.5B, 7B and 32B, Llama-3-8B and Mistral 123B. All score below their snapshot-Fresh means, by 7.6 F1 on average. The slope of Fresh accuracy against log-parameters is 4.46 per decade on these questions and 8.00 on the same six-model snapshot panel. These are accuracy slopes, not boundary-gap slopes. With only two weekly clusters, this check remains descriptive (`results/heldout_future.json`).

O The scale gradient

Figure 4 plots the Static-minus-Fresh closed-book contrast against parameter count on a logarithmic axis: each equal horizontal step is a tenfold size increase. Open circles use the 2025 boundary, filled circles the 2023/2026 contrast, and the gray line fits the open circles across seven rungs. The 32B point moves near zero under the year-pair contrast while the largest rung remains positive. The legend's p value permutes model-level outcomes. Separately, the joint quiz bootstrap gives a slope of 5.9 [3.1, 8.7] F1 per decade, conditional on this roster. No confidence band is plotted, and neither result isolates scale from family or training differences.

P Nearest work, claim by claim

Table 21 distinguishes the questions addressed by nearby work from the claims tested here. It compares scope rather than ranking methods.

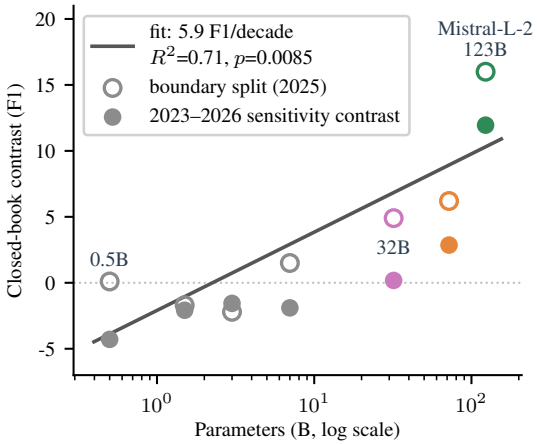


Figure 4: Closed-book contrasts across seven parameter-known rungs. Open markers use the 2025 boundary, filled markers the 2023/2026 contrast. The line fits the boundary trend.

Q Index interventions

The stale-index and leave-one-out interventions are reported in §5.3 (Table 2) as the primary Layer-3 evidence. Both change the *evidence* for the same questions and measure the response. Both estimate total effects of the tested retrieval condition, because an index change moves the document pool, the ranking and topical coverage together. Neither removes what a generator already produces closed-book, so a residual score under leave-one-out reflects closed-book performance together with whatever the remaining corpus supplies, and is not a measure of intrinsic item difficulty. Two readings follow. First, the stale index's Fresh point estimates are similar across the 40× parameter range (21.6, 21.7 and 21.3 F1, with equality untested), while its position relative to closed-book varies (−2.1 at 3B, −7.8 at 32B, −11.4 at 123B). Across these three rungs, the loss relative to closed-book is greater for models that answer more questions without retrieval. Second, the stale index retains 61% hit@5 while leave-one-out retrieves no linked relevant documents yet scores higher on Fresh questions (Table 17). Hit@5 alone therefore does not explain the harm under stale-index retrieval. Static questions are a partly treated comparison, not a placebo: the stale index changes 184 of 404 Static context sets, yet Static F1 moves by less than one point (Table 2).

Stale evidence and lexical-label precision. Hand-adjudicating 90 stale-index failures shows a lexical "grounded in stale evidence" label is right

Table 21: **Nearest prior work, claim by claim.** What each establishes and what it cannot, for the claims this paper makes.

Work	Establishes	Not supported by it
RealTimeQA (Kasai et al., 2023)	dated questions and search on news	no cross-model cutoff contrast
FreshQA (Vu et al., 2024)	freshness by answer volatility	all models judged at one date
Dated Data (Cheng et al., 2024)	resource-specific effective cutoffs	no shared relative axis
Time-aware LMs (Dhingra et al., 2022)	time as a model input	no retrieval contrast
Head-to-Tail (Sun et al., 2024)	recall varies with popularity	popularity, not time
AntiLeakBench (Wu et al., 2025)	whether an item was seen	not decay around a horizon
HoH (Ouyang et al., 2025)	outdated evidence harms RAG	supplies the stale fact without support conditioning
DYNAMICQA (Marjanovic et al., 2024)	knowledge conflict	no per-model cutoff versioning
This work: shown-text support diagnostics and temporal/index interventions across multiple models		

for only 33.3% of the items it fires on (Wilson [25, 44]). The modal failure is an *echo* of a never-correct entity rather than a faithfully reported superseded fact. Faithful reporting therefore explains 14 to 17% of stale-index failures. The two categories call for different interventions. Precision orders with model size (24%, 30%, 47%) but an exact Cochran–Armitage test gives p=0.081, so we report the pooled figure and treat the ordering as exploratory (Appendix Q.1).

Gold-Evidence as a provenance reference. Gold-Evidence contains a question's own evidence, unranked and truncated to k=5. Appendix J establishes its role as a provenance reference.

Q.1 Stale-evidence taxonomy

Every generation record stores the document ids shown to the model, allowing us to inspect the evidence associated with each failure. Taking every item where the stale index scores ≥ 10 F1 below the live one, a lexical rule based on content-word overlap between the answer and retrieved text labels 53.6% of 32B failures, 56.6% of 3B failures, and 37.2% of 123B failures as grounded in stale evidence (Table 7). Our audit distinguishes genuine superseded facts from lexical echoes. Across two random samples totaling 90 labeled items, including 60 over the Qwen pair and 30 over the 123B model, **30** (33.3%, Wilson [25, 44]) contain an answer that was once correct, such as Ron Klain for "who is the White House chief of staff". Of the rest, 41 are **echoes**: salient entities of roughly the right type lifted from stale news that were never a correct answer, such as Jimmy Carter for "which economist died at 100". Another 10 are type errors and 9 are non-answers.

Precision orders with model size at 24%, 30%, and 47%. An exact Cochran–Armitage trend test

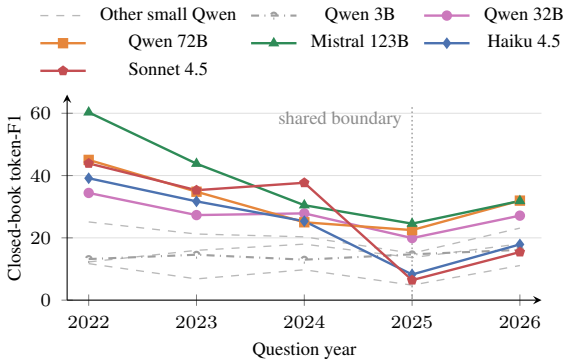


Figure 5: Closed-book F1 by year for nine generators. Eight dip in 2025 and recover in 2026, with Qwen2.5-3B the exception.

over the 90 adjudications gives p=0.081, and a Fisher test on the extreme pair gives p=0.071. Detecting this difference with 80% power requires approximately 71 items per model, compared with at most 33 here. We therefore center the result on the pooled estimate and treat the ordering as exploratory. After scaling each generator by audited precision, faithful reporting of a superseded fact explains 14 to 17% of stale-index failures. The distinction suggests different interventions. Superseded facts call for recency-aware retrieval, while unsupported echoes call for abstention or verification. One author completed this unblinded adjudication, so we use it to calibrate the pooled lexical label rather than as an agreement study.

R Naive-split diagnostics

Table 22 compares closed-book scores at the shared 2025 boundary. Δ_{cut} is the Static minus Fresh closed-book performance contrast. The final column reports the conservative 2023/2026 contrast. Intervals excluding zero are unadjusted. Multiplicity-adjusted tests of the 2025 contrast appear in Table 10.

Table 22: Question-pooled closed-book token-F1 by boundary. The 2023/2026 contrast has unadjusted 95% quiz-bootstrap intervals.

Generator	Static	Fresh	Δ_{cut}	2023/2026	95% CI
Qwen2.5-0.5B	9.5	9.4	$+0.1$	-4.3	$[-9.6, 1.1]$
Qwen2.5-1.5B	15.2	16.9	-1.7	-2.1	$[-10.3, 6.1]$
Qwen2.5-3B	13.6	15.8	-2.2	-1.6	$[-10.1, 6.9]$
Qwen2.5-7B	22.4	20.9	$+1.5$	-1.9	$[-12.5, 8.2]$
Qwen2.5-32B	30.1	25.2	$+4.9$	$+0.2$	$[-11.2, 11.1]$
Qwen2.5-72B	35.6	29.4	$+6.2$	$+2.9$	$[-7.1, 12.5]$
Mistral 123B	45.9	29.9	$+16.0$	$+12.0$	$[0.5, 22.4]$
Haiku 4.5	32.5	15.3	$+17.3$	$+13.8$	$[1.7, 25.3]$
Sonnet 4.5	39.2	13.0	$+26.2$	$+19.9$	$[8.4, 31.0]$
GPT-4o-mini	43.3	29.8	$+13.5$	$+10.8$	$[1.3, 20.1]$

