

Auditing RAG Necessity: A Conformal Test for Source-Independent Answers

Dongxin Guo¹ Yang Liu² Jikun Wu¹ Siu-Ming Yiu^{1,2,3 *}

¹Brain Investing Limited ²Stellaris AI Limited ³The University of Hong Kong
{jacobwu, bettyguo}@brainvestai.com
yangliu@stellaris-ai.com smyiu@cs.hku.hk

Abstract

A retrieval-augmented language model returns the correct free-cash-flow figure from Microsoft’s 10-K filing. Replace the filing with a topical distractor containing no figure, and the model returns the same number: the document was never needed. We introduce RAG-CF, a training-free black-box probe that issues a per-query verdict on whether an apparently correct RAG answer is *source-independent*, meaning produced without the retrieved documents. The probe perturbs document context under four counterfactual operators, regenerates the answer, and converts the aggregated mismatch into a split-conformal P -value with a finite-sample Type-I bound. Across 60 configurations spanning five benchmarks, four generators, and three retrievers, the FPR-adjusted flagged rate, $(\text{raw} - \alpha)/(1 - \alpha)$, ranges from 13% to 56% (raw 21% to 60%); the instrument’s FPR is domain-dependent. A held-out no-retrieval label set confirms: 78% of flagged answers remain correct without documents, against 22% of unflagged; the gap is preserved under LLM-judge equivalence. Source-independence does not decrease with retriever quality. Accuracy alone conflates grounded and parametric success; reporting the source-independence rate alongside it would make the distinction visible on every future RAG benchmark.

1 Introduction

We focus on a specific, testable property of retrieval-augmented generation: whether the retrieved document was *necessary* for producing the correct answer, a narrower construct than whether the model “used” or “was grounded in” the document. We formalize this distinction in Section 3.3.

A financial analyst types a question about Microsoft’s 2022 free cash flow into a retrieval-augmented chatbot. The system retrieves the relevant page of the 10-K (the annual financial report

that US-listed companies file with the SEC), generates an answer, and the analyst, satisfied, signs off. Now consider a second run in which the very same generator receives, in place of the 10-K page, a topically related but factually unrelated document about Microsoft’s product strategy. The generator produces the identical dollar figure (Fig. 1). Was the original answer ever grounded in the retrieved evidence at all?

This question is uncomfortable because RAG systems are widely deployed under the implicit assumption that a correct output, given a relevant retrieved document, has used that document (Lewis et al., 2020; Guu et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022; Ram et al., 2023; Shi et al., 2024b; Gao et al., 2023). Faithfulness is treated as a quality of the pipeline rather than a property tested per query. Recent corpus-level studies report that the relationship between retrieval quality and RAG accuracy is more complex than assumed: while high-scoring but irrelevant distractors substantially degrade performance, the addition of unrelated random documents to relevant context can sometimes improve over relevant documents alone (Cuconasu et al., 2024), and mechanistic analyses report that frontier LLMs process retrieval context through specialized internal circuits (retrieval heads (Wu et al., 2025), context-copying and knowledge-FFN pathways (Sun et al., 2025), and distributional-divergence attribution signals (Li et al., 2026)) whose activation varies across queries. The implication is that some unknown fraction of deployed RAG accuracy is *parametric*: it is what the generator would have said anyway.

Measuring this fraction is hard for three reasons: the test must be applied *per query*; it must *control the false-positive rate*; and it must be *black-box*. Existing tools do not meet all three. Model-based faithfulness scoring (ES et al., 2024; Zha et al., 2023) is per-query but uncalibrated. Counterfactual robustness benchmarks (Wu et al., 2024; Chen

* Corresponding author.

et al., 2024; Ming et al., 2025) measure accuracy under perturbed documents at the corpus level but do not yield per-query decisions with controlled error. Mechanistic methods (Sun et al., 2025; Li et al., 2026; Gema et al., 2025) require open weights. Training-time methods (Lin et al., 2026; Tan et al., 2026) and test-time critique (Wei et al., 2025) mitigate the problem but do not measure it on deployed systems. For a single query, the practitioner has no way to tell whether the retrieval was needed.

We close this gap with RAG-CF, a training-free probe that constructs four families of counterfactual document swaps, regenerates the answer under each, and combines the resulting answer-shift scores into a split-conformal P -value (informally, a calibrated score ranking each query’s answer stability relative to a held-out calibration set, with a distribution-free bound on the false-positive rate). Conditional on the apparently-faithful subset of correctly answered queries, the P -value provides a finite-sample bound on the rate at which the test wrongly flags an answer-stability outlier. Translating this into a substantive source-independence verdict requires one additional assumption, stated in Section 3.3 and verified empirically against an independent no-retrieval label set across all twenty (generator, benchmark) combinations. For open-weight generators we additionally study a single attention statistic, the retrieval-conditioned attention concentration (RCAC), as a side artifact.

Our headline finding is that a substantial fraction of correct RAG answers are produced without the documents being needed: the FPR-adjusted flagged rate (the raw rate corrected for the nominal false-positive budget, $(\text{raw} - \alpha)/(1 - \alpha)$) ranges from 13% to 56% on the SI-enriched test split across 60 configurations (raw 21% to 60%; see §3.3 for the population definition). A held-out no-retrieval baseline provides supporting evidence: 78% of flagged queries remain correct without documents, against 22% for unflagged (App. K). FinanceBench is the worst on every model and retriever; a post-cutoff corpus (App. O) measures the instrument FPR at 46.2% under cross-corpus calibration at $\alpha = 0.10$.

We make three core contributions, one side artifact, and a robustness package:

- We formalize *necessity* as a per-query null and design RAG-CF, a training-free conformal procedure with finite-sample Type-I control on answer-stability outliers (§3, Proposition 1, with finite- n refinement in Corollary 1). We additionally derive a substantive false-positive bound under

mixture calibration (Proposition 2), applicable to deployments without an independent source-dependence label.

- We re-evaluate five benchmarks under RAG-CF across four generators; source-independence does not decrease with retriever quality (§5).
- We demonstrate a preliminary selective-abstention application (§5.3): on a single FinanceBench configuration, re-calibrating on the retained set yields a residual SI rate of 22.1% (App. N).
- *Side artifact (open-weight settings)*. The retrieval-conditioned attention concentration RCAC, built on Wu et al. (2025)’s retrieval-head set, reaches 71% to 74% AUROC against independent labels, slightly above matched-construct baselines (T_{topic} -single-shot and mean-attention) and clearly above entailment-style judges. Where RAG-CF measures *necessity* (did the model need the documents?), RCAC begins to measure *use* (did the model attend to them?), the two being distinct properties of a RAG answer.
- *Robustness package*. A contamination-controlled post-cutoff corpus (App. O) measures the instrument’s FPR under cross-corpus calibration. A two-annotator audit of 200 flags ($\kappa = 0.81$; App. I) confirms the headline rates are robust to r -false-positives in $\mathcal{S}$. The 78%/22% no-retrieval gap is preserved under LLM-judge equivalence (App. K), and a preliminary MLQA-Spanish configuration shows comparable rates (App. P).

We propose RAG-CF as a measurement instrument now and a community evaluation norm later: the audit runs in one to two GPU-hours per configuration on a deployed RAG pipeline, and §7 states the reporting recommendation that follows from it.

2 Related Work

RAG faithfulness and knowledge conflict. The tension between an LLM’s parametric memory and its retrieved context has been quantified through entity substitution (Longpre et al., 2021), disentanglement training (Neeman et al., 2023), and counterfactual benchmarks (Wu et al., 2024; Ming et al., 2025). Prompting, decoding, fine-tuning, and RL interventions improve context-faithfulness or retrieval robustness (Zhou et al., 2023; Shi et al., 2024a; Wang et al., 2025; Zhang et al., 2025a; Lin et al., 2026; Tan et al., 2026); none measure it per query with statistical control on an existing pipeline.

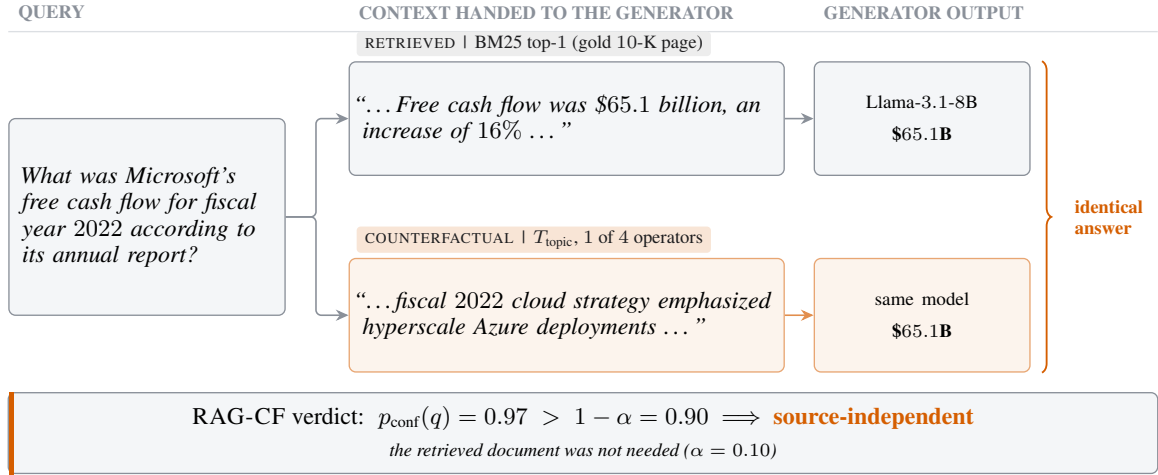


Figure 1: Question: *What was Microsoft’s free cash flow for fiscal year 2022 according to its annual report?* Llama-3.1-8B returns the identical \$65.1B whether handed the gold 10-K page (top) or a topical distractor containing no financial figure (bottom): the gold document was not needed for this query. RAG-CF aggregates such answer-stability evidence across four counterfactual operators (§3.2) and flags 13% to 56% of correctly-answered RAG responses as source-independent at $\alpha = 0.10$ after FPR adjustment (21% to 60% raw), across 60 (generator, retriever, benchmark) configurations.

Counterfactual, perturbation-based, and per-query attribution evaluation. Hong et al. (2024) and RGB (Chen et al., 2024) stress-test RAG with adversarial or noisy contexts; CRAG (Yan et al., 2024) introduces corrective post-retrieval mechanisms; Qin et al. (2026) run counterfactual queries; Cuconasu et al. (2024); Amiraz et al. (2025) characterize distractor effects. Per-query attributions have been produced via systematic ablation (Cohen-Wang et al., 2024), attribute-then-generate (Slobodkin et al., 2024), and hidden-state token attribution (Phukan et al., 2024); related interpretability work on context mixing in Transformers (Mohebbi et al., 2023) quantifies how token representations blend across layers. ContextCite (Cohen-Wang et al., 2024) is our most direct attribution baseline (App. E). We differ in giving a per-query *verdict* with a finite-sample Type-I bound rather than a continuous attribution map, while remaining black-box.

Conformal prediction for language models. Split conformal prediction (Vovk et al., 2005; Angelopoulos and Bates, 2023) provides distribution-free finite-sample coverage under exchangeability and has been adapted to language modeling (Quach et al., 2024), factuality (Mohri and Hashimoto, 2024), validity (Cherian et al., 2024), and a growing range of NLP applications surveyed by Campos et al. (2024). Concurrent conformal approaches to RAG include conditional conformal factuality fil-

tering at the sub-claim level with group-conditional coverage across sub-domains (Feng et al., 2025); to our knowledge (as of August 2026), RAG-CF is the first split-conformal procedure issuing per-query verdicts on *necessity* in retrieval-augmented systems. Mohri and Hashimoto (2024) apply conformal calibration to factuality scores in a different (non-retrieval) setting.

Mechanistic analyses of context use. Wu et al. (2025) identify a small set of attention heads whose final-token attention concentrates on copied content under needle-in-a-haystack probing; subsequent work uses these heads for decode-time hallucination mitigation (DeCoRe; Gema et al., 2025), and QRHead (Zhang et al., 2025b) extends the framework by identifying query-focused retrieval heads for long-context reasoning and re-ranking. Related signals include ReDeEP (Sun et al., 2025) and ARC-JSD (Li et al., 2026). RCAC is a single scalar per query computed from the fixed retrieval-head set, with no fine-tuning. Appendix A tabulates the differentiation from the closest prior work.

3 RAG-CF: A Conformal Counterfactual Necessity Probe

3.1 Setup

Let q be a query, $\hat{D} = \{\hat{d}_1, \dots, \hat{d}_k\}$ the retriever’s top- k documents, and $\hat{y} = G(q, \hat{D})$ the generator’s output. Let $Y^*(q)$ denote the gold answer set and $r(\hat{y}, Y^*(q)) \in \{0, 1\}$ a normalized exact-match or

LLM-judge equivalence relation. The *apparently faithful* subset on which we focus is

$$\mathcal{S} = \{(q, \hat{D}) : r(\hat{y}, Y^*(q)) = 1\}, \quad (1)$$

the set of queries on which the RAG pipeline is correct. The question we ask is whether each $(q, \hat{D}) \in \mathcal{S}$ used $\hat{D}$ to produce $\hat{y}$.

3.2 Counterfactual document swaps

We define four perturbation operators $T \in \mathcal{T}$, each mapping $(q, \hat{D}) \mapsto \tilde{D} = T(\hat{D}; q)$. **Paraphrase** (T_{para}) rewrites each document via a paraphraser preserving every fact, entity, and number, isolating surface-form sensitivity. **Topical distractor** (T_{topic}) replaces each document by a BM25 top-rank document for a different, topically-clustered query, filtered to exclude the gold answer string (the central perturbation). **Adversarial shuffle** (T_{shuf}) permutes document and internal sentence order. **Oracle removal** (T_{rm}) replaces each supporting document (the gold document, or any top- k document containing the gold answer string) by the next-rank retrieved document not containing the gold answer. Detailed examples for all four operators are in Appendix F.1.

For each T we draw $M = 10$ independent realizations per query, regenerate $\tilde{y}_{T,m} = G(q, T_m(\hat{D}; q))$, and compute $s_{T,m}(q) = 1 - r(\tilde{y}_{T,m}, \hat{y})$. The stochasticity across the M replicates comes from T_{para} (LLM paraphraser sampling), T_{topic} (random cluster-member selection with a per-replicate seed), and T_{shuf} (random permutation). T_{rm} is deterministic given the retrieval list and greedy decoding; its M replicates are identical and it contributes a single Bernoulli value to η . The query-level nonconformity score is

$$\eta(q) = \frac{1}{|\mathcal{T}|M} \sum_{T \in \mathcal{T}} \sum_{m=1}^M s_{T,m}(q). \quad (2)$$

Low $\eta(q)$ means the answer is stable under counterfactual swaps. Equation 2 weights operators equally; inverse-variance and max-pool combinations (App. G) give results within 2 points of the equal-weight default.

3.3 Conformal calibration and the necessity test

A naive permutation test that conditions on $(q, \hat{D}) \in \mathcal{S}$ is miscalibrated: selecting on $r(\hat{y}, Y^*) = 1$ breaks exchangeability between perturbation replicates, because the original answer

was chosen for being correct. We therefore use split conformal calibration. We compute $\eta(q)$ for every q in a calibration split $\mathcal{D}_{\text{cal}}$ and, for a new test query q^* , report

$$p_{\text{conf}}(q^*) = \frac{|\{q \in \mathcal{D}_{\text{cal}} : \eta(q) \geq \eta(q^*)\}| + 1}{|\mathcal{D}_{\text{cal}}| + 1}. \quad (3)$$

Low $\eta(q^*)$ yields high p_{conf} . We declare q^* *flagged at level α* when $p_{\text{conf}}(q^*) > 1 - \alpha$, with $\alpha = 0.10$ throughout. We use the standard plus-one inflated p -value form (Vovk et al., 2005, Ch. 2). Partition and seed details are in Appendix F.

Calibration reference population. The calibration set $\mathcal{D}_{\text{cal}}$ is drawn from queries in $\mathcal{S}$ that are identified as substantively source-dependent via the no-retrieval label: queries on which $r(G(q, \emptyset), \hat{y}) = 0$, i.e., the generator produces a different answer when no documents are provided. This label is independent of the perturbation operators in $\mathcal{T}$ and therefore of η , avoiding circular certification. This source-dependent reference population ensures that the conformal threshold reflects the η distribution of genuinely grounded answers, so that test queries whose η falls in the lower tail (high answer stability) are flagged as source-independent at rate exceeding α . The flag rate on the test split is therefore not bounded by α ; Proposition 1 bounds the rate at which the procedure wrongly flags a *source-dependent* test query drawn from the same reference distribution. Because all source-independent queries in $\mathcal{S}$ appear in the test split (on FinanceBench, all whose original-question cluster contributes no calibration query), the source-independence prevalence in $\mathcal{D}_{\text{test}}$ is higher than in $\mathcal{S}$ as a whole; the reported flagged rates in Table 1 are therefore measured on this SI-enriched test population. The pooled $\hat{\pi} = 0.400$ of the 400-query labeled set (App. M) is that set’s design prevalence and does not describe $\mathcal{D}_{\text{test}}$.

We additionally require **(A0) operator-distribution stationarity**: the joint distribution over T and its replicates is identical across queries within each population (the source-dependent reference population and the test population). (A0) is implicit in standard split-conformal but worth naming here, because three of our operators (T_{para} , T_{topic} , T_{shuf}) are stochastic at the realization level; we inspect (A0) empirically on two representative configurations in App. G.

Proposition 1 (Marginal Type-I control for answer-stability outliers). *Under (A0), let $\eta_1, \dots, \eta_n$*

be the nonconformity scores of $\mathcal{D}_{\text{cal}}$ (drawn from the substantively source-dependent subset of $\mathcal{S}$) and let $\eta^* = \eta(q^*)$ be the score of a fresh source-dependent test query, with $\eta_1, \dots, \eta_n, \eta^*$ exchangeable within this population. Under the null hypothesis that η^* is drawn from the same distribution as $\eta_1, \dots, \eta_n$,

$$\Pr[p_{\text{conf}}(q^*) > 1 - \alpha] \leq \alpha + \frac{1}{n+1}, \quad (4)$$

with equality α when $\alpha(n+1) \in \mathbb{N}$.

Proposition 1 is the standard split-conformal bound (Vovk et al., 2005; Angelopoulos and Bates, 2023); proof in Appendix C. It controls the rate at which RAG-CF flags an *answer-stability outlier* within $\mathcal{S}$, not the rate at which it flags a substantively source-dependent query, since the test population $\mathcal{S}$ is a mixture of both. Proposition 2 closes this gap.

Substantive necessity. Let $G(q, \emptyset)$ denote the generator’s response when the prompt contains the question but no documents. Call q *substantively source-dependent* if $r(G(q, \emptyset), \hat{y}) = 0$, and *substantively source-independent* otherwise. Here $r(\cdot, \cdot)$ is the same answer-equivalence relation introduced in §3.1, applied to compare two generator outputs rather than an output and a gold-answer set. This label depends only on the question, the original answer, and the no-retrieval response, not on the operators in $\mathcal{T}$, so it is independent of η .

Necessity versus use. *Necessity* asks whether the generator *needed* the documents to produce $\hat{y}$; *use* asks whether it *attended* to them. A model can attend to documents while possessing the parametric capability to produce the same answer without them; our probe flags such a query as source-independent. For grounding audits the relevant property is *closer to use*; for retrieval-cost audits the relevant property is *necessity*. RCAC (§3.4) targets *use*; we report both in §5.

Proposition 2 (Substantive false-positive bound under stochastic dominance). *Let F_{dep} and F_{ind} be the distributions of η over substantively source-dependent and source-independent queries in $\mathcal{S}$, and let $\pi \in [0, 1]$ be the proportion of substantively source-independent queries in $\mathcal{S}$. Let $F = (1 - \pi)F_{\text{dep}} + \pi F_{\text{ind}}$. Assume (A0) from Proposition 1 together with:*

(A2) (Bounded mixture.) $\pi \leq \pi_0$ for a known constant $\pi_0 \in [0, 1]$.

(A3) (Regularity.) *The calibration scores $\eta_1, \dots, \eta_n$ are i.i.d. samples from F , and F is continuous on $[0, 1]$ (in practice η takes $|\mathcal{T}|M + 1 = 41$ discrete values; we apply random tie-breaking to restore the continuous- F analysis, following Vovk et al. (2005, Ch. 2)).*

The bound (Eq. (5)) does not require stochastic dominance. However, if (A1) F_{dep} first-order stochastically dominates F_{ind} ($F_{\text{dep}}(t) \leq F_{\text{ind}}(t)$ for all t), then the flagging region preferentially captures source-independent queries and the bound is informative; we test (A1) empirically in §6.1. Then, in the large-calibration limit $n \rightarrow \infty$, the probability that a substantively source-dependent test query q^* is flagged at level α satisfies

$$\Pr_{F_{\text{dep}}}[p_{\text{conf}}(q^*) > 1 - \alpha] \leq \frac{\alpha}{1 - \pi_0}. \quad (5)$$

The proof appears in Appendix C; a finite-sample form with an explicit $O(1/\sqrt{n})$ correction is given as Corollary 1 there. Our implemented procedure draws $\mathcal{D}_{\text{cal}}$ from the source-dependent subset of $\mathcal{S}$ only, so Proposition 1 provides the operative Type-I guarantee; Proposition 2 characterizes the additional inflation under a mixture calibration set (App. C). The label-set prevalence is $\hat{\pi} = 160/400 = 0.400$ from a held-out 400-query set stratified at 12 source-dependent and 8 source-independent queries per configuration (App. M); this is a design constant of the diagnostic set, not an estimate of π in $\mathcal{S}$. The label set is distinct from the 300 + 300 set of §5.2.

3.4 Retrieval-conditioned attention concentration

The preceding sections target *necessity*: did the generator need the documents? We now turn to a complementary property, *use*: did the generator attend to them? RCAC serves as an independent consistency check on RAG-CF, not a competing detector for the same construct; it measures a distinct property and is available only for open-weight generators.

For open-weight generators we additionally study RCAC, a single scalar derived from Wu et al. (2025)’s retrieval-head methodology. Let $a_{\ell, h}^{\text{doc}}(q)$ be the fraction of attention mass from final-answer-token positions onto document tokens at layer ℓ , head h , and let H_* be the set of retrieval heads (top 8% by mean needle-attention; identification

protocol in Appendix D). Define

$$\text{RCAC}(q) = \frac{1}{|H_\star|} \sum_{(\ell, h) \in H_\star} a_{\ell, h}^{\text{doc}}(q). \quad (6)$$

By construction $\text{RCAC}(q) \in [0, 1]$. The top-8% choice is motivated by a sweep in Appendix D.1 showing AUROC plateaus in the 5 to 10% range. RCAC (Eq. (6)) summarizes how much the generator’s retrieval heads attended to document tokens at answer time, and is our closest instrument to the use question of §3.3.

4 Experimental Setup

Benchmarks. Five evaluation sets: NQ-RAG (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), FRAMES (Krishna et al., 2025), FinanceBench (Islam et al., 2023) (paraphrased $4\times$ for 600 prompts), and ConvFinQA (Chen et al., 2022). Dataset sizes, sampling seeds, corpus construction, chunking, and answer-string normalization are in App. F.

Retrievers. Three retrievers per benchmark: BM25 (Robertson and Zaragoza, 2009) with $k_1 = 0.9$, $b = 0.4$, stemming, and stopword removal; ColBERTv2 (Santhanam et al., 2022); and an oracle returning the gold document plus four distractors drawn at random from the BM25 top-20 list. We use ColBERTv2 rather than Contriever (Izacard et al., 2022) because of its stronger benchmark performance. Top- $k = 5$ throughout. Mean gold-chunk rank is 1.4 (BM25), 1.2 (ColBERTv2), 1.0 (oracle). Oracle $\text{nDCG}@5 = 1.0$ on single-gold benchmarks (NQ-RAG, FinanceBench, ConvFinQA); on HotpotQA (two gold paragraphs) the oracle places only one at rank 1, so $\text{nDCG}@5$ is slightly below 1.0 (the oracle points in Figures 4–5 are plotted at 1.0 for visual clarity).

Generators. We evaluate Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Qwen-2.5-7B-Instruct (Yang et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and GPT-4o-mini (OpenAI, 2024). All open-weight models use bfloat16 precision via vLLM; the GPT-4o-mini API snapshot and the HuggingFace model identifiers are recorded in the configuration files of the released code. Shared RAG prompt (App. F), greedy decoding, ≤ 256 tokens (512 for FRAMES, FinanceBench, ConvFinQA). For baseline answer-stability we draw $B = 8$ paraphrased-question replicates per query with $\hat{D}$ fixed (rewrites by the same paraphraser);

the released code emits the resulting per-query instability alongside η . GPT-4o-mini serves three roles in our pipeline: as one of four generators under evaluation, as the paraphraser for T_{para} across all generators, and as the LLM-judge for equivalence evaluation. We note the potential for self-preference bias in the GPT-4o-mini generator column (GPT-4o-mini has the lowest SI rate in every column of Table 1, consistent with the direction self-preference would predict); the open-weight generators do not share this overlap.