Figure 5 plots annual mean token-F1, with four small Qwen rungs grouped as gray dashed curves and the other five generators identified by color and marker. The dotted vertical line marks the descriptive 2025 boundary, which is descriptive rather than a fitted change point. Lines connect annual means and carry no confidence bands. The boundary split inherits year-to-year difficulty drift. **Eight of the nine** generators dip at 2025 and rebound at 2026, including Qwen2.5-0.5B ($4.8 \rightarrow 11.1$), which has low closed-book performance. Qwen2.5-3B rises $12.99 \rightarrow 14.71 \rightarrow 16.14$ from 2024 to 2026. Together, these trajectories are consistent with changes in question difficulty without identifying a training cutoff. The leave-one-out condition agrees. With each question's gold removed, 2026 scores highest (23.8) and 2025 lowest (13.9), an observed swing of $\approx$10 F1, twice the 32B's gap. Item composition also drifts: mean answer length falls $2.63 \rightarrow 1.90$ words and the proper-noun share rises $0.58 \rightarrow 0.74$ from 2023 to 2026. Controlling for it preserves the contrast. Centering scores within answer-type$\times$length cells, or matching 131 2023/2026 pairs on answer type and length, leaves the seven-rung slope at 3.59 and 3.77 against 3.46 raw. The raw and residualised estimates weight quizzes equally within 2023 and 2026, whereas Table 22 pools questions. This explains Mistral's $+9.7$ raw gap versus $+12.0$ in that table. Its residualised and matched gaps are $+10.5$ and $+11.0$. Matching estimates the pooled within-pair difference. The exact rung-permutation p values are 0.052 (residualised), 0.045 (matched) and 0.053 (raw), so these controls preserve direction without establishing equivalence or removing unmeasured difficulty. The **conservative contrast** of §5.1 therefore runs alongside the split in Table 22, and it changes the panel's middle: the 32B's contrast falls from 4.9 to **0.2** and the 72B's from 6.2 to 2.9, with both intervals including zero. The difference between the two 32B estimates is itself unresolved ($+4.8$ $[-2.2, 12.1]$, §5.1), so we read these split values as sensitive to bin selection rather than as resolved effects.

R.1 Why Δ_{mask} is near zero

The admissible-subset check in Appendix R.2 complements this paired mask comparison. Two explanations could account for the near-zero Δ_{mask}. First, the main index pools the sampled questions' own evidence. If this leaves little competition for answer-bearing documents, the mask might simply replace one answer-bearing passage with another. Alternatively, most displaced passages might carry no answer in the first place. To examine these possibilities, we measure answer coverage on questions whose context set changed.

The mask changes the ordered top-5 for 334 of 878 questions (328 unordered sets), displacing 569 passages. Of those, 7.9% carry the gold span in the shown text, against 6.3% of the 569 passages substituted for them. Shown-text answer coverage changes by -0.5 percentage points on the changed questions ($[-1.5, 0.3]$, quiz-clustered) and by -0.2 over the full set ($[-0.6, 0.1]$). Coverage is 16.5% before and 16.0% after on Static, 18.5% and 18.5% on Fresh. The original full-text analysis reported a resolved 1.4-point coverage loss computed on full passage text. That claim no longer holds when coverage is measured on the shown text.

Two observations follow. Most displaced passages were not answer-bearing, so there was little for the mask to remove, and the coverage change is unresolved. This reading is consistent with Δ_{mask} intervals that include zero for 18 of 20 model-by-slice cells. The exposure was concentrated among passages with limited answer support. Static questions account for 238 of the 334 changes, matching

Table 23: Question-pooled Static retrieval gains on all questions and on the selected subset with originally admissible contexts. This is not the paired masking experiment.

	All (n=404)	Admissible (n=179)
Generator	ΔF1 [95% CI]	ΔF1 [95% CI]
Qwen2.5-1.5B	+19.6 [+14.9, +24.4]	+21.3 [+14.2, +28.6]
Qwen2.5-3B	+22.3 [+17.5, +27.1]	+21.0 [+13.9, +28.0]
Qwen2.5-7B	+14.3 [+9.8, +19.0]	+14.1 [+7.6, +20.8]
Qwen2.5-32B	+8.7 [+3.7, +13.8]	+15.3 [+8.2, +23.0]
Qwen2.5-72B	+4.5 [−0.3, +9.5]	+7.6 [+0.7, +14.6]
Llama-3-8B	+7.0 [+2.1, +11.9]	+8.2 [+0.5, +15.5]
Mistral 123B	−3.9 [−8.5, +0.6]	+4.0 [−2.4, +10.6]
GPT-4o-mini	−3.3 [−7.8, +1.4]	+2.4 [−4.8, +9.2]
Sonnet 4.5	−5.0 [−9.6, −0.2]	+1.2 [−5.7, +8.1]

the 19.8% against 4.8% exposure split, so the correction lands on the slice with the exposure rather than uniformly.

The main corpus holds 6,692 documents. The full corpus holds 26,611, with 19,919 same-period distractors, and applies the mask per question by construction. The full-index analysis therefore estimates masked performance rather than Δ_{mask}. Adding distractors lowers retrieval hit rates while leaving F1 changes unresolved (§K), another instance in which retrieval and answer metrics respond differently.

R.2 Admissible-subset diagnostic

Table 23 re-estimates the retrieval effect on the subset of model-question pairs whose *original* contexts were temporally admissible. This diagnostic describes how exposure is distributed. Because selection follows the treatment, 179 of 404 Static questions remain compared with 398 of 474 Fresh questions, so the subset compares different questions before and after, and a generator's score on the admissible subset differs from its score on the full slice through both exposure and composition. The disappearance of Static-side harm on this subset reflects that selection.

Our paired correction in Table 1 re-ranks and re-answers the same questions under a per-query mask, so Δ_{mask} is paired within question. That experiment keeps the Static-side harm (§5.2).

S Context budget

Every Layer-3 result fixes the context budget at top-k=5. Because the support label requires at least one of the k shown passages to contain the gold span, k changes membership in the supported and unsupported groups.

The original k=10 arm is withdrawn. The generation script truncated inputs at 1,536 tokens from the right. The k=3 and k=5 prompts fit (maximum 697 and 1,050 tokens), but 639 of 878 k=10 prompts exceeded the limit (median 1,594), so they lost the answer instruction and 480 lost the question itself. The resulting cached answers were passage continuations. The original values (4.0 to 12.4 F1) and the claim that the interaction "falls to a third or less" are invalid. The token-length audit covers all 50 cached conditions: 36 local conditions have exact counts from the model's own tokenizer or GGUF vocabulary, and the 14 API conditions (Claude Haiku, Sonnet, GPT-4o-mini) use a Qwen2.5 tokenizer as a proxy. Only the three k=10 arms exceed any limit. Generation-time library versions were not recorded.

Corrected rerun. We re-answered the same questions with the same three generators under one untruncated configuration (HF transformers, bfloat16, greedy), including a closed-book arm under the same settings. The k=3 and k=5 outputs reproduce the frozen caches exactly for 96–98% of questions. We recompute Δ_{sup} with question-pooled means and quiz-clustered intervals (Table 24). Two k=10 arms were run. The *prefix* arm keeps the five k=5 passages in place and appends the next five distinct fused candidates, so it is the addition contrast. The *re-fused* arm uses the original fusion at depth ten. It contains all five k=5 documents for 878/878 questions but reorders the first five for 724, so the re-fused minus prefix difference is an order contrast (mean F1 −0.5 to +0.0), not an addition of passages.

Table 24 compares three context budgets and four arms. The estimand is the Static support interaction Δ_{sup}: the retrieval gain (retrieval minus closed-book token-F1, paired within question) on supported minus unsupported questions, among the 404 Static questions, in F1 points (0–100). We report question-pooled arm means from the untruncated rerun, using masked rankings and labels on the 600-character shown text. The 95% percentile intervals resample weekly quizzes (5,000 draws,

Table 24: Untruncated context-budget comparison. The prefix arm adds passages, while the re-fused arm also changes their order. Interactions and intervals use question-pooled estimates.

	Δ_{sup} by context budget				k=10 prefix − k=5	
Generator	k=3	k=5	k=10 prefix	k=10 re-fused	own partition	k=5 partition fixed
Qwen2.5-1.5B	+56.6 [47.5, 65.6]	+47.5 [38.2, 56.7]	+44.5 [35.6, 53.1]	+44.1 [35.0, 52.8]	−3.0 [−8.2, 2.2]	+2.3 [−1.1, 6.1]
Qwen2.5-3B	+54.8 [45.6, 63.8]	+48.8 [40.0, 57.5]	+44.3 [35.9, 52.8]	+44.0 [35.7, 52.2]	−4.5 [−9.7, 0.5]	+0.7 [−2.7, 4.4]
Qwen2.5-7B	+51.7 [43.2, 59.8]	+48.9 [39.9, 57.7]	+40.4 [31.5, 48.9]	+40.8 [31.9, 49.3]	−8.6 [−14.8, −2.7]	−0.1 [−3.7, 3.7]

seed 20260917). The *prefix* arm keeps the five k=5 passages in place and appends five more (878/878 contain all k=5 passages, 878/878 in the same first five positions), so k=5 vs prefix is an addition contrast. The *re-fused* arm rebuilds the fusion at depth 10 (878/878 contain the k=5 passages, 724/878 reorder the first five), so it is an order contrast against the prefix arm. Going from k=5 to k=10 moves 33 questions into the supported arm and 0 out of it over all 878 questions, 17 of them Static. The last column recomputes the difference with the k=5 partition held fixed. Those fixed-partition intervals include zero, which permits nonzero effects and does not establish partition movement as the only explanation. Every arm interval excludes zero.

The interaction stays large at every budget: 51.7–56.6 F1 at k=3, 47.5–48.9 at k=5 and 40.4–44.5 at k=10 (prefix), with the lowest interval bound at 31.5. Under the prespecified rule (the k=10 point inside the k=5 interval) there is no evidence of attenuation. The secondary paired difference on each arm's own partition is -3.0 $[-8.2, 2.2]$, -4.5 $[-9.7, 0.5]$ and -8.6 $[-14.8, -2.7]$ (resolved for 1 of three generators). Between k=5 and k=10, 33 questions move from unsupported to supported over all 878 questions (17 of the 404 Static ones) and 0 move back. Holding the partition at its k=5 definition, every difference includes zero in all 18 cells (three generators, three regimes, two weightings). Some questions therefore change support membership, and holding the k=5 partition fixed yields no resolved interaction difference. Because those intervals also permit nonzero effects, they do not establish partition movement as the sole explanation of the own-partition decrease. Mean F1 rises by 0.8 to 1.6 at k=10 and falls by 0.9 to 1.6 at k=3. We therefore keep the Layer-3 result context-budget dependent: its magnitude varies with the budget, and it is resolved at every budget tested.

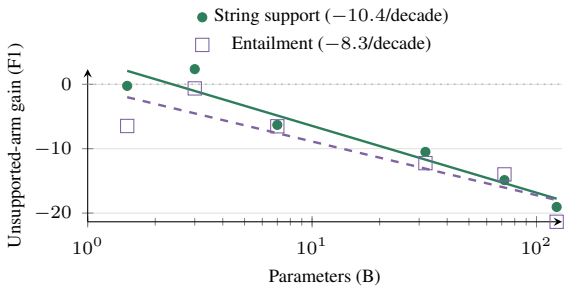


Figure 6: Unsupported-arm retrieval gain by parameter count. Results use masked Static contexts and support labels on the 600-character shown text.

S.1 Unsupported-arm scale pattern

We inspect the interaction arms separately to locate the model-level pattern. Figure 6 plots unsupported-arm retrieval gain against parameter count in the masked Static slice. Green circles and a solid fit use string support, while hollow purple squares and a dashed fit use entailment. The horizontal zero line is parity with closed-book: points below it indicate retrieval harm within the corresponding unsupported group. The log axis and negative fitted slopes show increasing harm across the roster, without plotted uncertainty bands. Across 6 known-parameter rungs, string support gives -10.4 F1 per decade with exact p=0.006, while entailment gives -8.3 with p=0.018. This roster-conditional association combines variation in model family, training, and scale, and support itself is gold-conditioned and not randomised. The association characterizes this panel. Our descriptive result is the partition of within-question retrieval gains in §5.3, supported by the index interventions in §5.3.

S.2 Pre-answer routing gates

The unsupported-arm scale pattern (Appendix S.1) uses the answer, so it is not a deployable rule. We evaluate five inexpensive gates available before generation. We apply Holm correction across nine generators (Appendix F). Lexical overlap performs best, yet its paired gain remains significant after

Holm correction for **three of nine**, Mistral 123B $+3.45$, Sonnet $+3.05$ and Qwen2.5-72B $+2.66$, the strongest closed-book generators. The best inexpensive signal recovers a median $\mathbf{15}\%$ of the per-question ceiling. These results define the available gain from the tested answer-independent gates. The date-policy selection study in §5.4 then asks whether the audit changes policy choices.

T Validation studies

The studies below cover correctness and adjudication (§T.2), passage support, freshness coding (§T.4), metric error, and failure types (§T.7). They distinguish completed legacy annotation from the outstanding exact-input collection.

This appendix reports the full numerical detail for our validation studies.

T.1 Human correctness study

Three annotators independently labelled 395 answers drawn by blinded stratified sampling, and a fourth adjudicated the items they split on, blind to the vote counts. Inter-annotator reliability is Krippendorff's α=0.904, with pairwise κ from 0.88 to 0.92. A majority resolves every item with no three-way ties. This α measures agreement among annotators. We evaluate metric-to-human agreement separately.

The audit evaluates the thresholded decision $\mathbf{1}[\text{token-F1} \geq 0.5]$. Its estimand is classification agreement at this threshold, which complements the paper's continuous mean contrasts.

We weight the item strata by their population sizes and obtain design-based intervals from a within-stratum bootstrap (5,000 common draws, seed 20260802). Using the same 34 adjudicated items and 0.5 threshold throughout gives:

- human-correct rate 0.391 $[0.327, 0.482]$
- metric-to-human agreement 0.882 $[0.791, 0.946]$, κ=0.36
- **conditional miss rate, given a human-correct answer, $\mathbf{0.240}$** $[0.096, 0.383]$
- false-credit rate, given a not-human-correct answer, 0.039

The denominator changes the interpretation. Human-correct answers scored below threshold account for 9.4% of *all* weighted outputs. Dividing that prevalence by the 39.1% human-correct rate gives the 24.0% conditional miss rate: the share of correct answers that the metric misses.

Calibration strata. Two strata, sampled at 20 items each, carry weights of 196 and 504 and together represent 84% of the item population. They are probability samples from their strata and are therefore included in the population estimate, which is why the intervals above are design-based rather than i.i.d. Their large population weights dominate the width of each interval.

Metric error patterns. Recurring causes are surface-form differences, including non-ASCII quotation marks (*"Les Misérables"* against *Les Misérables*), plurals, currency formatting and hyphenation. SQuAD normalisation strips ASCII punctuation but leaves Unicode punctuation, so these score zero.

T.2 Adjudication of contested items

The adjudicator saw the competing labels but not the vote split, so the majority could not simply be ratified. On correctness they agreed with the majority on 47% of contested items, against 50% expected from choosing between the candidates at random. The two procedures therefore have similar agreement on these contested items. The adjudicator changes 18 of these 34 labels. This set represents 8.6% of the sample, and the population estimates move by less than half a point. Departures have a consistent direction, with PARTIAL→CORRECT ten times against one in reverse, i.e. a more generous threshold at that boundary. On evidence support the adjudicator agreed with the majority 64% of the time against a 51% baseline.

T.3 Agreement with human labels

Both automatic support labels are compared with the human ones at the passage level. Task B contains 900 passage annotations from 300 question–context instances, representing 254 unique questions. Each instance contributes three passages, with repeated questions arising from separately sampled model contexts. The masked pack contains 630 passages from 210 distinct questions. The human label is the majority of three, with the adjudicator substituted on contested passages. Table 30 reports Cohen's κ against a strict human YES.

What the human statistics establish. Annotators saw 700 characters per passage, with newlines replaced by spaces in Task B and whitespace collapsed in the masked pack. Generators saw the raw first 600 characters, which we also use for

Table 25: Static support interactions with 95% quiz-bootstrap intervals. Conditions and annotation coverage differ across columns, as in Table 3.

Generator	Masked string	Masked entailment	Unmasked human
Qwen2.5-1.5B	+48.9 [39.8, 57.7]	+49.7 [43.2, 56.7]	+55.0 [39.1, 69.9]
Qwen2.5-3B	+49.0 [40.4, 57.4]	+43.5 [35.5, 51.4]	+55.2 [40.1, 69.7]
Qwen2.5-7B	+48.9 [40.0, 57.6]	+38.1 [28.8, 47.7]	+41.8 [26.5, 57.3]
Llama-3-8B	+38.0 [27.2, 48.8]	+36.2 [26.9, 45.6]	+41.8 [25.4, 58.0]
Qwen2.5-32B	+44.7 [34.8, 54.2]	+37.8 [28.0, 47.8]	+54.7 [37.1, 72.1]
Qwen2.5-72B	+46.6 [36.1, 56.6]	+34.2 [26.2, 42.3]	+44.1 [27.9, 61.5]
Mistral 123B	+37.7 [29.0, 47.3]	+33.5 [25.0, 42.4]	+30.6 [15.7, 46.1]
Haiku 4.5	+35.8 [24.9, 46.6]	+28.4 [17.2, 40.3]	+38.7 [22.6, 55.6]
Sonnet 4.5	+31.8 [21.9, 41.5]	+34.4 [25.4, 43.4]	+34.5 [18.7, 50.4]
GPT-4o-mini	+38.6 [27.5, 49.4]	+33.9 [24.2, 44.0]	+45.0 [30.3, 60.5]

every automatic label. The statistics are therefore agreement with legacy 700-character labels, not validation on the exact model input. On the shown text the string label has precision 0.82, recall 0.64 and κ=0.65 [0.55, 0.73] on Task B, and precision 0.81, recall 0.78, κ=0.74 [0.66, 0.82] on the masked pack. Entailment has precision 0.84, recall 0.85, κ=0.80 [0.74, 0.85] on Task B and precision 0.71, recall 0.87, κ=0.72 [0.64, 0.80] on the masked pack. The submitted values (κ 0.56 and 0.65, precision 0.63 and 0.62) compared labels computed on more text than the generators saw with excerpts of 700 characters. Matching the human window gives the string label κ=0.68. Eleven Task B rows (ten questions) and six masked-pack rows carry the gold string only in characters 600–700. Excluding them changes each κ by at most 0.03 (string 0.65 $\rightarrow$ 0.67, entailment 0.80 $\rightarrow$ 0.81). This exclusion is an incomplete sensitivity check for window effects, because other content in those characters can also change a judgement. An exact-input human annotation pack (900 Task B and 630 masked rows at 600 characters) is prepared under `annotation/exact_input_600_2026-09-20/` and not yet collected. Exact-input human validation therefore remains outstanding.

Question-level agreement. Humans labelled three passages per question–context instance, although generators saw five. On exactly the three human-labelled passages, a restricted annotation subset, the question-level string label agrees with the human label on 0.82 of Task B questions (κ=0.62) and entailment on 0.93 (κ=0.85). Over all five shown passages, the rates are 0.78 and 0.89. On the masked pack the three-passage rates are 0.86 for both labels and the five-passage rates 0.80 and 0.82.

Dated hypothesis. Replacing the undated hypothesis with "As of *question date*, the answer to … is …" keeps every support interaction positive with intervals excluding zero (nine of nine generators unmasked, ten of ten masked) but lowers agreement with the undated human labels to κ=0.52 [0.42, 0.61] (Task B) and 0.48 (masked pack). Date resolution has not been validated, so the paper's entailment label is *answer entailment* and the dated variant is a sensitivity analysis.