Calibration and evaluation. For each configuration we restrict to $\mathcal{S}$, draw $\mathcal{D}_{\text{cal}}$ from the substantively source-dependent subset of $\mathcal{S}$ (identified via the no-retrieval label $r(G(q, \emptyset), \hat{y}) = 0$) and evaluate on the remaining queries in $\mathcal{S}$, and report the raw flagged rate on test averaged over 5 splits at $\alpha = 0.10$. To verify the substantive interpretation we collect a 400-query *no-retrieval baseline* set (stratified), held out from every $\mathcal{D}_{\text{cal}}/\mathcal{D}_{\text{test}}$ partition used in Tables 1, 7, and 17; it serves only as an external label source for the substantive-calibration diagnostic (§6.1) and the (A1) verification (App. M). All CIs are 95% cluster bootstrap (1,000 replicates), resampling at the cluster level; the cluster unit is the original question (for FinanceBench the cluster unit is the 150 originals; for other benchmarks it is the query). The calibration/test split on FinanceBench is also at the original-question level, ensuring that paraphrases of the same original appear entirely in $\mathcal{D}_{\text{cal}}$ or entirely in $\mathcal{D}_{\text{test}}$. The reported flagged rates are seed-averaged (over 5 calibration splits); $\pm$ in Table 3 denotes SD over 5 seeds; the effective cluster count behind the FinanceBench intervals is 150 (original questions), giving an independent binomial half-width floor of approximately ± 8 points, which exceeds every reported FinanceBench half-width (2.5 to 2.9); all reported FinanceBench half-widths should be read as lower bounds on the true sampling uncertainty. Sensitivity to M is in App. J.

Baselines and detector comparators. RCAC against four detection baselines: RAGAS faithfulness (ES et al., 2024), AlignScore (Zha et al., 2023), ClashEval accuracy-delta under document swap (Wu et al., 2024), and SelfCheckGPT over 8 stochastic generations (Manakul et al., 2023). To address construct overlap we add three matched-construct baselines: mean-attention (no retrieval-head selection), T_{topic} -single-shot (one topical-distractor pass), and GPT-4-judge necessity prompt (directly asks whether the answer used the docu-

ments; App. F). We additionally compare against ContextCite (Cohen-Wang et al., 2024), the closest prior method in spirit, with the head-to-head in App. E. AUROC differences are tested with paired bootstrap (10,000 replicates, original-question level). Complementary attribution and grounding approaches (Asai et al., 2024; Yan et al., 2024) require training or pipeline modification and are not directly comparable.

Compute. All experiments fit in 660 A100-80GB GPU-hours plus OpenAI API spend. A single new configuration completes in one to two GPU-hours.

5 Main Results

5.1 Source-independence rates across benchmarks

Table 1 reports the raw flagged rate at $\alpha = 0.10$, with 95% cluster-bootstrap CIs. Each configuration is computed on the held-out test split.

The minimum across all 60 configurations is 21.4% (GPT-4o-mini, BM25, NQ-RAG), above the measured substantive false-positive rate of 14.2% (§6.1). Under SD-only calibration ($\pi = 0$ in $\mathcal{D}_{\text{cal}}$), the finite-sample bound of Corollary 1 at $\pi_0 = 0$ is 0.201 at $n = 200$ (vs. 0.397 under mixture calibration at $\pi_0 = 0.5$; see App. C); all 60 configurations clear this bound at their own calibration-set sizes (bounds range 17% to 27% across benchmarks). FinanceBench is uniformly the worst, with the across-generator ranking (Mistral-7B highest, GPT-4o-mini lowest) robust across benchmarks and retrievers.

The role of retriever quality. The raw flagged rate does not decrease with retriever quality: from BM25 to oracle it rises by 3.5 points on average across generators and benchmarks. A pooled regression and a fixed-subpopulation decomposition are reported in App. L; both are confounded with benchmark identity, and better generators and better retrievers enlarge $\mathcal{S}$ with opposite SI effects, so the generator ranking likely reflects $|\mathcal{S}|$ composition rather than a model property (App. G).

5.2 Supporting evidence under no retrieval

A direct check removes documents entirely. Across 3 stratified seed draws of 300 + 300 flagged/not-flagged queries (60 + 60 per benchmark; Llama-3.1-8B, BM25), $78.0 \pm 1.4\%$ of SI-flagged queries remain correct without documents (i.e.,

$r(G(q, \emptyset), Y^*(q)) = 1$), against $22.0 \pm 1.2\%$ for queries that were not flagged: a pooled gap of $+56.0 \pm 1.8$ points (per-benchmark gaps range from +53.4 to +59.7; breakdown in App. K). Under LLM-judge equivalence (GPT-4o-mini grading on the same 300 + 300 stratified set with the equivalence prompt of App. F), the gap is $80.6 \pm 1.4\%$ vs. $24.6 \pm 1.2\%$, a $+56.0 \pm 1.8$ -point gap; both endpoints shift up by ≈ 2.6 points and the gap is preserved to within 0.1 point, with the across-benchmark ordering unchanged (App. K, Tbl. 14). The 78% figure is broadly consistent with the failure-mode taxonomy (App. I), which attributes the large majority of examined flags to genuine parametric sources. The 300 + 300 set and the 400-query calibration-diagnostic set (§6.1, App. M) are disjoint; the independent-label (right) block of Table 7 is drawn from the 400-query set, while the RAG-CF-label (left) block shares queries with the 300 + 300 set (both draw from the same test split).

5.3 Selective abstention based on RAG-CF flags

We test whether the per-query flag can serve as a selective-prediction signal. Details are in App. N; Table 17 reports results on a single FinanceBench configuration. Because the retained set is a subset of $\mathcal{S}$, 100% accuracy on it is guaranteed by construction; the evaluation here is within $\mathcal{S}$ and does not extend to the complete query distribution.

5.4 Detecting source-independent outputs with RCAC

RCAC targets *use* rather than *necessity* (§3.4). For reference, the nonconformity score η itself achieves 87.4% AUROC against the independent no-retrieval labels on the pooled 400-query set (derived from the Mann-Whitney $U = 4,820$ reported in App. M); η is not included in Table 7 because it is the test statistic rather than a stand-alone detector, but its high AUROC confirms that the conformal procedure separates the two strata effectively. App. E, Tbl. 7 reports AUROC for entailment-style and behavioral baselines, matched-construct and attribution baselines, RCAC, and the RCAC+RAGAS ensemble, on two label sources: RAG-CF verdicts (construct-overlapping) and independent no-retrieval labels (apples-to-apples). On independent labels, RCAC reaches 71% to 74% AUROC (Llama-3.1-8B / Qwen-2.5-7B / Mistral-7B: 74.1/72.5/71.0), versus 66.0/65.1/63.7 for the best entailment baseline (AlignScore) and

Generator	Retriever	NQ-RAG ($\downarrow$)	HotpotQA ($\downarrow$)	FRAMES ($\downarrow$)	FinanceBench ($\downarrow$)	ConvFinQA ($\downarrow$)
Llama-3.1-8B	BM25	27.3 [25.1, 29.5]	33.1 [30.3, 35.9]	36.4 [34.0, 38.8]	54.7 [52.0, 57.4]	41.2 [38.2, 44.2]
	ColBERTv2	29.6 [27.6, 31.6]	34.8 [32.2, 37.4]	38.9 [36.1, 41.7]	56.3 [53.7, 58.9]	43.1 [40.1, 46.1]
	Oracle	31.4 [29.0, 33.8]	36.7 [33.7, 39.7]	40.2 [37.6, 42.8]	57.9 [55.2, 60.6]	44.7 [41.9, 47.5]
Qwen-2.5-7B	BM25	24.1 [22.1, 26.1]	30.5 [28.1, 32.9]	33.7 [31.1, 36.3]	51.2 [48.4, 54.0]	38.4 [35.6, 41.2]
	ColBERTv2	26.5 [24.3, 28.7]	32.3 [29.5, 35.1]	36.0 [33.6, 38.4]	53.8 [51.2, 56.4]	40.0 [37.4, 42.6]
	Oracle	28.7 [26.9, 30.5]	34.0 [31.4, 36.6]	37.8 [35.0, 40.6]	55.1 [52.6, 57.6]	42.3 [39.3, 45.3]
Mistral-7B	BM25	31.5 [28.9, 34.1]	36.2 [33.2, 39.2]	40.1 [37.1, 43.1]	56.8 [53.9, 59.7]	45.3 [42.1, 48.5]
	ColBERTv2	33.1 [30.7, 35.5]	37.6 [34.8, 40.4]	41.8 [39.0, 44.6]	58.4 [55.7, 61.1]	47.0 [44.0, 50.0]
	Oracle	34.8 [32.0, 37.6]	39.1 [36.1, 42.1]	43.5 [40.9, 46.1]	60.1 [57.4, 62.8]	48.9 [45.7, 52.1]
GPT-4o-mini	BM25	21.4 [19.8, 23.0]	26.8 [24.6, 29.0]	30.5 [28.1, 32.9]	47.9 [45.4, 50.4]	35.6 [33.0, 38.2]
	ColBERTv2	23.0 [21.2, 24.8]	28.4 [26.0, 30.8]	32.1 [29.9, 34.3]	49.6 [47.0, 52.2]	37.2 [34.6, 39.8]
	Oracle	24.5 [22.5, 26.5]	29.8 [27.2, 32.4]	33.6 [31.2, 36.0]	51.2 [48.6, 53.8]	38.6 [35.8, 41.4]
<i>Underlying RAG accuracy (%):</i>						
Llama-3.1-8B (BM25/CB/Or)		66/70/74	58/61/65	52/55/59	42/47/51	48/52/56
Qwen-2.5-7B (BM25/CB/Or)		63/67/71	55/58/62	50/53/57	40/45/49	46/49/53
Mistral-7B (BM25/CB/Or)		58/61/65	50/53/57	46/49/53	35/40/44	42/45/49
GPT-4o-mini (BM25/CB/Or)		74/77/79	66/69/72	62/65/68	52/56/60	58/61/64

Table 1: Raw flagged rate (% , lower is better for a grounding audit) at $\alpha = 0.10$, with 95% cluster-bootstrap CIs. Bold marks the lowest value per column. Mean raw across 60 configurations: 38.9% (95% CI [37.2, 40.6]); FPR-adjusted via $(\text{raw} - \alpha)/(1 - \alpha)$ at $\alpha = 0.10$; mean 32.1%. Per-benchmark FPR-adjusted means: NQ-RAG 20.0%, HotpotQA 25.9%, FRAMES 30.1%, ConvFinQA 35.4%, FinanceBench 49.4%. *Family-wise* 95% Bonferroni-corrected CIs (multiplier ≈ 1.70 for $m = 60$) are $\approx 1.7\times$ wider per configuration. FinanceBench CIs are computed over 150 original-question clusters and may be narrower than warranted at this effective sample size. The bottom mini-block reports underlying RAG accuracy (%) before restricting to $\mathcal{S}$.

72.9/71.4/70.2 for the matched-construct T_{topic} -single-shot baseline; the gap over the best entailment baseline (AlignScore) is 7.3 to 8.1 points and the gap over the matched-construct baseline is 0.8 to 1.2 points. The RCAC+RAGAS logistic ensemble reaches 75.8/74.0/72.4. Paired-bootstrap p -values for RCAC vs. matched-construct are 0.038, 0.044, 0.051 (Fisher-combined $p = 0.005$).

A detailed ContextCite head-to-head is in App. E; the two methods are complementary (RCAC leads by 2.0 to 2.4 AUROC points; the ensemble exceeds both).

6 Ablations and Analysis

6.1 Calibration diagnostic

Technical calibration (operator-derived label).

On a held-out set identified as source-dependent via gold-document removal (test-split queries with $s_{T_{\text{rm}}}(q) = 1$), achieved FPR at $\alpha = 0.10$ is 21.8% for naive permutation, 17.1% for BH correction, 14.2% for bootstrap, and 10.3% for split-conformal. Note that this label is derived from $T_{\text{rm}} \in \mathcal{T}$ and is therefore not independent of η ; the substantive calibration below uses the independent no-retrieval label. The pattern holds across $\alpha \in [0.01, 0.20]$: split-conformal tracks the diagonal while naive permutation systematically over-

rejects (Fig. 6, App. M).