Exact-input judge. Appendix X.3 reports a separate open-weight judge applied to the exact 600-character input and calibrated against the same legacy human labels. This is automated calibration against existing annotations. The judge shares a model family with much of the generator panel and receives question and passage dates. Temporal relation validity remains unevaluated. We separately evaluate the question-level interaction in Appendix T.6.

T.4 Freshness re-pilot

The initial freshness rubric yields α=0.471 over three raters. Two construct issues explain this value. Two thirds of rows present an off-topic document, producing α=0.29 compared with 0.57 on-topic. The initial rule also asks whether a fact could change, which concerns the world rather than the observed text. Our revised rule classifies the sentence as an ongoing STATE, a completed dated EVENT, or TIMELESS, and permits SUPERSEDED only for a STATE. Piloted on 150 on-topic rows it reaches α=0.847 where all three raters agree that the evidence is in scope. Overall α=0.643 combines two raters at κ=0.91 with a third at 0.51. We use the rubric to validate the publication-date construct and reserve human-judged staleness estimates for a further pilot with the explicit scope rule.

Coded as *did this evidence post-date the question*, human agreement is α=0.893. This establishes that the post-question evidence construct is consistently codable. We then compare human and mechanical labels directly. Against the publication-date rule, the human majority agrees on $\mathbf{99.0}\%$ of 871 passages, with κ=$\mathbf{0.949}$, recall 0.97, and precision 0.94 for the mechanical flag. Annotators reading only the excerpt independently identify essentially the passages the metadata flags, which is what makes §5.2 a finding about the index rather than about the date field.

T.5 Metric and abstention errors

All metric-error summaries use identical adjudicated labels, population weights, threshold, and bootstrap draws. Human-correct answers below the 0.5 threshold comprise 14.4% $[2.8, 30.2]$ of *all* Static outputs and 3.6% $[2.5, 5.4]$ of all Fresh outputs. Their difference is 10.84 percentage points $[-1.05, 26.45]$ (two-sided stratified-bootstrap $p = 0.1304$). Conditional on a human-correct answer, the miss rates are instead 29.8% $[7.1, 49.5]$ and 12.6% $[8.5, 20.2]$, compared with an overall conditional rate of 24.0%. A threshold audit cannot fully characterize bias in continuous token-F1 contrasts. Slice-dependent metric error remains unresolved at this sample size.

The prefix-based abstention detector catches 35 of 71 adjudicated human abstentions (recall 0.493). Its reported rates describe detected refusals. Recall alone cannot establish an upper bound on true abstention, and a lower bound would also require control of false positives.

T.6 Human-labeled support

`pipeline/human_support_interaction.py` recomputes the L3 interaction with the automatic support proxy replaced by the human passage labels, taking a question to be supported when at least one of its labelled passages was called YES under the adjudicated labels (majority with contested-label overrides). Uncertainty is a quiz-clustered bootstrap resampling whole weekly quizzes, so the supported and unsupported arms move together within a draw.

Four design properties guide interpretation. The sampling strata were formed on the submitted full-text string label. The design weights remain valid, because the sample is a probability sample of those strata, but 23 of 254 questions would fall in another stratum under the shown-text label (9 Static, 14 Fresh). Task B showed annotators up to three of the top five passages. Missing a supporting passage can move a question from the supported group into the unsupported group, changing the estimated interaction. The labeled set contains 254 questions, with 127 per slice, which determines interval width. Labels correspond to the shipped, unmasked rankings, so this study validates the unmasked contrast. Results are in `results/human_support_interaction.json`. The strict YES and the looser YES-or-PARTIAL variants are both reported there, as are the Fresh-slice and pooled estimates.

T.7 Support-label errors

`pipeline/support_label_typology.py` classifies each of the 900 human-labelled passages into agreement, automatic-only (a false positive) and human-only (a false negative), and reports the rates by slice and by the generator whose retrieved context the passage came from.

The rates below are recomputed on the 600 characters shown, with a human PARTIAL counted as supported. The human side remains the legacy 700-character label. Two properties matter for Layer 3. First, both labels err mostly by omission on the shown text. Exact-string containment disagrees with humans on 21.2% of passages $[18.7, 23.9]$, with 174 human-only against 17 automatic-only passages, which is the paraphrase blindness the main text describes together with answers that appear only beyond the shown window. Entailment disagrees on 15.8% $[13.6, 18.3]$, with 125 human-only against 17 automatic-only.

Second, the slice dependence differs by label. Containment disagrees on 18.0% $[14.7, 21.6]$ of Static passages against 24.4% $[20.7, 28.2]$ of Fresh, so its omissions are more frequent on Fresh passages. Entailment disagrees on 15.6% $[12.4, 19.1]$ against 16.0% $[12.9, 19.6]$, with overlapping intervals. Omissions move some supported questions into the unsupported arm. Their effect on the interaction depends on the gains of the misclassified questions, which omission frequency alone cannot reveal. Disagreement shows no monotone relationship with generator size, ranging from 15.7% to 27.8% (string) and 11.9% to 20.8% (entailment) across the panel without ordering by parameter count.

T.8 Masked human-label study

Human support labels for the *masked* passages were collected in a second round, built by

annotation/build_masked_pack.py and scored by annotation/masked_support_2026-08-04/score_masked_support.py. The analysis was fixed in docs/protocols/masked_support.md before any label was collected, including the outcome that would have falsified the joint claim.

The pack contains 200 stratified questions, 50 from each cell of slice crossed with automatic support, plus 10 previously annotated questions for cross-round consistency. Each has three passages (630 rows total). The strata were formed on the submitted full-text string label, so their design weights remain valid. Under the shown-text label, 30 of the 200 stratified questions would fall in another stratum (10 Static and 20 Fresh, all supported to unsupported). The population support rate uses inverse inclusion weights from sampling_manifest.json on the 200 stratified questions. The ten consistency cases have no design weight and are excluded from this estimate. The corrected rate is 0.3821. The cached interactions instead use unweighted question-pooled arm means on all 210 questions, with quiz-cluster intervals. They describe this sample and are not population-weighted estimates. Human-unsupported means no support in the three annotated passages, not in all five passages shown to the generator.

Three annotators labelled all 630 passages independently, blind to the model, the condition and the automatic label, giving α=0.933 with no contested rows requiring adjudication. Agreement is higher than the 0.84 to 0.90 of the unmasked round on the same rubric, which may reflect the absence of post-question passages or differences in sample composition. The two rounds do not isolate a masking effect on agreement.

U Analysis panels

Analyses use different subsets of the roster, because not every model has every cache or a published parameter count. The machine-readable inventory, docs/model_inventory.json, is generated by scripts/build_model_inventory.py. It declares panel membership and checks unique-question coverage for every required cached condition. Unavailable settings remain explicit. The additional analyses in Appendix X link these panels to the protocol chronology.

Table 26 gives the inventory. The retriever sweep uses only Qwen2.5-32B, so these analyses do not cross every generator with every retriever.

$|M|$ is the number of models with usable caches for that analysis. Claude models and GPT-4o-mini publish no parameter count, which is why the scale analyses use fewer models than the retrieval analyses. Only models with a published count are called *rungs*.

The cutoff–scale panel contains the seven parameter-known models in the original cutoff panel. Support–scale and the item-level fit exclude 0.5B (no complete matching retrieval run) and Llama-3-8B (outside that original ladder), leaving six. API models have no published parameter counts and are excluded from scale fits. Masked string/NLI interactions cover all ten Hybrid models. The unmasked NLI panel has nine complete models and excludes Haiku. Human unmasked and masked checks cover the ten Hybrid models on 254 and 210 labelled questions, respectively. The latter interaction describes the sample. The eight-model capability centering set adds GPT-4o-mini and Sonnet to the six-model fit.

V Claude Haiku 4.5 reference

Our detailed six-retriever table uses the local open-weight Qwen2.5-32B replicate (Table 16). For reference we report the same view for a *frontier* generator, Claude Haiku 4.5 (via the Anthropic API), whose closed-book contrast is +17.3 F1 (vs a non-significant +4.9 for the 32B), the slice-by-retrieval difference-in-differences is +12.9 to +16.5 F1 across the six retrievers, compared with +4.2 to +7.6 for the 32B, and the closed-book−Gold-Evidence difficulty differential is +14.2 F1, compared with +4.6 for the 32B. These measures track the larger gap of the frontier reference (Section 5.1). Signs follow the main tables. Δ is Static−Fresh, so positive means a drop across the 2025 boundary.

Haiku provides a frontier reference, while Mistral 123B anchors our capability claim. Haiku abstains more often on Fresh questions (50% → 76%), and its *answered*-only F1 is high and similar across these date slices (64 Static / 61 Fresh), so its aggregate contrast is largely a pattern associated with rising refusal rates (Appendix M). The open-weight Mistral 123B rarely abstains, so its observed performance contrast is not accompanied by the same refusal pattern. Neither observation identifies a memory mechanism.

Table 27 reports token-F1 and answer coverage (Cov.), the fraction of retrieved full documents containing the gold span. Coverage therefore differs

Table 26: Analysis panels verified against cached outputs. $|M|$ counts models and Qs counts unique questions. Membership is analysis-specific.

Analysis	Estimand	$\|M\|$	Generators	Qs	Status
Cutoff panel	Δ_{cut} at 2025	9	0.5B, 1.5B, 3B, 7B, 32B, 72B, 123B, Haiku, Sonnet	878	complete
Full-scale closed-book	Δ_{cut}, no per-year cap	3	32B, 72B, 4o-mini	4176	complete
Unmasked Hybrid	Δ_{RAG}	10	1.5B, 3B, 7B, L3-8B, 32B, 72B, 123B, Haiku, Sonnet, 4o-mini	878	complete
Masked Hybrid (paired)	Δ_{mask}	10	1.5B, 3B, 7B, L3-8B, 32B, 72B, 123B, Haiku, Sonnet, 4o-mini	878	complete
Support (string/NLI)	Δ_{sup}, masked	10	1.5B, 3B, 7B, L3-8B, 32B, 72B, 123B, Haiku, Sonnet, 4o-mini	878	complete
Unmasked support	Δ_{sup}, string/NLI	9	1.5B, 3B, 7B, L3-8B, 32B, 72B, 123B, Sonnet, 4o-mini	878	complete
Decision	held-out policy gain	9	1.5B, 3B, 7B, L3-8B, 32B, 72B, 123B, Sonnet, 4o-mini	878	complete
Cutoff–scale	joint slope	7	0.5B, 1.5B, 3B, 7B, 32B, 72B, 123B	878	complete
Support–scale	unsupported gain slope	6	1.5B, 3B, 7B, 32B, 72B, 123B	878	complete
Item-level fit	capability interaction	6	1.5B, 3B, 7B, 32B, 72B, 123B	878	complete
Controlled family	Δ_{cut} at a disclosed cutoff	6	L3.2-1B, L3.2-3B, L3.1-8B, L3.1-8B-q4, L3.1-70B, L3-8B	878	complete
Retriever sweep	Δ_{RAG} by retriever	1	32B only	878	6 retrievers

Table 27: Claude Haiku 4.5 (frontier reference), 2025 boundary. ΔF1 is Static−Fresh. †: two-sided quiz-bootstrap $p < 0.001$.

System	Static		Fresh		ΔF1
	F1	Cov.	F1	Cov.	
Closed-book	32.5	n/a	15.3	n/a	+17.3†
BM25	32.9	0.24	28.5	0.25	+4.3
DPR	27.8	0.22	26.8	0.24	+1.0
SPLADE	33.5	0.23	30.0	0.27	+3.5
ColBERT v2	34.9	0.23	31.1	0.27	+3.8
BGE	33.9	0.24	33.1	0.28	+0.8
Hybrid	35.0	0.24	32.1	0.28	+2.8
Gold-Evidence	32.9	0.22	29.8	0.25	+3.1

from support in the 600-character excerpt. Every retrieval row has a smaller Static-minus-Fresh point contrast than closed-book, largely through higher Fresh scores. Because the table describes a single frontier reference, its retrieval rankings and refusal behavior need not extend to other generators.

W Qualitative examples

Tables 28 and 29 show examples from the cached per-example outputs in the public artifact. Each is a Fresh (2026) question that the closed-book generator answers incorrectly and the retrieval-augmented system (Hybrid) answers correctly. We label the generator explicitly in each table, because the two behave differently in a way that matters for interpreting the closed-book contrast (Appendix M). In these examples, Qwen2.5-32B supplies an incorrect answer, sometimes an outdated one, while Claude Haiku often abstains. Predictions are reproduced as logged, except that long refusals are elided with "…". Counting a *rescue* as a Fresh question (2025 boundary, n=474) with closed-book token-F1 <0.30 and hybrid token-F1 >0.80, Qwen2.5-32B has 88 and Claude Haiku 92. The arrows compare logged answers to the same question. Dates fix the time at which the supplied gold answer is evaluated. These selected rescues illustrate different error behaviors, not their frequencies or proof that the generator used a particular retrieved passage.

X Additional analyses

This appendix holds the detail behind the supplementary checks: what was prespecified for the new experiments, the correction manifest for the support analyses, the exact-input judge calibration, and the full window, replication, context-budget and decision results summarised in §5.

Table 28: **Qwen2.5-32B**. The model commits to outdated answers, including the 2022 host city for the Winter Olympics item.

Question (Fresh, date)	Gold	Closed-book → Hybrid
Where will the Winter Olympics be held next month? (2026-01-15)	*Italy*	**Beijing, China** (stale) → **Italy** ✓
Which former women's tennis champion came out of retirement to play doubles at Queen's Club? (2026-06-12)	*Serena Williams*	**Martina Navratilova** (wrong) → **Serena Williams** ✓
How many moons does Saturn have? (2026-02-26)	*274*	**82** (stale) → **274** ✓
Which AI company may be headed for a major stock-market debut? (2026-06-11)	*OpenAI*	**Anthropic** (wrong) → **OpenAI** ✓

Table 29: Claude Haiku 4.5 on selected Fresh questions. These examples include abstentions and an outdated answer to the Saturn item.

Question (Fresh, date)	Gold	Closed-book → Hybrid
Which tech company laid off 10% of its workforce this week to focus more on AI? (2026-05-22)	*Meta*	"I don't have access to current information…" → **Meta** ✓
How many moons does Saturn have? (2026-02-26)	*274*	**146** (stale) → **274** ✓
Which AI company may be headed for a major stock-market debut? (2026-06-11)	*OpenAI*	"I don't have access to documents or info…" → **OpenAI** ✓

X.1 Documented prespecification

We documented the primary E2/E3/E6 and E5 protocols before their designated analyses. Subsequent corrections and additional analyses are identified separately, with deviations recorded. The primary protocols for the window (E3), replication (E2), context-budget (E6) and decision (E5) studies are written down with their decision rules in `docs/protocols/primary_analyses.md` before the corresponding generation ran. That file is a dated local plan, which is documented prespecification rather than an independently authenticated public registration, and it also records additions made after a preliminary inspection of E6. The later R-series corrections and sensitivity analyses (masked entailment repair, visible-text relabelling, the Mistral native-template run, the selection-refit bootstrap and the judge calibration) were *not* prespecified and are reported as follow-up analyses throughout. The shared configuration fixes the generator settings (HF transformers, bfloat16, greedy, 32 new tokens, no truncation, with the script aborting on any prompt over $4{,}096$ tokens), byte-identical prompts to the original pipeline, token-F1 as the primary endpoint, paired quiz resampling ($B{=}5{,}000$, seed 20260917), Holm correction across generators within each primary family, and the shown-text string label. The decision rules fixed in advance were:

- **E3 (window).** Arms $w \in \{0, 1, 3, 7\}$ days and unmasked, mask applied before fusion. The three-day arm must reproduce the shipped masked rankings and the unmasked arm the shipped rankings (both reproduced exactly). Primary contrasts are quiz-weighted paired differences. A window is "practically equivalent" to another only if the 90% interval of the paired difference lies inside ± 2 F1, otherwise "not shown equivalent" or "different" (95% interval excludes zero). Fixed in advance: the corpus was source-filtered at three days, so the seven-day arm tests sensitivity within the retained document universe only.

- **E2 (replication).** 800 unused question ids, 200 per year for 2022–2025, seed 20260917, with inclusion probabilities stored per item. Both samples use the frozen $26{,}611$-document corpus and closed-book, masked Hybrid and unmasked Hybrid arms. Stability margins: ± 2 F1 for Δ_{mask} and ± 5 for Δ_{sup}, judged by the 90% interval of the new-minus-original difference from joint resampling of shared quizzes. Otherwise the verdict is "stability not established" or "changed".

- **E6 (context budget).** Arms $k{=}3$ (prefix of $k{=}5$), $k{=}5$, the original re-fused $k{=}10$ list, and $k{=}10$ prefix. The primary family is the Static support interaction per arm. Report the support transition matrix regardless of outcome and conclude "no evidence of attenuation" if the $k{=}10$ prefix point lies inside the $k{=}5$ interval.

- **E5 (decision).** Policies closed-book, retrieve, and closed-book-before-b for eight half-year boundaries. Naive selection uses the first two policies and audit selection all candidates, ranked by development-fold quiz-weighted F1.

Use 20 repeats of five-fold cross-fitting grouped by weekly quiz and Holm correction across nine generators. A positive Holm-adjusted difference means the audit changes the decision and improves held-out utility. Report a negative difference as a failure and identical selections or an interval including zero as "unchanged or not resolved".

Deviations and additions are logged with their dates. A paired bootstrap of the interaction difference was added to E6 after a preliminary run had used the frozen closed-book cache (the prespecified rule is still reported), and the two $k{=}10$ arms were reinterpreted as an order contrast once it was found that every re-fused list contains all five $k{=}5$ documents. The 1.5B model was added to E3 because E6 needed its closed-book arm. The 32B and Mistral arms ran on the original Metal configuration, whose frozen outputs reproduced exactly ($40/40$ and $20/20$ checked questions). The selection-refit bootstrap for E5 (Appendix X.6) was added retrospectively and is labelled as such.

X.2 Support-label corrections

Checking the original pipeline exposed two measurement errors. Although prompts show the first 600 characters of each passage, the string support label had read the whole stored passage and the entailment label 1,500 characters. The masked entailment arm had also reused labels from unmasked contexts. Every support analysis was relabelled on the shown text. The manifest (`docs/corrections.md`, generated by `results/analysis/CORRECTION_MANIFEST.json`) lists 36 analyses: 22 recomputed on the visible text, 7 that depend on legacy 700-character human labels and are reported as agreement with those labels, 6 not affected, and 1 withdrawn (the truncated $k{=}10$ arm, Appendix S). The effect of the relabelling is uniform: every support interaction remains positive with an interval excluding zero and grows. The unmasked Static string interaction moves from 17.4–37.1 to 29.6–51.1 F1 across nine generators, the unmasked entailment interaction from 19.4–34.6 to 32.0–50.0, the masked string interaction from 19.7–36.5 to 31.8–49.0 (ten generators) and the masked entailment interaction, now labelled on the masked contexts, from $+20.7$–$+35.2$ to $+28.4$–$+49.7$. The string-supported arm shrinks from $214/264$ to $167/208$ questions (Static/Fresh), every move being supported to unsupported. Context-level agreement between the two labels rises from 77.6% to 86.4%. Two consequences of the original full-text analysis are recorded: the claim that strong models do not gain given support no longer holds (Sonnet's Static gain given support is 10.8 $[2.4, 19.2]$ and Mistral's 19.2 $[12.0, 26.3]$), and the mask-null coverage claim is removed (Appendix R.1). The answer-coverage retrieval statistic of §4 still reads full documents and is not a support label. The correction manifest and replay scripts are public. Appendix A records the current traced input inventory, successful and failed analysis stages, and intentionally skipped model-label generation. The corrected cached-analysis artifact is public (Appendix A). Fresh tokenization and human collection remain incomplete.