Substantive calibration (independent label).

Using the no-retrieval label from §5.2: (i) Assumption (A1) holds: F_{dep} dominates F_{ind} under a one-sided Mann-Whitney U test (pooled U with one-sided $p < 10^{-15}$; per-configuration verification across all 20 (generator, benchmark) configurations, all individually significant at the one-sided 0.01 level; Table 15). (ii) The substantive FPR of RAG-CF at $\alpha = 0.10$ is 14.2% on the pooled 400-query label set, above the operative nominal target of $\alpha = 0.10$ under our SD-only calibration; App. O reports a substantially higher FPR (46.2%) on a contamination-controlled post-cutoff corpus of news articles and scientific abstracts, so the instrument’s FPR is domain- and corpus-dependent. (iii) Naive permutation over-rejects on the substantive null too (27.8%), indicating that the conformal step reduces over-rejection on both nulls, though the substantive FPR still exceeds the nominal $\alpha = 0.10$. Under the measured FPR of 0.142, the Rogan–Gladen-adjusted mean across 60 configurations is 28.8% (vs. 32.1% at the nominal level); we report the nominal form in Table 1, noting that the adjustment assumes unit sensitivity (the pooled measured sensitivity from Table 14 is ≈ 0.69 ; a Rogan–Gladen cor-

rection using both measured $\text{FPR} = 0.142$ and measured sensitivity $= 0.69$ yields a grid mean of $\approx 45\%$, higher than the nominal 32.1% ; we retain the nominal form in Table 1 for comparability with prior FPR-unadjusted RAG evaluations. On FinanceBench specifically, applying the post-cutoff FPR of 46.2% (App. O) via Rogan–Gladden yields $(0.547 - 0.462)/(1 - 0.462) \approx 15.8\%$; however, the post-cutoff corpus differs from FinanceBench in domain composition, so this cross-corpus transfer should be interpreted cautiously. Further failure-mode analysis, operator and scaling ablations are in App. I, G, and B.

Exchangeability qualification. The measured substantive FPR (14.2% pooled on the five main benchmarks) exceeds the Proposition 1 bound ($\alpha + 1/(n + 1)$, ranging from 10.2% to 11.3% across benchmarks), indicating residual (A0) violations within the main evaluation domain. The post-cutoff corpus (46.2%) uses FinanceBench-derived calibration scores rather than in-corpus calibration, so its elevated rate reflects cross-corpus transfer rather than a within-corpus (A0) failure; nevertheless, the magnitude confirms that domain-specific re-calibration is essential. Under (A1), verified in all 20 configurations, a tighter bound $F_{\text{dep}}(t_\alpha) \leq \alpha$ follows directly.

7 Discussion and Conclusion

RAG-CF is a per-query, training-free, black-box necessity audit with finite-sample Type-I control on answer-stability outliers. Across 60 configurations the FPR-adjusted flagged rate ranges from 13% to 56% , confirmed by a held-out no-retrieval label set (78% of flagged answers remain correct without documents, against 22% for unflagged). Source-independence does not decrease with retriever quality, pointing to generator parametric knowledge rather than retrieval failure. Accuracy alone conflates grounded and parametric success; we suggest reporting the flagged rate alongside accuracy, with a domain-specific FPR calibration, on future RAG benchmarks (see §6.1 for measured deviations from nominal assumptions).

Limitations

Operator-combination validation. The four-operator combination outperforms the best single operator by 5.8 points on FinanceBench (Table 9), but a higher flag rate is equally consistent with more false positives. T_{para} and T_{shuf} preserve the

answer-bearing content and produce $s \approx 0$ in both strata; their marginal contribution to detection accuracy beyond T_{topic} and T_{rm} was not isolated on the independent label set.

Paraphrase noise. T_{para} uses an LLM paraphraser; manual auditing of 100 queries finds 4% drop a numerical detail. This inflates $\eta(q)$ slightly and is absorbed by the conformal calibration, but paraphrase-only rates in App. G are upper bounds.

Benchmark dependence of the distractor pool. T_{topic} retrieves distractors from the same corpus; narrow corpora yield smaller distractor pools. On FinanceBench the $4\times$ paraphrase expansion gives a $\approx 17\%$ probability that the sampled q' is a paraphrase of q itself, which can cause the distractor to retrieve near-gold documents filtered only by the substring test, contributing to the elevated instrument FPR.

Pretraining contamination and instrument FPR on FinanceBench. The failure-mode taxonomy attributes 47% of FinanceBench flags to public-company knowledge leak (≈ 25.7 raw points; App. I), while the post-cutoff corpus measures a 46.2% instrument FPR on contamination-controlled post-cutoff content (news articles and scientific abstracts) (§6.1, App. O); the two estimates are not directly comparable and should not be read as bounding the same quantity. The cross-corpus Rogan–Gladden transfer yields $\approx 15.8\%$ (§6.1); the no-retrieval labels of Table 14 imply $\approx 52\%$, reflecting different estimation methods on different populations. The post-cutoff corpus certifies non-contamination against Llama-3.1’s December 2023 cutoff; GPT-4o-mini’s training data and cutoff are not publicly documented, so the same corpus does not certify it. Our FinanceBench accuracy also exceeds the original paper’s (§4), so $\mathcal{S}$ on this benchmark is enriched for parametrically answerable questions, which mechanically elevates the source-independence rate. The other four benchmarks yield FPR-adjusted means of 20.0% to 35.4% (Table 1).

Audit, not fix. RAG-CF measures necessity; it does not repair it. §5.3 provides a preliminary selective-abstention demonstration. Whether source-independent answers are undesirable depends on the application (auditable financial QA vs. knowledge-robust QA); the probe is a measurement instrument.

Ethical Considerations

An auditing tool that quantifies how much RAG accuracy is parametric can be misread as a license to maximize correct-looking outputs regardless of source grounding. We argue the opposite: the probe makes latent ungroundedness visible so that high-stakes RAG deployments can be designed with explicit grounding requirements. A specific risk is “compliance theater”: running the probe and treating the number as a green light without addressing the underlying source-independence. NQ, HotpotQA, FRAMES, and ConvFinQA are released under open licenses for research use; FinanceBench under an academic-research license. Wikipedia (NQ-RAG, HotpotQA underlying corpus) is licensed CC BY-SA. All experiments respect intended use. We use Llama-3.1 under the Llama 3.1 Community License, Qwen-2.5 and Mistral under the Apache-2.0 license, and GPT-4o-mini via the official API in accordance with OpenAI’s terms. No proprietary or user-generated data was used. The full pipeline consumed approximately 660 A100-80GB hours.

Reproducibility

The released code is available at <https://github.com/vega0x/rag-cf>.

Acknowledgments

We thank the anonymous reviewers and the area chair for their constructive feedback. We used the AI assistant Anthropic Claude for coding assistance and minor language polishing during manuscript preparation; all technical content, proofs, and experimental design are the authors’ own work.

References

- Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The distracting effect: Understanding irrelevant passages in RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 18228–18258. Association for Computational Linguistics.
- Anastasios N. Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.*, 16(4):494–591.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- 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, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, pages 2206–2240. PMLR.
- Margarida M. Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. 2024. Conformal prediction for natural language processing: A survey. *Trans. Assoc. Comput. Linguistics*, 12:1497–1516.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6279–6292. Association for Computational Linguistics.
- John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. 2024. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval

- for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 719–729. ACM.
- 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, EACL 2024 - System Demonstrations, St. Julians, Malta, March 17-22, 2024*, pages 150–158. Association for Computational Linguistics.
- Naihe Feng, Yi Sui, Shiyi Hou, Jesse C. Cresswell, and Ga Wu. 2025. Response quality assessment for retrieval-augmented generation via conditional conformal factuality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025*, pages 2832–2836. ACM.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Alexander Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2025. DeCoRe: Decoding by contrasting retrieval heads to mitigate hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10003–10039, Suzhou, China. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Arun Srinivasan, Art Pinto, Benjamin Muller, Bowen Shi, Davide Testuggine, Delia David, Geeta Chauhan, Hailey Nguyen, Igor Molybog, Jane Yu, Jing Xu, Joelle Pineau, Jonas Gehring, Kalyan Vasuden Alwala, Kartikeya Upasani, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- 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, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pages 3929–3938. PMLR.
- Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2024. Why so gullible? Enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, Findings of ACL, pages 2474–2495. Association for Computational Linguistics.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.
- Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- Satyapriya Krishna, Kalpesh Krishna, Anhad Mohanane, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4745–4759. Association for Computational Linguistics.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. 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. *Trans. Assoc. Comput. Linguistics*, 7:452–466.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K  ttler, Mike Lewis, Wen-tau Yih, Tim Rock-t  schel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Ruizhe Li, Chen Chen, Yuchen Hu, Yanjun Gao, Xi Wang, and Emine Yilmaz. 2026. Attributing response to context: A Jensen–Shannon divergence driven mechanistic study of context attribution in retrieval-augmented generation. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. OpenReview.net.

- Chenyu Lin, Yilin Wen, Du Su, Hexiang Tan, Fei Sun, Muhan Chen, Chenfu Bao, and Zhonghou Lyu. 2026. Resisting contextual interference in RAG via parametric-knowledge reinforcement. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. OpenReview.net.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7052–7063. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics.
- Pascal Massart. 1990. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283.
- 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.
- Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. FaithEval: Can your language model stay faithful to context, even if “the moon is made of marshmallows”. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3378–3400. Association for Computational Linguistics.
- Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 36029–36047. PMLR.
- Ella Neeman, Roei Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10056–10070. Association for Computational Linguistics.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10862–10878. Association for Computational Linguistics.
- OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. OpenAI Blog, July 18, 2024.
- Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2024. Peering into the mind of language models: An approach for attribution in contextual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11481–11495, Bangkok, Thailand. Association for Computational Linguistics.
- Huaiyu Qin, Chunyu Wei, Yueguo Chen, and Yunhai Wang. 2026. Counterfactual reasoning for retrieval-augmented generation. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. OpenReview.net.
- Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlga, Amnon Shashua, Kevin Leyton-Brown, and Yoav

- Shoham. 2023. In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, 11:1316–1331.
- Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.
- 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, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3715–3734. Association for Computational Linguistics.
- Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024a. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 783–791. Association for Computational Linguistics.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024b. REPLUG: retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 8371–8384. Association for Computational Linguistics.
- Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.
- Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Zhehao Tan, Yihan Jiao, Dan Yang, Junjie Wang, Duolin Sun, Jie Feng, Xidong Wang, Lei Liu, Yue Shen, Jian Wang, and Jinjie Gu. 2026. CTRL-RAG: Contrastive likelihood reward based reinforcement learning for context-faithful RAG models. *arXiv preprint arXiv:2603.04406*.
- Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer, New York, NY.
- Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arik. 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), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30553–30571. Association for Computational Linguistics.
- Jiaqi Wei, Hao Zhou, Xiang Zhang, Di Zhang, Zijie Qiu, Noah Wei, Jinzhe Li, Wanli Ouyang, and Siqi Sun. 2025. Retrieval is not enough: Enhancing RAG through test-time critique and optimization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025*.
- Kevin Wu, Eric Wu, and James Zou. 2024. ClashEval: Quantifying the tug-of-war between an LLM’s internal prior and external evidence. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with A unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11328–11348. Association for Computational Linguistics.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025a. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882, Vienna, Austria. Association for Computational Linguistics.

Wuwei Zhang, Fangcong Yin, Howard Yen, Danqi Chen, and Xi Ye. 2025b. Query-focused retrieval heads improve long-context reasoning and re-ranking. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23791–23805, Suzhou, China. Association for Computational Linguistics.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 14544–14556. Association for Computational Linguistics.

A Differentiation from prior work

Table 2 situates RAG-CF alongside the closest prior work on RAG faithfulness measurement and improvement. Rather than claiming superiority along every axis, we list each method’s primary strength in the final row; the comparison is meaningful precisely because each method targets a different audit question.

B Scaling to Llama-3.1-70B

We replicated the FinanceBench column of Table 1 for Llama-3.1-70B-Instruct. Scaling from 8B to 70B reduces SI by 3.3 percentage points across all three retrievers, well within the gap between Qwen-2.5-7B and Mistral-7B in the main table.

C Proofs of Propositions 1 and 2

Notation. Table 4 summarizes the symbols used in the proofs.