Table 30: Agreement with legacy human support labels. These statistics do not establish exact-input human validation or superiority of a labeling method.

Automatic label	Agree	κ	Prec.	Recall
String, 700 (human)	0.890	0.676 [0.58, 0.76]	0.819	0.684
String, 600 (shown)	0.882	0.646 [0.55, 0.73]	0.819	0.641
Entailment, 600 (shown)	0.928	**0.800** [0.74, 0.85]	0.842	**0.854**
LLM judge, 600 (shown)	**0.932**	0.792 [0.74, 0.84]	**0.987**	0.722

X.3 Exact-input judge calibration

Table 30 compares automatic labels with human judgements on 900 passage rows from 254 questions (Task B), using strict *yes* on both sides. Agree is agreement, Prec. is precision, and κ is Cohen's kappa, with quiz-clustered 95% intervals beneath it. Annotators judged 700-character excerpts, whereas automatic labels use the 600-character text shown to generators. These rates measure agreement with legacy human labels rather than validation on the exact model input (11 rows carry the gold string only in characters 600–700). The judge row is a separate open-weight model (Qwen2.5-14B-Instruct) applying the annotators' rubric. Sharing a model family with much of the panel may correlate its errors with the generators'. Its prompt includes question and passage dates, but temporal relation validity was not separately evaluated. Bold marks the highest point estimate per column among the 600-character rows. Overlapping κ intervals and the absence of a paired comparison leave the relative accuracy of the labels unresolved.

For automated calibration, a separate open-weight judge (Qwen2.5-14B-Instruct, greedy, out-

put constrained to YES/PARTIAL/NO) applied the annotators' Task B rubric to the exact 600-character input of every annotated passage and, for a window-matched comparison, to the 700-character excerpt the annotators saw. Against a strict human YES the judge reaches κ=0.79 [0.74, 0.84] on Task B and 0.83 [0.77, 0.89] on the masked pack at 600 characters, and 0.83 and 0.87 at 700 (Table 30 gives the Task B rows beside the string and entailment labels). On the masked pack, the string label reaches 0.74 [0.66, 0.82] and entailment 0.72 [0.64, 0.80]. All three κ intervals overlap within each pack. We report point estimates with intervals, leaving relative agreement unresolved because no paired comparison was run. On Task B the entailment point estimate is nominally the higher one. The separate open-weight judge shares a model family with much of the generator panel, so its errors may be correlated with theirs. Its errors are omissions: precision 0.99 and 0.94, recall 0.72 and 0.80. Applied to all 4,390 shown passages, the judge label puts 165 of 404 Static and 186 of 474 Fresh questions in the supported arm (string: 167 and 208, entailment: 207 and 266). Under the judge label the support interaction excludes zero for all ten generators in both the unmasked and the masked condition and is larger than under our labels: the unmasked Static medians are 43.1 (judge), 40.4 (string) and 36.7 (entailment) question-pooled, and the judge minus entailment difference is positive for ten of ten generators. Twenty passages have exact ties between two label probabilities. Reversing the tie-breaking choice moves at most three questions between arms. The judge prompt supplies the question date and each passage's publication date, but we did not separately evaluate temporal relation validity. Establishing temporal reasoning therefore requires further validation. This study calibrates a judge against legacy human labels. Exact-input human validation remains outstanding.

X.4 Admissibility windows in full

Table 31 compares each admissibility window with no mask for five generators. Each window cell is the paired *masked-window minus unmasked* difference in quiz-weighted token-F1 points (0–100), pooled over the 404 Static and 474 Fresh questions and resampled over the 199 weekly quizzes. The last column is the paired zero-day minus three-day difference with its 95% interval. Equivalence uses a different criterion: the 90% interval must lie inside ±2 F1. All 20 overall window-versus-unmasked intervals meet that criterion, although several Static or Fresh intervals do not. The 1.5B, 3B and 7B arms use historical HF/bfloat16 reruns with untruncated prompts, not the frozen answers in Table 1. The 32B and Mistral arms reuse frozen answers where rankings match. For frozen 1.5B answers, the overall three-day effect is +0.81, with 90% interval [−0.61, 2.35] under 5,000 quiz resamples (seed 20260803). It is not equivalent within ±2. Equivalence here applies to the window-study outputs. The public `results/reporting_summary.json` records the frozen check.

Window choice also changes which passages reach the generator. It changes the ordered top-5 for 41 (0 d), 22 (1 d) and 25 (7 d) of 878 questions relative to three days, against 334 for the unmasked ranking, and the shown-text support rate moves between 0.419 and 0.427 (one question gains and two lose support at zero days). All 20 window-versus-unmasked contrasts are practically equivalent within ±2 F1 because their 90% intervals lie inside that margin. Holm p=1.0 refers separately to zero-difference tests and does not establish equivalence. Within Static alone, the 1.5B intervals are "not shown equivalent" for every window (for example +0.78 [−1.34, 3.08] at three days), and within Fresh the 32B intervals are "not shown equivalent" at one, three and seven days (for example −0.69 [−2.13, 0.46]) while the 7B three- and seven-day contrasts are "different" but small (+0.69 [0.10, 1.49] and +0.59 [0.09, 1.37], masked better). Zero days against three days is "different" for the 1.5B and 3B models (−0.28 [−0.51, −0.08] and [−0.50, −0.10], 90% intervals, Holm p=0.040 and 0.036) while 90% intervals establish practical equivalence for the 7B, 32B and Mistral (−0.08, +0.11, −0.05). Seven days against three days is at most 0.03 in absolute value. On the 41 questions whose context changes at zero days, question-pooled F1 falls by 8.2, 9.6, 3.8, 0.6 and 5.0 for the 1.5B, 3B, 7B, 32B and Mistral models (harmed/rescued counts 6/1, 5/0, 2/0, 3/3, 3/1): removing post-question documents that fall inside the grace window hurts the affected questions, and the loss is invisible on average.

X.5 Question-disjoint replication

Table 32 reports the question-disjoint replication for the common years 2022–2025, quiz-weighted estimates pooled over Static and Fresh, in token-F1 points (0–100), with 95% percentile intervals

Table 31: Admissibility-window sensitivity. Overall quiz-weighted masking effects and the zero-day versus three-day contrast. Intervals are 95%.

Generator	0 d	1 d	3 d (study)	7 d	0 d − 3 d
Qwen2.5-1.5B	+0.13	+0.32	+0.41	+0.43	−0.28 [−0.56, −0.05]
Qwen2.5-3B	−0.05	+0.13	+0.23	+0.24	−0.28 [−0.55, −0.07]
Qwen2.5-7B	+0.25	+0.24	+0.33	+0.29	−0.08 [−0.23, 0.03]
Qwen2.5-32B	−0.41	−0.40	−0.52	−0.55	+0.11 [−0.25, 0.59]
Mistral 123B	−0.06	−0.11	−0.02	+0.00	−0.05 [−0.30, 0.23]

Table 32: Question-disjoint, within-source replication on 800 unused questions. Estimates and interval-based stability checks account for shared weekly quizzes.

Generator	Δ_{sup} original (532 q)	Δ_{sup} new (800 q)	Δ_{mask} original	Δ_{mask} new	Difference within margin?
Qwen2.5-3B	42.9 [33.5, 52.3]	48.9 [42.4, 55.7]	+0.82	+0.01	not established
Qwen2.5-7B	48.9 [39.9, 57.7]	47.3 [39.5, 54.9]	−0.22	+1.24	not established
Qwen2.5-32B	45.8 [36.7, 54.8]	48.4 [40.9, 56.1]	−0.02	+0.47	not established
Mistral 123B	39.9 [30.7, 49.0]	42.7 [34.8, 50.8]	+0.38	+0.43	yes (Δ_{mask})

resampling the weekly quizzes shared by the two samples. Δ_{sup} is the masked-context support interaction (retrieval gain on supported minus unsupported questions) and Δ_{mask} is masked minus unmasked retrieval F1. The new sample is 800 questions, compared with 532 in the original common-year sample. Margins fixed in advance: ± 2 F1 for Δ_{mask} and ± 5 F1 for Δ_{sup}, judged on the 90% interval of the new-minus-original difference. Stability within those margins is established only for Mistral's Δ_{mask}.