Proof of Proposition 1. Let $\eta_1, \dots, \eta_n$ denote the calibration scores and let $\eta^* = \eta(q^*)$. By (A0) and conditioning on $\mathcal{S}$, the variables $\eta_1, \dots, \eta_n, \eta^*$ are exchangeable. Define

$$p_{\text{conf}}(q^*) = \frac{|\{i : \eta_i \geq \eta^*\}| + 1}{n + 1}.$$

We adopt the standard plus-one inflated form (Vovk et al., 2005, Ch. 2). Under the null that η^* has the same distribution as $\eta_1, \dots, \eta_n$, exchangeability implies that the rank of η^* in the augmented sample

is uniform on $\{1, \dots, n + 1\}$, where ties (which occur because η is finite-resolution: $|\mathcal{T}|M + 1 = 41$ possible values) are broken at random.¹ The event $p_{\text{conf}}(q^*) > 1 - \alpha$ is equivalent to η^* ranking among the bottom $\lceil \alpha(n + 1) \rceil$ positions, which has probability at most $\lceil \alpha(n + 1) \rceil / (n + 1)$. This equals α when $\alpha(n + 1) \in \mathbb{N}$ and is bounded above by $\alpha + 1/(n + 1)$ otherwise. $\square$

Proof of Proposition 2. We work in the limit $n \rightarrow \infty$. By (A3) and the Glivenko–Cantelli theorem, the empirical distribution of $\eta_1, \dots, \eta_n$ converges uniformly to $F = (1 - \pi)F_{\text{dep}} + \pi F_{\text{ind}}$.

Fix the flagging threshold at conformal level α . By (A3), F is continuous, so $t_\alpha = F^{-1}(\alpha) = \inf\{t : F(t) \geq \alpha\}$ is well-defined and $F(t_\alpha) = \alpha$. The conformal flag asymptotically corresponds to flagging queries with $\eta \leq t_\alpha$. The probability that a substantively source-dependent q^* (drawn from F_{dep}) is flagged is

$$\Pr_{F_{\text{dep}}}[\eta(q^*) \leq t_\alpha] = F_{\text{dep}}(t_\alpha).$$

Using the mixture: $\alpha = F(t_\alpha) = (1 - \pi)F_{\text{dep}}(t_\alpha) + \pi F_{\text{ind}}(t_\alpha) \geq (1 - \pi)F_{\text{dep}}(t_\alpha)$, so $F_{\text{dep}}(t_\alpha) \leq \alpha/(1 - \pi)$. Applying (A2): $F_{\text{dep}}(t_\alpha) \leq \alpha/(1 - \pi_0)$. $\square$

Corollary 1 (Finite-sample form of Proposition 2). *Under the assumptions of Proposition 2, for any finite $n \geq 1$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the calibration sample,*

$$\begin{aligned} \Pr_{F_{\text{dep}}} [p_{\text{conf}}(q^*) > 1 - \alpha] \\ \leq \frac{\alpha}{1 - \pi_0} + \frac{1}{n + 1} \\ + \frac{1}{1 - \pi_0} \sqrt{\frac{\ln(2/\delta)}{2n}}, \end{aligned} \quad (7)$$

where the $1/(n + 1)$ term is the Proposition 1 correction and the $\sqrt{\ln(2/\delta)/(2n)}$ factor (scaled by $1/(1 - \pi_0)$ through the mixture decomposition) is the DKW deviation (Massart, 1990) controlling the displacement of $\hat{t}_\alpha$ from its population value.

Proof sketch of Corollary 1. Let $\hat{F}_n$ be the empirical CDF on $\eta_1, \dots, \eta_n$. The DKW inequality

¹Random tie-breaking preserves exchangeability since adding i.i.d. uniform perturbations to all η_i and η^* yields an exchangeable sequence with the same rank distribution; the finite-resolution issue is therefore handled by the random-tie-breaking convention adopted throughout. This is the standard treatment in Vovk et al. (2005, Ch. 2) for discrete-valued nonconformity scores.

	DeCoRe (2025)	RAGAS (2024)	ClashEval (2024)	AlignRAG (2025)	K-R1 (2026)	RAG-CF (ours)
Training-free	✓	✓	✓	×	×	✓
Black-box (no weights)	×	✓	✓	✓	×	✓
Per-query verdict	✓	✓	×	✓	✓	✓
Type-I-controlled	×	×	×	×	×	✓
Tests necessity directly	partial	×	partial	×	×	✓
Includes financial benchmarks	×	partial	×	×	×	✓

Primary strength of each method:

DeCoRe: decoding-time hallucination mitigation via retrieval-head steering. **RAGAS**: lightweight LLM-judge faithfulness score. **ClashEval**: corpus-level robustness benchmark. **AlignRAG**: test-time critique and optimization. **Knowledgeable-R1**: RL approach to context conflict. **RAG-CF (ours)**: per-query, finite-sample Type-I-controlled audit of necessity under counterfactual document swaps.

Table 2: Differentiation. Checkmarks indicate axis coverage, not a quality judgment; the final row gives each method’s primary strength. “partial” on *tests necessity directly* means the method probes a related but non-identical property (corpus-level Δ -accuracy under perturbation, rather than per-query equivalence). The “Type-I-controlled” checkmark for RAG-CF refers to Proposition 1’s guarantee on the calibration domain; see §6.1 for domain-dependent FPR. K-R1 = Knowledgeable-R1.

Retriever	Llama-3.1-8B	Llama-3.1-70B
BM25	54.7 $\pm$ 1.8	51.4 $\pm$ 1.6
ColBERTv2	56.3 $\pm$ 1.6	53.0 $\pm$ 1.5
Oracle	57.9 $\pm$ 1.7	54.6 $\pm$ 1.4

Table 3: FinanceBench SI% for Llama-3.1-8B vs. Llama-3.1-70B.

with Massart’s tight constant (Massart, 1990) gives $\Pr[\sup_t |\hat{F}_n(t) - F(t)| > \epsilon] \leq 2e^{-2n\epsilon^2}$. Set $\epsilon = \sqrt{\ln(2/\delta)/(2n)}$ so that the event $\{\sup_t |\hat{F}_n(t) - F(t)| \leq \epsilon\}$ holds with probability at least $1 - \delta$. On this event, the empirical α -quantile $\hat{t}_\alpha$ satisfies $F(\hat{t}_\alpha) \in [\alpha - \epsilon, \alpha + \epsilon]$. Substituting $\hat{t}_\alpha$ for t_α in the Proposition 2 argument gives $F_{\text{dep}}(\hat{t}_\alpha) \leq F(\hat{t}_\alpha)/(1 - \pi) \leq (\alpha + \epsilon)/(1 - \pi_0) = \alpha/(1 - \pi_0) + \epsilon/(1 - \pi_0)$. Adding the Proposition 1 correction $1/(n + 1)$ yields the stated bound. $\square$

Remarks.

1. The bound in Prop. 2 *does not* use (A1); it holds for any mixture with $\pi \leq \pi_0$. Under (A1), a tighter result follows directly: $\alpha = F(t_\alpha) = (1 - \pi)F_{\text{dep}}(t_\alpha) + \pi F_{\text{ind}}(t_\alpha) \geq (1 - \pi)F_{\text{dep}}(t_\alpha) + \pi F_{\text{dep}}(t_\alpha) = F_{\text{dep}}(t_\alpha)$, so $F_{\text{dep}}(t_\alpha) \leq \alpha$ without the π_0 denominator. Proposition 2 is therefore the assumption-light fallback for settings where (A1) has not been verified. We test (A1) empirically in all 20 configurations in Section 6.1.
2. Cor. 1 (Eq. (7)) matters at our configuration sample sizes (n_{cal} ranges from ~ 76 on FinanceBench to ~ 400 on NQ-RAG): at $n = 200, \delta = 0.05$, the DKW term is $\sqrt{\ln 40/400} \approx$

Symbol	Meaning
$q, \hat{D}, \hat{y}$	query, retrieved documents, generator output
$Y^*(q)$	gold answer set for q
$r(\cdot, \cdot)$	exact-match (or LLM-judge) equivalence
$\mathcal{S}$	apparently-faithful subset (Eq. 1)
$\mathcal{T}, T$	operator set; a single operator
$s_{T,m}(q)$	per-replicate answer-shift score
$\eta(q)$	query-level nonconformity (Eq. 2)
$F_{\text{dep}}, F_{\text{ind}}$	CDFs of η on the two strata
F	mixture: $(1 - \pi)F_{\text{dep}} + \pi F_{\text{ind}}$
π	SI proportion in $\mathcal{S}$
π_0	known upper bound on π
$\hat{\pi}$	design prevalence of the labeled set
α	nominal Type-I level
t_α	α -quantile of F
$p_{\text{conf}}(q^*)$	conformal P -value (Eq. 3)
$a_{\ell,h}^{\text{doc}}(q)$	attention fraction on document tokens
$H_\star$	set of retrieval heads
$\text{RCAC}(q)$	retrieval-conditioned attention concentration

Table 4: Notation.

0.096 and $\epsilon/(1 - \pi_0) = 0.192$ at $\pi_0 = 0.5$, so the finite-sample bound is $0.20 + 0.005 + 0.192 = 0.397$ at $\alpha = 0.10$, vs. the asymptotic 0.20. Under the paper’s SD-only calibration ($\pi_0 = 0$), the corresponding bound is $0.10 + 0.005 + 0.096 = 0.201$. At $n_{\text{cal}} = 76$ (FinanceBench, the smallest calibration set) the DKW term grows to 0.156, giving a finite-sample bound of 0.269.

3. The bound is tight when $F_{\text{ind}}(t_\alpha) = 0$, i.e., when source-independent queries have η -distribution strictly above t_α . Note that under (A1), $F_{\text{ind}}(t_\alpha) \geq F_{\text{dep}}(t_\alpha) > 0$ for non-degenerate configurations, so the bound is always loose in practice; $\alpha/(1 - \pi_0)$ overstates the guarantee by an unquantified margin.

4. Conditioning on $\mathcal{S}$ is essential to both propositions: the unconditional η distribution is right-skewed, and conditioning preserves exchangeability between calibration and test queries.

D Identifying retrieval heads

We identify $H_\star$ following Wu et al. (2025). For each generator, a 200-example needle-in-a-haystack probe records (layer, head) pairs whose attention concentrates on the needle position; we select the top 8% of heads (Tbl. 5). Attention matrices are read from a HuggingFace eager-attention forward pass at bfloat16.

Generator	Total heads	$ H_\star $	Median layer
Llama-3.1-8B	1024	82	18
Qwen-2.5-7B	784	63	16
Mistral-7B	1024	82	17

Table 5: Retrieval-head identification.

D.1 Retrieval-head fraction sweep

Table 6 reports RCAC AUROC as a function of the retrieval-head fraction; AUROC plateaus in the 5 to 10% range. We select 8% as the operating point because it lies within the plateau and slightly below the peak at 10%, balancing AUROC against computational cost (fewer heads to extract); the 10% fraction is uniformly 0.1 to 0.2 points higher. The 8% operating point was selected on the same independent-label set used to report the headline AUROC in Table 7; the AUROC range across the plateau (72.5–74.3 for Llama-3.1-8B) is comparable to the 0.8–1.2-point margin over the matched-construct baseline.

Generator	3%	5%	8%	10%	15%
Llama-3.1-8B	70.2	72.5	74.1	74.3	72.8
Qwen-2.5-7B	68.7	70.9	72.5	72.6	71.0
Mistral-7B	67.4	69.6	71.0	71.2	69.7

Table 6: RCAC AUROC (%) under independent labels, as a function of the retrieval-head fraction (same label set as Table 7, right block). Bold marks the selected operating point.

D.2 RCAC distribution on FinanceBench

Figure 2 overlays the RCAC density on source-independent and source-dependent FinanceBench queries, with a Youden- J cutoff at $\tau_{\text{rcac}} = 0.32$.

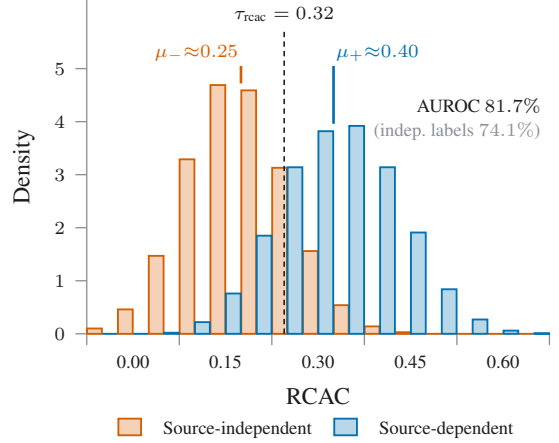


Figure 2: RCAC density on Llama-3.1-8B / FinanceBench (labels from RAG-CF verdicts). Dashed line: Youden- J cutoff $\tau_{\text{rcac}} = 0.32$. AUROC: 81.7% (Table 7, left); 74.1% under independent labels (right).