The new sample (800 questions: 600 Static, 200 Fresh) has a shown-text support rate of 0.464 against 0.425 for the original common-year questions, and unmasked exposure of 17.5% of contexts against 12.0%. Table 32 gives the estimates. The new-minus-original differences (quiz-weighted, pooled) are: Δ_{sup} $+6.0$ $[-3.8, 16.1]$, -1.6 $[-11.4, 8.4]$, $+2.7$ $[-6.9, 12.3]$ and $+2.9$ $[-6.8, 12.4]$ for the 3B, 7B, 32B and Mistral models (90% intervals), none inside ± 5. For Δ_{mask}, the differences are -0.8 $[-2.7, 1.0]$, $+1.5$ $[0.1, 2.9]$, $+0.5$ $[-1.0, 2.0]$ and $+0.05$ $[-1.7, 1.8]$, only the last inside ± 2. Question-pooled, the 3B Δ_{sup} difference is "changed" and larger in the new sample ($+10.2$ $[0.2, 20.1]$, 95%) and its Static Δ_{mask} likewise (-2.31 $[-4.64, -0.02]$), while the 32B and Mistral pooled Δ_{mask} are "stable within margin". Inverse-probability weighted population estimates on the new sample give Δ_{sup} of 51.1, 46.8, 46.0 and 38.0 F1 and Δ_{mask} of -0.3, $+1.3$, $+0.5$ and $+0.5$. The 43 questions in quizzes absent from the original sample have a support rate of 0.581 and are reported descriptively only. A large support interaction and a small mask effect thus recur on question-disjoint items, but these sample sizes do not establish quantitative stability.

X.6 Date-policy selection

Table 33 reports quiz-weighted token-F1 (0–100) on held-out weekly quizzes for nine generators, pooling Static and Fresh questions. Retrieval uses masked Hybrid rankings, and 95% percentile intervals resample the 199 weekly quizzes. *Naive* chooses between always-closed-book (CB) and always-retrieve (RAG), while *audit* may also choose a date boundary. Choice columns show the two most frequent choices, when available. Omitted choices account for any remaining share. The *20-repeat estimate, conditional CI* column uses $20\times$ five-fold cross-fitting and treats the stored cross-fitted predictions as fixed. The Holm p column belongs to that analysis. The *5-repeat estimate, selection-refit CI* column re-runs the whole cross-fitting and selection inside each of 1,000 outer bootstrap draws with $5\times$ five-fold cross-fitting, and its point estimate comes from that five-repeat procedure, so the two columns differ in repeat count as well as in what is resampled (for example Mistral 123B $+4.33$ against $+3.35$). All Holm-adjusted p values for the refit procedure are at least 0.54. Under the 20-repeat conditional inference: 2 gains, 2 losses, 1 unresolved, 4 with identical actions. Under the 5-repeat selection-refit inference there are 0 gains, 0 losses, 5 unresolved, 4 with identical actions.

Both gains and losses are reported in Table 33. Under the 20-repeat conditional inference the Holm-adjusted p values are 0.004 (Mistral 123B), 0.016 (Sonnet), 0.043 (Qwen2.5-72B),

Table 33: Exploratory date-policy selection. Conditional and selection-refit intervals differ. No selection-refit utility difference is resolved after Holm correction.

A. Held-out utility and retrieval rates

Generator	20-repeat estimate conditional 95% CI	Holm p 20-repeat	5-repeat estimate selection-refit 95% CI	Retrieval rate naive → audit
Qwen2.5-1.5B	+0.00 [0.00, 0.00]	1.000	+0.00 [0.00, 0.00]	1.00 → 1.00
Qwen2.5-3B	+0.00 [0.00, 0.00]	1.000	+0.00 [0.00, 0.00]	1.00 → 1.00
Qwen2.5-7B	+0.00 [0.00, 0.00]	1.000	+0.00 [−0.22, 0.00]	1.00 → 1.00
Qwen2.5-32B	+0.00 [0.00, 0.00]	1.000	+0.00 [−2.62, 0.00]	1.00 → 1.00
Qwen2.5-72B	−0.77 [−1.46, −0.18]	0.043	−0.46 [−2.82, 1.63]	1.00 → 0.94
Mistral 123B	+4.33 [2.22, 6.55]	0.004	+3.35 [−1.27, 6.69]	0.60 → 0.55
Llama-3-8B	−0.23 [−0.49, −0.04]	0.034	+0.00 [−3.15, 0.65]	1.00 → 0.99
GPT-4o-mini	+3.22 [0.64, 5.93]	0.068	+3.07 [−0.66, 7.78]	0.05 → 0.53
Sonnet 4.5	+5.39 [1.88, 9.00]	0.016	+5.30 [−0.25, 10.28]	0.95 → 0.54

B. Most frequent choices across training folds

Generator	Naive choice	Audit choice
Qwen2.5-1.5B	RAG 100%	RAG 100%
Qwen2.5-3B	RAG 100%	RAG 100%
Qwen2.5-7B	RAG 100%	RAG 100%
Qwen2.5-32B	RAG 100%	date 2022-07-01 80%, RAG 20%
Qwen2.5-72B	RAG 100%	date 2022-07-01 57%, date 2023-07-01 26%
Mistral 123B	RAG 60%, CB 40%	date 2025-01-01 85%, date 2025-07-01 7%
Llama-3-8B	RAG 100%	RAG 95%, date 2023-01-01 4%
GPT-4o-mini	CB 94%, RAG 6%	date 2025-01-01 92%, date 2025-07-01 8%
Sonnet 4.5	RAG 95%, CB 5%	date 2025-01-01 100%

0.034 (Llama-3-8B) and 0.068 (GPT-4o-mini). Under the 5-repeat selection-refit inference every Holm-adjusted p is at least 0.54, and the refit draws are centred below the original-data estimates where selection matters. A resampled data set holds about 63% distinct quizzes, which may affect selection without establishing conservative interval coverage. Question-pooled estimates agree in sign (stored: Mistral $+4.0$ $[2.3, 5.7]$, Sonnet $+2.9$ $[0.7, 5.2]$, GPT-4o-mini $+5.1$ $[2.2, 7.8]$, 72B -0.7, Llama -0.3). Under refitting, only GPT-4o-mini is resolved before correction, at $+4.9$ $[0.8, 7.0]$. Subgroups: the 72B and Llama losses are Static only (-1.2 $[-2.3, -0.3]$ and -0.4 $[-0.8, -0.1]$ stored). Against always-closed-book on Static questions the naive choice costs Sonnet 7.9, Mistral 4.8 and GPT-4o-mini 0.9 F1. The audit choice removes that cost for Sonnet and GPT-4o-mini and leaves Mistral at -0.9 $[-1.6, -0.3]$. Retrieval call rates move from 0.05 to 0.53 (GPT-4o-mini), 0.95 to 0.54 (Sonnet), 0.60 to 0.55 (Mistral), 1.00 to 0.94 (72B) and 1.00 to 0.99 (Llama). These call rates do not measure latency. For the three generators with positive conditional gains, the modal audit choice is the 2025-01-01 boundary, which is the paper's descriptive split and was originally read off the full-data curve: the selection is cross-fitted, but the candidate set is not independent of the data, a residual circularity. The incremental value of Layer-2 masking inside the date-conditioned class is small: masking changes 0 to 3.6% of held-out actions, and the corresponding F1 differences range from -0.35 to $+0.33$ with no Holm-significant cell. A secondary leakage variant, the naive procedure scored on unmasked against masked outputs, changes the closed-book versus retrieve choice in 0 of 100 fits for six generators (15 for Mistral, 7 for Sonnet, 2 for GPT-4o-mini) with a resolved difference for Sonnet only ($+1.2$ $[0.3, 2.2]$, masked better). The study is internal validation over random held-out quizzes of an already-inspected dataset, not a prospective test.

X.7 Mistral prompt format

The Mistral 123B407 GGUF lacks `tokenizer.chat_template`, so `llama-cpp-python` rendered every Mistral prompt in its llama-2 fallback format (`[INST] <<SYS>>...<</SYS>>...[/INST]`). The deviation is identical across conditions and is not a truncation issue. For the check, all 878 questions were re-answered closed-book and with masked Hybrid retrieval under Mistral's v3 instruct format, with the same GGUF, context size and decoding. Token construction was verified against the official `mistral_common` v3 tokenizer on 878 of 878 prompts. Exact-answer agreement between

formats is 80.2% closed-book and 76.3% with retrieval. Mean F1 differences are $+0.6$ $[-0.5, 1.7]$ closed-book and $+0.1$ $[-0.8, 1.0]$ with retrieval. The Layer 1 contrast is $16.0 \rightarrow 16.9$ question-pooled ($+0.9$ $[-1.3, 3.2]$) and $15.9 \rightarrow 17.5$ quiz-weighted ($+1.6$ $[-0.6, 3.7]$). The conservative 2023/2026 contrast is $12.0 \rightarrow 14.3$. The Static support interaction falls from 38.2 to 30.7 quiz-weighted (-7.5 $[-14.0, -1.7]$, pooled regime -4.1 $[-8.2, -0.5]$) and is unresolved question-pooled (-2.9 $[-7.9, 2.0]$). The Fresh interaction is unresolved under both weightings. Thus the Layer 1 anchor has no resolved change, although its intervals permit differences of a few F1 points in either direction. Mistral's Layer 3 magnitudes depend on the tested configuration.

X.8 Retrieved-but-hidden answers

Figure 2 separates three visibility groups. Each row is one generator: green circles mark shown answers, orange squares hidden answers, and gray triangles answers absent from the retrieved passages. Horizontal position is RAG minus closed-book F1, with zero denoting no gain. Bars are 95% percentile intervals from 5,000 joint weekly-quiz bootstrap draws. Class sizes are fixed across generators, and means weight questions equally within each class. The hidden group loses performance despite retrieval of an answer-bearing document, which distinguishes document retrieval from exposure to useful text. For a concrete example, the retrieved text for the 2026-04-24 meteor-shower question contains "Lyrids" only after character 1,453, beyond the shown prefix.