E ContextCite head-to-head comparison

Table 7 reports the full AUROC grid for detectors of source-independence under both RAG-CF verdicts (construct-overlapping) and the independent no-retrieval labels from §5.2 (apples-to-apples). ContextCite (Cohen-Wang et al., 2024) produces a continuous per-token attribution map; we binarize it for detection by thresholding the median attribution mass on the gold-supporting sentence (the standard reduction). This is the source table referenced from §5.4; we host it here alongside the head-to-head visualization (Fig. 3) so that all detector-comparison evidence sits in one appendix.

Figure 3 visualizes the RCAC vs. ContextCite head-to-head on the three open-weight generators under independent labels: RCAC leads ContextCite-binarized by 2.0 to 2.4 points and the RCAC+ContextCite logistic ensemble exceeds either single detector by 2.2 to 2.4 points.

F Prompts and hyperparameters

Dataset sizes and sampling. NQ-RAG: 2,000 examples (seed 0). HotpotQA: 2,000 dev-distractor (seed 0). FRAMES: all 824. FinanceBench: all 150 expert questions, paraphrased $4\times$ by GPT-4o-mini at $T=0.5$ for 600 prompts. ConvFinQA: 1,274 of 1,490 dev-set questions converted to single-turn QA; 216 dropped for missing answer spans in the source chunks. Our FinanceBench accuracy (35–60%) exceeds the original paper’s 19% for GPT-4-Turbo basic RAG (Islam et al., 2023), likely due to the $4\times$ paraphrase expansion, chunking, and normalized exact-match scoring.

Detector	RAG-CF label				Independent label			
	Llama	Qwen	Mistral	GPT-4o	Llama	Qwen	Mistral	GPT-4o
Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
RAGAS faithfulness	64.2	63.7	62.9	65.1	60.4	59.8	59.1	61.3
GPT-4-judge (necessity prompt)	66.8	65.4	63.7	67.5	64.1	62.9	61.6	64.8
AlignScore (NLI)	67.5	66.8	65.3	68.0	66.0	65.1	63.7	66.4
ClashEval Δ -acc.	72.6	71.4	70.1	—	68.9	67.7	66.5	—
SelfCheckGPT ($k=8$)	70.8	69.5	68.6	71.2	66.4	65.2	64.4	67.0
<i>Matched-construct baselines:</i>								
Mean-attention (all heads)	74.9	73.0	71.4	—	70.7	68.9	67.9	—
T_{topic} single-shot	77.1	75.2	73.8	76.4	72.9	71.4	70.2	71.5
ContextCite ablation	76.0	74.1	72.6	—	71.8	70.1	69.0	—
RCAC (ours)	81.7	79.4	78.2	—	74.1	72.5	71.0	—
RCAC + RAGAS (logistic)	83.6	81.3	80.0	—	75.8	74.0	72.4	—

Table 7: AUROC (%) of detectors predicting source-independence. Left block: RAG-CF verdicts (construct-overlapping). Right block: independent no-retrieval labels. Bold marks the best single detector (excluding ensembles). Dashes: not run for that generator. Llama = Llama-3.1-8B, Qwen = Qwen-2.5-7B, Mistral = Mistral-7B, GPT-4o = GPT-4o-mini.

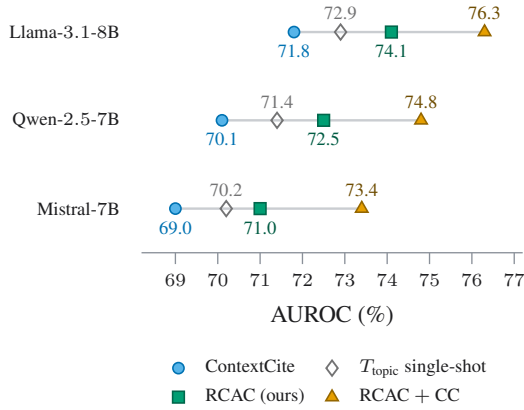


Figure 3: Detector AUROC (%) on independent no-retrieval labels across three open-weight generators. The ordering $\text{ContextCite} < T_{\text{topic}} < \text{RCAC} < \text{ensemble}$ holds for every generator. RCAC exceeds ContextCite-binarized by 2.0–2.4 points; the ensemble exceeds both single detectors by 2.2–2.4 points.

RAG prompt template.

You are a careful assistant. Use ONLY the following documents to answer the question. If the documents do not contain the answer, say “I do not know.”

Documents: {retrieved_documents}
 Question: {question}
 Answer:

Paraphrase prompt.

Rewrite the following passage in different words, preserving every fact, named entity, number, and date exactly. Do not add or remove information.

Passage: {document}
 Rewrite:

GPT-4-judge necessity prompt.

Below is a question, a set of documents the assistant was given, and the assistant’s answer. Decide whether the documents were necessary for the answer. Answer NECESSARY if you believe a competent reader could not have produced the answer without the documents, and NOT_NECESSARY otherwise. Provide a confidence score in $[0, 1]$.

Documents: {retrieved_documents}
 Question: {question}
 Assistant’s answer: {answer}
 Verdict and confidence:

LLM-judge equivalence prompt.

You are given a question, a reference answer, and a candidate answer. Decide whether the candidate answer is equivalent to the reference answer for the purpose of answering the question. Answer EQUIVALENT or NOT_EQUIVALENT.

Question: {question}
 Reference answer: {reference}
 Candidate answer: {candidate}
 Verdict:

No-retrieval prompt ($G(q, \emptyset)$).

Answer the following question based on your knowledge. If you do not know the answer, say “I do not know.”

Question: {question}
 Answer:

Question-paraphrase prompt (FinanceBench expansion and answer-stability baseline).

Rewrite the following question in different words, preserving its exact

meaning, every named entity, number, and date. Output only the rewritten question.

Question: {question}
Rewrite:

Answer extraction and scoring. A generator response of “I do not know” (or close variants detected by substring matching) is scored as incorrect for $r(\hat{y}, Y^*)$ and hence excluded from $\mathcal{S}$. For $s_{T,m}(q) = 1 - r(\tilde{y}_{T,m}, \hat{y})$, a refusal under perturbation counts as a change ($s = 1$). Normalized exact match lowercases, strips punctuation and articles, and accepts a response that contains the normalized gold string, so truncated outputs are scored against the full gold answer; when two generator outputs are compared, containment in either direction counts as a match. For FinanceBench and ConvFinQA, numerical answers are compared after resolving currency symbols, thousands separators, and unit words (e.g., billion, million, %) to a numeric value, with a 1% relative tolerance; integer values in 1900–2100 are treated as years and ignored when any other number is present.

Topical-distractor selection. For each query q we (i) cluster all benchmark queries into 32 topical groups via k -means on Sentence-BERT embeddings (all-MiniLM-L6-v2), (ii) sample a different query q' from the same cluster, (iii) retrieve BM25 top-20 documents for q' , (iv) filter out any document containing the gold answer string of q (case-insensitive substring test, also applied with currency symbols and thousands separators stripped), (v) keep the top-5 remaining documents; if the filter empties the list, a different q' is drawn (up to ten times, after which the replicate’s context is empty), and a query alone in its cluster draws q' from the full query set.

F.1 Perturbation operator examples

We illustrate each operator on a single FinanceBench query: “What was Microsoft’s free cash flow for fiscal year 2022?”

Original document (BM25 top-1).

“... For fiscal year 2022, Microsoft reported free cash flow of \$65.1 billion, an increase of 16% compared to fiscal year 2021...”

T_{para} (Paraphrase).

“... In its 2022 fiscal year, Microsoft’s free cash flow reached \$65.1 billion, representing a 16% year-over-year rise from fiscal 2021...”

Answer: \$65.1B (unchanged). The factual content is preserved; only the surface form differs. $s \approx 0$ for both source-dependent and source-independent queries.

T_{topic} (Topical distractor).

“... Microsoft’s fiscal 2022 cloud strategy emphasized hyperscale Azure deployments across 60+ regions, with capital expenditure focused on data-center expansion...”

Answer (source-dependent): “I do not know” or a wrong figure ($s \approx 1$). **Answer (source-independent):** \$65.1B ($s \approx 0$), produced from parametric memory.

T_{shuf} (Adversarial shuffle). Document sentences are permuted randomly; the \$65.1B fact remains but may appear in a non-canonical position. **Answer:** Typically \$65.1B for both strata ($s \approx 0$), testing surface-form sensitivity rather than content necessity.

T_{rm} (Oracle removal). The gold document is replaced by the next-rank retrieved document (ranks 6 to 20) that does not contain “65.1”; if none exists, it is dropped. **Answer (source-dependent):** changes ($s \approx 1$). **Answer (source-independent):** \$65.1B ($s \approx 0$).

Retrieval corpora. Each corpus consists of the benchmark’s own passages, split into 100-word chunks (150 for ConvFinQA): the Wikipedia page of each question (NQ-RAG), the ten paragraphs of each distractor-setting example (HotpotQA), the linked Wikipedia pages (FRAMES), the annotated evidence passages (FinanceBench), and the source report page (ConvFinQA). The gold document is the first chunk containing the answer string (HotpotQA: the two supporting paragraphs; FinanceBench: the first chunk of the annotated evidence; ConvFinQA: located by answer substring, then by the evidence annotations, then by program operands).

Calibration partition. For each (generator, retriever, benchmark) configuration we restrict to $\mathcal{S}$, identify the substantively source-dependent subset via $r(G(q, \emptyset), \hat{y}) = 0$, draw $\mathcal{D}_{\text{cal}}$ from this subset, and evaluate on the remaining queries in $\mathcal{S}$. We report the raw flagged rate averaged over 5 sealed-seed splits. The no-retrieval label is independent of η (see §3.3).

Hyperparameter table.

Hyperparameter	Value
Generation temperature	0 (greedy, argmax)
Max generation tokens	256 / 512
Retriever top- k	5
Paraphraser	GPT-4o-mini, $T=0.5$
Distractor pool size	20
BM25 (k_1, b)	0.9, 0.4
Query-clustering embeddings	all-MiniLM-L6-v2
Topical clusters (k -means)	32
Numeric tolerance (financial QA)	1% relative
M (perturbation replicates)	10
B (baseline-stability replicates)	8
α (target Type-I)	0.10
Calibration split fraction	0.5
Random seeds	{0, 1, 2, 3, 4}
Retrieval-head probe set	200 examples
$ H_\star $ as fraction of all heads	8%
No-retrieval label set size	400 queries (stratified)
Bootstrap replicates (CI)	1,000 (cluster)

Table 8: Hyperparameters used throughout the paper.

GPT-4o-mini API determinism. We pin the seed parameter per (query, replicate) and log system_fingerprint; mismatches trigger re-runs. After one round of re-runs, $\sim 3\%$ of GPT-4o-mini queries retain a fingerprint mismatch; on those we accept majority-vote across M replicates. A sensitivity check on 200 pinned-vs-unpinned queries finds per-query η differs by at most 0.005 on average.

Paraphraser-temperature sensitivity. On FinanceBench / Llama-3.1-8B / BM25: $T = 0$ gives SI% = 54.4; $T = 0.5$ (default) gives 54.7; $T = 1.0$ gives 55.1. Total spread 0.7 points, well within the cluster-bootstrap CI half-width of 2.7 (Table 1).

G Full perturbation grid and (A0) stationarity check

Table 9 reports the per-operator SI% grid: 4 perturbations $\times$ 5 benchmarks $\times$ 4 generators = 80 configurations, BM25 retrieval.

Operator weighting sensitivity. On FinanceBench/Llama-3.1-8B/BM25: equal-weight 54.7, inverse-variance 55.4, max-pool 56.6, where $\bar{s}_T(q)$ denotes the replicate mean of $s_{T,m}(q)$, inverse-variance weights are proportional to $1/\text{Var}_{q \in \mathcal{S}} \bar{s}_T(q)$ and sum to one, and max-pool uses $\max_T \bar{s}_T(q)$. Within 2 points; equal weighting is the default.