For each question we classify the shipped Hybrid top-5 as SHOWN (the gold string occurs within the first 600 characters of at least one passage, 375 questions), HIDDEN (it occurs only beyond the shown window, 103) or NONE (400). The SHOWN minus HIDDEN retrieval gain is 45.4 to 57.1 F1 question-pooled (41.0 to 55.6 quiz-weighted) and excludes zero for all ten generators, pooled and within Static and Fresh. The HIDDEN gain itself is negative (-7.5 to -20.8). HIDDEN minus NONE is negative for eight of ten generators pooled, but HIDDEN questions have higher closed-book F1 (Qwen2.5-32B: 33.2 against 17.0) and similar retrieval F1. This baseline difference prevents attributing the lower gain to hidden answers. Class membership is not randomised: relative to SHOWN, HIDDEN questions have a worse best rank of the answer-bearing passage (2.45 against 1.45), lower panel closed-book F1 (29.8 against 35.6), a median gold offset of 1,313 against 92 characters, and similar years and answer lengths. The masked rankings give the same picture (369/102/407 questions).

Y Detailed scope and limitations

We define the scope of each claim through its design, endpoint, and evaluation panel. The supporting analyses appear in the Results and Appendix T.

Observational scope. Our contrasts identify where retrieval effects concentrate within the observed questions, contexts, and models. The stale and leave-one-out indices are interventions on the retrieval condition, and their effects are total effects that change the document pool, ranking, and topical coverage together. Support status is an offline, gold-conditioned diagnostic of the shown text. It correlates with entity salience and popularity, and it does not establish evidence use. Model size also co-varies with family, data, and training procedure. Provider cutoffs are documented dates rather than measured boundaries. Causal identification of evidence use would require human-verified entailing, neutral, and contradicted evidence crossed with current, superseded, and anachronistic time status within a fixed-cutoff model family.

Question time. The 2025 boundary comes from the observed closed-book curve, so we use it as a descriptive split, and the 2023 to 2026 pair provides the conservative sensitivity analysis. Seven of nine generators have undisclosed cutoffs, which keeps their contrast at the calendar level. Layer 1 measures a closed-book performance contrast between question dates. Knowledge and memorization require separate identification. The permutation test rejects its specified answer-permutation null. **The scale association is roster-conditional** and endpoint-sensitive: six of seven model-level gap intervals include zero, and the quiz-weighted 2023/2026 choice-grounded slope yields p=0.27. The controlled family fixes release, documented cutoff, and quantisation for its closest pair, while training dynamics may still vary with size. Its cutoff arm has a minimum detectable effect of 7.7 F1, which sets the resolution of that analysis.

Evidence time. Our pooled index filtered documents relative to their source questions, and the per-query mask aligns each document with the question being answered. The three-day grace convention is a choice. Windows of 0, 1, 3, and 7 days keep the overall quiz-weighted masking effect within

± 2 F1 for the five generators tested, with wider Static and Fresh intervals, and a zero-day rule costs the two smallest models 0.28 F1. Across the 10 complete generators in Table 1, Δ_{mask} remains near zero. This empirical range defines the mask-effect claim.

Context budget. The support interaction is context-budget dependent. Adding five passages after the $k{=}5$ list lowers the Static interaction by 3 to 9 F1 on each budget's own partition for the three generators tested. Partition movement contributes to this comparison: 33 questions become supported, none lose support, and every difference on the fixed $k{=}5$ partition includes zero. Because those intervals also permit nonzero effects, partition movement need not be the only explanation.

Answer metrics. The human study evaluates the thresholded decision $\mathbf{1}[\text{token-F1} \geq 0.5]$. Among human-correct answers, the threshold identifies 76% and misses 24%, with a wide design-based interval because two high-leverage strata dominate the weighting. The point estimate of metric error is 14.4% of all Static outputs and 3.6% of all Fresh outputs (human-correct but below threshold), while the interval for their difference includes zero. We therefore pair continuous token-F1 with Exact Match and choice-grounded accuracy. Our abstention detector has 0.493 recall, so its reported rates cannot bound true abstention.

Support labels. Containment and entailment are computed on the 600-character text each generator saw. Annotators judged 700-character excerpts, so our human statistics are agreement with those legacy labels rather than validation on the exact input: string $\kappa{=}0.65$ and entailment $\kappa{=}0.80$ at the passage level, changing by at most 0.03 when the 11 rows whose answer appears only in characters 600–700 are excluded. A separate open-weight judge applying the annotators' rubric to the exact input reaches $\kappa{=}0.79$ $[0.74, 0.84]$ and 0.83 $[0.77, 0.89]$ against the human labels on the two packs. These intervals overlap those of the string and entailment labels. Relative agreement remains unresolved without a paired comparison. The judge receives the question date and the passage publication date, but we did not separately evaluate temporal relation validity, and it shares a model family with much of the generator panel, so its errors may be correlated with theirs. Its role is judge calibration against existing annotations. Exact-input human validation remains outstanding. A human pass on the exact 600-character input (900 Task B and 630 masked rows) is prepared and not yet collected. The initial human freshness field reaches $\alpha{=}0.47$, which places it below our reliability threshold of 0.6 and excludes it from the staleness estimate.

Prompt format. Mistral 123B ran in llama.cpp's llama-2 fallback format because its GGUF file carries no chat template. Re-answering all 878 questions in the model's own format does not resolve a change in condition means or in the Layer 1 contrast (16.9 vs 16.0 F1, a change of $+0.9$ $[-1.3, 3.2]$), although the intervals permit nonzero differences. The Static support interaction falls under quiz weighting ($38.2 \rightarrow 30.7$, a change of -7.5 $[-14.0, -1.7]$). Mistral's Layer 3 magnitudes are therefore scoped to the tested configuration.

Retrieval corpus and source. The main index pools evidence linked to sampled questions, creating a restricted retrieval setting. The 26,611-document index adds same-period distractors and preserves the main conclusions, while remaining smaller than a deployed news index. At that scale we re-run BGE, Hybrid and BM25 for retrieval quality, and Hybrid and BM25 for generation. We use direct zero-shot DPR and a fixed query without reranking, and relevance comes from each question's linked evidence rather than exhaustive corpus annotation. Gold documents contain an answer-bearing passage for 52.3% of questions, which limits the Gold-Evidence reference but not retrieval from the pooled corpus. The main three-layer audit uses RealTimeQA: 878 English questions from Western news outlets in one evaluation snapshot. The question-disjoint replication draws 800 unused questions, but 172 of its 182 quizzes are weeks already in the working sample, so it is a within-source, largely same-week replication rather than an independent temporal test. The supplementary FreshQA analysis tests only the closed-book contrast. Multilingual, cross-domain, and longitudinal evaluation is needed to establish the broader range of the reported effect sizes.

Decision study. The date-policy study in §5.4 is internal validation on the questions and policy family the audit already studied. It quantifies model-specific trade-offs for date-conditioned policies, a Layer 1 consequence. Deployment improvement and the joint decision value of all three layers require further evaluation. Calibrated cross-encoders, MiniCheck-style scoring (Tang et al., 2024), Astute RAG (Wang et al., 2025), and context-compliance methods (Chen et al., 2026) provide stronger comparison points for a dedicated routing study.

Z Ethics and provenance

Data and licensing. We build the benchmark from RealTimeQA (Kasai et al., 2023). Its questions and annotations are released under CC BY-SA 4.0, which we take from the dataset's own datasheet because the upstream repository carries no licence file. The article text is a separate matter: upstream collects it under a United States fair-use rationale, which is a defence rather than a licence and does not transfer to us. The public artifact includes question manifests, cached model outputs, and article excerpts needed for the shown-text analyses. Its data-terms file distinguishes upstream questions, model outputs and third-party article text. The redistribution basis for article excerpts remains unconfirmed, as the upstream rationale does not itself grant redistribution rights. Document IDs, publication dates and source URLs support attribution and removal requests.

Model usage. We run open-weight models locally at temperature 0 and access the frontier references through the Anthropic and OpenAI APIs under their usage policies. We cache every response. The closed-book condition measures answer availability without retrieval, and we describe its variation by question date as a performance contrast over public questions rather than as an estimate of what a model knows, which keeps the analysis at the behavioural level. Mistral 123B ran in llama.cpp's llama-2 fallback prompt format for every condition because its GGUF file carries no chat template. A full rerun in the model's own format (Appendix X.7) does not resolve a change in the Layer 1 contrast and lowers its Static support interaction under quiz weighting, so we scope its Layer 3 magnitudes to the tested configuration.

Annotation. Three annotators independently completed the blinded pack of 395 correctness items and 900 evidence rows under the guidelines in `annotation/`. We combine labels by majority vote. A fourth annotator adjudicates contested items while blind to the vote split. All four volunteered without monetary compensation. Annotators saw 700-character excerpts, while generators saw 600. A further pack with the exact 600-character input (900 Task B and 630 masked rows) is prepared under the same guidelines and is not yet collected. Items omit model and condition identifiers, and the study collects no personal data. One author separately completed an unblinded internal adjudication of 90 stale-evidence cases. We identify those judgements as an internal audit and report them separately from the independent labels.

Representation and bias. RealTimeQA is English-language and Western-news-centric, which carries the geopolitical and editorial perspectives of its source outlets. We scope our conclusions to this setting and identify multilingual and non-Western evaluation as a priority for broader validation.

AI-use disclosure. The main cutoff panel contains seven locally run open-weight models and two Claude models used as frontier references. GPT-4o-mini contributes analyses that use its documented cutoff. The controlled-cutoff family uses locally run Llama-3.x checkpoints. Appendix U enumerates every panel. Authoring tools supported pipeline implementation and manuscript drafting. The public artifact preserves cached outputs and analysis provenance. We report replay outcomes and remaining validation gaps separately (Appendix A).