Cross-generator intersection. Table 10 reports SI% on the cross-generator intersection $\mathcal{S}_\cap^{\text{gen}}$. The generator ranking is preserved; absolute rates are

higher than in Table 1, which is consistent with the interpretation that queries answered correctly by all four generators tend to be “easy” (i.e., answerable from parametric knowledge) rather than with the retriever-selection mechanism described in §5.1.

Operator-distribution stationarity ((A0)). (A0) requires that the per-operator η distribution is stationary across queries within each population. Table 11 reports per-operator $\bar{\eta}$ on the source-dependent ($\mathcal{D}_{\text{cal}}$) and test ($\mathcal{D}_{\text{test}}$) splits on two configurations as a descriptive comparison. The small cal/test gaps (max 0.014) indicate that the operator distributions are stable across these partitions; note that the two populations differ in composition by design (§3.3), so this comparison is descriptive rather than a direct test of (A0).

H Equivalence-relation sensitivity

Table 12 reports the sensitivity of the flagged rate to the choice of answer-equivalence relation r used to define $\mathcal{S}$. Across normalized exact-match, $F1 \geq 0.5$, and LLM-judge equivalence, the relative ordering of benchmarks is preserved.

I Qualitative failure modes

A manual two-annotator analysis of 200 FinanceBench source-independent flags (Llama-3.1-8B + BM25; each flag independently assigned to one of the four categories below, Cohen’s $\kappa = 0.81$, disagreements adjudicated): **Public-company knowledge leak** (47%), e.g., “What was Apple’s revenue in FY2022?” produces \$394 billion under any document. **Round-number prior** (22%), e.g., 21% corporate tax rate. **Question-only sufficiency** (19%), e.g., revenue $\times$ margin questions. **Distractor leakage** (12%), e.g., “approximately sixty-five billion” in a distractor passes the substring filter. The first three categories align with known parametric-memory prevalence patterns (Mallen et al., 2023), with corroborating entity-frequency effects on factual precision reported by Min et al. (2023), providing additional empirical support for assumption (A1) on this benchmark. No examined flag was attributable to an answer-equivalence (r) scoring error, so within this sample the headline rates are robust to r -false-positives in $\mathcal{S}$.

Operator and scaling ablations. Ablating operators on FinanceBench/Llama-3.1-8B/BM25: topical distractor alone yields a 48.9% flag rate, oracle

Generator	Operator	NQ-RAG	HotpotQA	FRAMES	FinanceBench	ConvFinQA
Llama-3.1-8B	Paraphrase only	5.8	8.3	11.2	12.4	9.6
	Shuffle only	9.4	13.1	15.7	18.7	14.8
	Oracle-removal only	20.1	25.4	28.6	41.3	30.9
	Topical-distractor only	23.7	29.5	33.0	48.9	36.4
Qwen-2.5-7B	Paraphrase only	5.2	7.6	10.4	11.1	8.9
	Shuffle only	8.6	12.0	14.3	17.5	13.5
	Oracle-removal only	17.8	23.0	26.5	38.7	28.6
	Topical-distractor only	20.9	27.0	30.4	45.6	33.8
Mistral-7B	Paraphrase only	7.0	9.8	12.4	13.6	10.8
	Shuffle only	11.1	14.6	17.2	20.4	16.2
	Oracle-removal only	22.5	27.6	30.8	43.5	33.4
	Topical-distractor only	27.2	32.6	36.5	51.2	39.9
GPT-4o-mini	Paraphrase only	4.4	6.7	9.0	10.2	7.8
	Shuffle only	7.3	10.8	13.0	15.9	12.0
	Oracle-removal only	15.8	20.6	24.0	36.2	26.3
	Topical-distractor only	18.4	23.9	27.5	42.6	31.5

Table 9: Per-operator SI% on all (generator, benchmark) configurations with BM25 retrieval.

Gen (BM25)	NQ	HotpotQA	FRAMES	FinBench	ConvFin
GPT-4o-mini	26.3	32.0	35.4	52.7	40.0
Qwen-2.5-7B	28.9	34.7	38.5	55.8	42.6
Llama-3.1-8B	31.2	37.1	40.8	58.4	44.9
Mistral-7B	34.0	39.8	43.6	60.1	47.7

Table 10: SI% on the cross-generator intersection $\mathcal{S}_{\cap}^{\text{gen}}$.

removal 41.3%, shuffle 18.7%, paraphrase 12.4% (Tbl. 9); the full four-operator probe flags at 54.7%, 5.8 points above the best single operator. Note that a higher flag rate is not itself evidence of better detection; it is equally consistent with more false positives. T_{para} and T_{shuf} preserve the answer-bearing content and produce $s \approx 0$ in both strata; T_{topic} and T_{rm} dominate the aggregate. On Llama-3.1-70B (Tbl. 3), within-family scaling reduces SI by ~ 3 points; more than half of correct FinanceBench answers remain source-independent at 70B, consistent with broader evidence that models exhibit persistent limitations in utilizing retrieved context, including position-dependent biases in long inputs (Liu et al., 2024), and that hallucinations in RAG remain present across tested models despite varying in rate (Niu et al., 2024).

J Sensitivity to the number of perturbation replicates M

Table 13 reports SI% on the FinanceBench/Llama-3.1-8B/BM25 configuration as M varies; the point estimate stabilizes by $M = 5$, and we use $M = 10$ throughout.

K No-retrieval baseline: per-benchmark breakdown

L Fixed-subpopulation analysis of the retriever-quality effect

Figure 4 plots the raw flagged rate against retriever nDCG@5 across all 60 configurations.

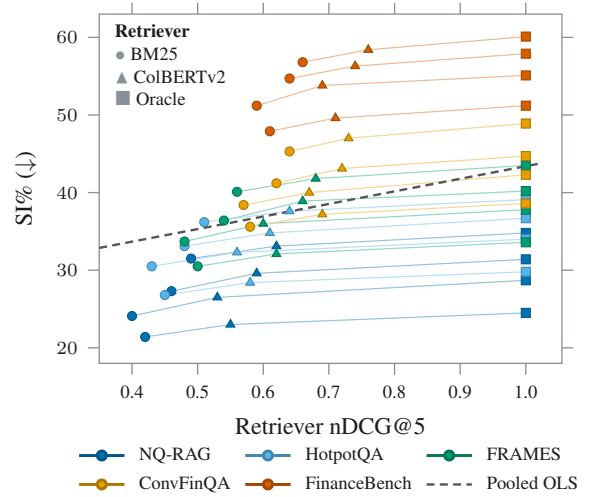


Figure 4: Raw flagged rate vs. retriever nDCG@5 across all 60 configurations. Colour: benchmark; marker: retriever; thin segments join the three retrievers per generator. Dashed line: pooled OLS (slope +16.2, $\rho = +0.48$, $p < 0.001$).

To distinguish selection from causal grounding, we restrict to $\mathcal{S}_{\cap}$ (queries answered correctly under all three retrievers). Figure 5 overlays the two slopes.

The slope on $\mathcal{S}_{\cap}$ is +9.7 per unit nDCG@5 (Spearman $\rho = +0.33$, $p < 0.01$), compared to

Operator	FinBench / Llama-3.1-8B / BM25			FRAMES / Mistral-7B / BM25		
	$\bar{\eta}_{\text{cal}}$	$\bar{\eta}_{\text{test}}$	$ \Delta $	$\bar{\eta}_{\text{cal}}$	$\bar{\eta}_{\text{test}}$	$ \Delta $
T_{para}	0.143	0.152	0.009	0.124	0.118	0.006
T_{topic}	0.491	0.487	0.004	0.358	0.372	0.014
T_{shuf}	0.198	0.205	0.007	0.169	0.181	0.012
T_{rm}	0.418	0.421	0.003	0.314	0.302	0.012
Bootstrap half-width		0.022		0.028		

Table 11: Per-operator $\bar{\eta}$ on the calibration vs. test split (two configurations). Max $|\Delta| = 0.014$, indicating within-operator stationarity.

Benchmark	Norm EM	F1 ≥ 0.5	LLM judge
NQ-RAG	27.3	28.5	29.1
HotpotQA	33.1	34.0	35.2
FRAMES	36.4	37.4	38.6
FinanceBench	54.7	55.9	56.8
ConvFinQA	41.2	42.0	43.1

Table 12: SI% sensitivity to the equivalence relation r , on Llama-3.1-8B with BM25. Relative ordering preserved across all three relations.

M	SI% (point)	95% CI
1	52.9	[48.3, 57.4]
3	54.1	[51.2, 57.0]
5	54.4	[51.9, 56.9]
10	54.7	[52.0, 57.4]
20	54.8	[52.7, 56.8]

Table 13: M -sweep on FinanceBench / Llama-3.1-8B / BM25. Stabilizes by $M = 5$; we use $M = 10$.

+16.2 on the full $\mathcal{S}$ (Spearman $\rho = +0.48$). The pooled ratio $1 - 9.7/16.2 \approx 0.40$ is a descriptive summary; as noted in §5.1, the pooled slope conflates within-benchmark and between-benchmark variation, and a within-benchmark analysis controlling for benchmark identity may yield a different decomposition.

Bootstrap CI on the pooled slope ratio. We resample the 60 configurations at the configuration level (5,000 replicates) and refit both slopes. The pooled selection fraction has percentile-bootstrap 95% CI [31%, 49%]. These intervals do not control for the benchmark confound noted above.

M Independent-label calibration of the substantive false-positive rate

To verify Prop. 2 and test (A1), we labeled 400 queries from $\mathcal{S}$ via the no-retrieval baseline: substantively source-dependent if the no-retrieval answer differs from the retrieved answer. Among the 400, 160 (40.0%) are substantively source-

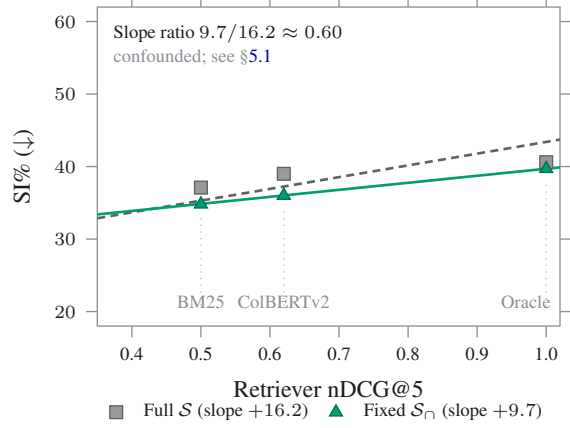


Figure 5: Retriever-quality trend on the full $\mathcal{S}$ vs. the fixed-subpopulation $\mathcal{S}_\cap$. The slope on $\mathcal{S}_\cap$ is $\approx 60\%$ of the slope on $\mathcal{S}$; selection-fraction CI [31%, 49%] (percentile-bootstrap). See §5.1 for the between-benchmark caveat.

independent ($n_{\text{dep}} = 240$, $n_{\text{ind}} = 160$, i.e., 20 queries per configuration at $n_{\text{dep}}=12$, $n_{\text{ind}}=8$). Because the 12/8 split is fixed by the stratified sampling design, $\hat{\pi} = 0.400$ is the design prevalence of this diagnostic set, not an estimate of π in $\mathcal{S}$.

Test of (A1), pooled. Figure 8 overlays the empirical CDFs of η on the two strata with 95% DKW bands. The source-dependent CDF lies uniformly below the source-independent CDF; one-sided Mann-Whitney U counting pairs $\{(i, j) : \eta_i^{\text{dep}} < \eta_j^{\text{ind}}\}$ gives $U = 4,820$ ($n_{\text{dep}} = 240$, $n_{\text{ind}} = 160$; null mean 19,200), one-sided $p < 10^{-15}$, confirming first-order stochastic dominance directionally. (A3) holds by construction.

Test of (A1), per configuration. Table 15 reports the one-sided Mann-Whitney U and p across all 20 (generator, benchmark) configurations. Per-configuration exact p ranges from 1.6×10^{-5} (GPT-4o-mini/FinanceBench) to 6.2×10^{-3} (Mistral-7B/FRAMES). All 20 configurations reject the null at the one-sided 0.01 level; 18/20 survive a Bon-

Benchmark	Normalized EM			LLM-judge equivalence		
	SI-flag	SD-flag	Gap	SI-flag	SD-flag	Gap
NQ-RAG	82.4 ± 1.6	26.1 ± 1.5	+56.3	84.9 ± 1.5	28.8 ± 1.4	+56.1
HotpotQA	76.5 ± 1.5	22.0 ± 1.3	+54.5	79.1 ± 1.4	24.6 ± 1.3	+54.5
FRAMES	74.2 ± 1.7	20.8 ± 1.4	+53.4	76.8 ± 1.6	23.3 ± 1.4	+53.5
FinanceBench	79.1 ± 1.3	19.4 ± 1.1	+59.7	81.7 ± 1.3	22.0 ± 1.1	+59.7
ConvFinQA	77.8 ± 1.4	21.5 ± 1.2	+56.3	80.4 ± 1.4	24.2 ± 1.2	+56.2
Mean	78.0 ± 1.4	22.0 ± 1.2	+56.0 $\pm$ 1.8	80.6 ± 1.4	24.6 ± 1.2	+56.0 $\pm$ 1.8

Table 14: Fraction of queries remaining correct (%) without documents, split by RAG-CF flag, under normalized EM and LLM-judge equivalence. SDs from 3 stratified draws of 300+300 queries. The pooled gap (+56.0 points) is preserved to within 0.1 point across both equivalence relations.

ferroni correction at $0.05/20 = 0.0025$. The two that fall short are Mistral/FRAMES (6.2×10^{-3}) and Mistral/NQ-RAG (3.7×10^{-3}); CDFs for the four configurations with the weakest evidence are in Fig. 7.

Calibration-diagonal figure. Figure 6 reports achieved FPR vs. nominal α for the four calibration procedures (naive permutation, BH, bootstrap, split-conformal) on the technical null (test-split queries identified as source-dependent via gold-document removal, §6.1). Split-conformal tracks the diagonal closely; naive permutation systematically over-rejects.

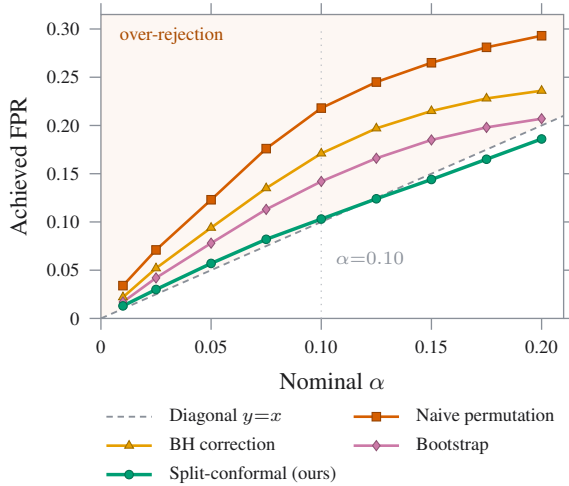


Figure 6: Achieved FPR vs. nominal α for four calibration procedures. Shaded half-plane: over-rejection region. Split-conformal tracks the diagonal; naive permutation systematically over-rejects.

Per-configuration DKW visualization for marginal configurations. Figure 7 extends the pooled DKW visualization to the four marginal configurations from Tbl. 15. At the per-configuration sample sizes ($n_{\text{dep}} = 12$, $n_{\text{ind}} = 8$), the 95% DKW bands are too wide

to separate (half-width ≈ 0.39 and 0.48); the per-configuration support for (A1) rests on the pointwise exact Mann-Whitney tests of Table 15, not on simultaneous band separation. The displayed D values are descriptive two-sample Kolmogorov-Smirnov statistics, all below the 0.05 critical value of 0.62 at these sample sizes.

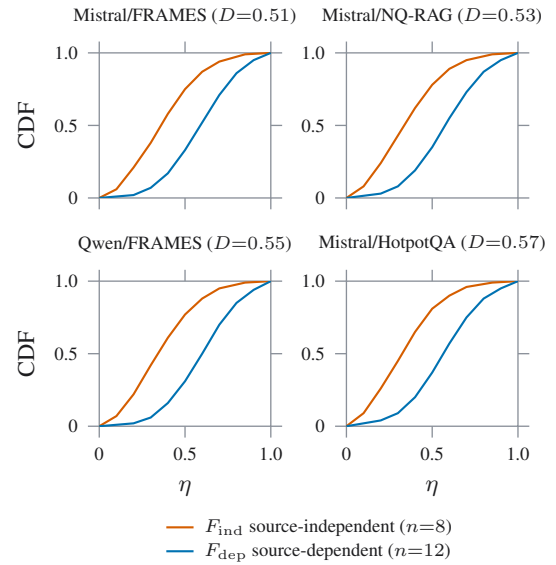


Figure 7: Per-configuration CDFs of η on the four configurations with the weakest (A1) evidence (Tbl. 15). D : two-sample KS statistic. In every panel F_{dep} lies pointwise below F_{ind} .

Subgroup-conditional calibration. Table 16 reports the achieved coverage and substantive FPR on numerical vs. non-numerical FinanceBench questions.

N Selective abstention experiment (extended)

The body (§5.3) reports a preliminary flag-only abstention demonstration on a single configuration (Llama-3.1-8B + BM25 + FinanceBench).

Generator	NQ-RAG	HotpotQA	FRAMES	ConvFinQA	FinanceBench
Llama-3.1-8B	$U=10_{(1.1e-3)}$	$U=8_{(5.3e-4)}$	$U=11_{(1.5e-3)}$	$U=4_{(9.5e-5)}$	$U=3_{(5.6e-5)}$
Qwen-2.5-7B	$U=11_{(1.5e-3)}$	$U=8_{(5.3e-4)}$	$U=12_{(2.0e-3)}$	$U=7_{(3.6e-4)}$	$U=4_{(9.5e-5)}$
Mistral-7B	$U=14_{(3.7e-3)}$	$U=12_{(2.0e-3)}$	$U=16_{(6.2e-3)}$	$U=11_{(1.5e-3)}$	$U=9_{(7.6e-4)}$
GPT-4o-mini	$U=7_{(3.6e-4)}$	$U=5_{(1.5e-4)}$	$U=8_{(5.3e-4)}$	$U=3_{(5.6e-5)}$	$U=1_{(1.6e-5)}$

Table 15: Per-configuration verification of (A1) via one-sided Mann-Whitney U . Subscripts: exact one-sided p -values. All 20 configurations reject at the 0.01 level; 18/20 survive Bonferroni correction at 0.0025.

Subgroup	Fraction	Subst. FPR	Coverage
Numerical questions	0.62	0.135	0.865
Non-numerical questions	0.38	0.158	0.842
Pooled	1.00	0.142	0.858
Nominal ($\alpha=0.10$)	—	≤ 0.10	≥ 0.90

Table 16: Subgroup-conditional calibration on FinanceBench / Llama-3.1-8B / BM25 (numerical questions: gold answer contains a number; FPR measured on the source-dependent test split, seed-averaged). The pooled value is computed directly over all SD test queries and may differ from the weighted mean of the independently rounded subgroup values. Under our SD-only calibration, the operative nominal target is $\alpha = 0.10$; both subgroups exceed the nominal target, consistent with the elevated domain FPR discussed in §6.1.

Operating point	Cov. (%)	Acc. (%)	Resid. SI%
Baseline (no abstention)	100.0	100.0	54.7
Flag-only abstention	45.3	100.0	22.1

Table 17: Selective-abstention demonstration (Llama-3.1-8B + BM25 + FinanceBench). Residual SI% is re-measured on the retained subset with a fresh calibration split. Coverage is $1 - \text{raw flagged rate}$ on the test split.

Setup. Flag-only abstention: on any query flagged by RAG-CF at $\alpha = 0.10$, return “I cannot determine the answer from the retrieved documents.”

Discussion. The residual SI% on the retained set is not zero even under flag-only abstention because the $\alpha = 0.10$ flag captures only the lower tail of η . The low coverage reflects the high SI rate on FinanceBench; on NQ-RAG (raw flagged rate 27.3%) the test-split coverage is 72.7%.

O Controlled-corpus replication

To isolate retrieval dilution from pretraining contamination on the high-rate FinanceBench configuration, we replicate the audit on a corpus of 200 documents constructed such that every gold answer is, by construction, absent from the public pre-

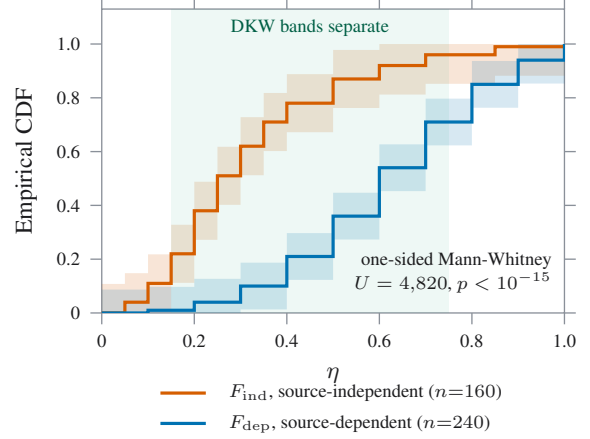


Figure 8: Pooled empirical step CDFs with 95% DKW bands. F_{dep} lies uniformly below F_{ind} ; bands separate across the shaded interior. One-sided Mann-Whitney $U = 4,820$, $p < 10^{-15}$.

training distribution. We curated post-knowledge-cutoff English news articles and scientific abstracts published between January and June 2024 (after Llama-3.1’s December 2023 training cutoff, verified against the Llama-3.1 release notes), and wrote 200 question-answer pairs in which the gold answer is a fact stated in the document. Every pair was spot-checked against a Llama-3.1-8B no-retrieval baseline: a question was retained only if the no-retrieval pass produced “I do not know” or a factually wrong answer, certifying that the gold information cannot be recovered parametrically. The remaining RAG-CF pipeline is unchanged ($M = 10$, $\alpha = 0.10$, 5 split seeds), with calibration scores transferred from the FinanceBench source-dependent partition (see the interpretation below). Table 18 reports the resulting flagged rates alongside the matched FinanceBench and NQ-RAG configurations.

Interpretation as an FPR calibration experiment. Because every query in the post-cutoff corpus is substantively source-dependent by construction ($\pi = 0$), the observed flag rate of 46.2% is the substantive false-positive rate of RAG-CF

Corpus	Raw SI%	FPR-adj. SI%	95% CI
Post-cutoff (this app.)	46.2	—	[42.1, 50.3]
FinanceBench (Tbl. 1)	54.7	49.7	[46.7, 52.7]
NQ-RAG (Tbl. 1)	27.3	19.2	[16.8, 21.7]

Table 18: Controlled-corpus replication (Llama-3.1-8B + BM25). Post-cutoff corpus: 200 post-Dec-2023 documents, 200 QA pairs; calibration transferred from FinanceBench. The FinanceBench and NQ-RAG intervals accompany the FPR-adjusted rates (the Table 1 intervals under the same transformation); the post-cutoff interval accompanies the raw rate. Since $\pi = 0$ by construction, the post-cutoff raw rate (46.2%) is the instrument’s FPR under cross-corpus calibration.

on this corpus, substantially above the nominal $\alpha = 0.10$. The calibration set $\mathcal{D}_{\text{cal}}$ for this experiment was drawn from FinanceBench source-dependent queries rather than from within the post-cutoff corpus, making this a cross-corpus transfer experiment; the elevated rate therefore reflects both retrieval dilution and the domain shift between FinanceBench and the post-cutoff corpus, and the Proposition 1 guarantee does not apply across this domain boundary. This excess is attributable to retrieval dilution: the T_{topic} operator’s substring filter on the gold answer string permits distractor leakage (the failure-mode taxonomy in App. I attributes 12% of flags to this mechanism), and within-domain priors allow the generator to produce plausible answers from question structure alone. The post-cutoff corpus therefore serves as a *diagnostic of the instrument’s FPR* on contamination-controlled post-cutoff content, not as a measurement of source-independence prevalence. The gap between the post-cutoff rate (46.2% raw) and the FinanceBench rate (54.7% raw) is ≈ 8.5 raw points; however, since both rates contain different false-positive contributions, the difference does not constitute a formal upper bound on contamination. The failure-mode taxonomy in App. I attributes a larger fraction (47% of flags, i.e., ≈ 25.7 raw points) to public-company knowledge leak; the two estimates are not directly comparable because they measure different quantities (residual rate difference vs. annotator-classified share of flags).

P Multilingual sensitivity

We run a single non-English configuration to begin probing cross-lingual transfer. **Setup.** Llama-3.1-8B-Instruct on MLQA-Spanish, 1,000 queries

(seed 0); BM25 with a Spanish stemmer; top- $k = 5$; $M = 10$; $\alpha = 0.10$; source-dependent calibration partition per §3.3, averaged over 5 sealed-seed splits. Topical-distractor clustering uses paraphrase-multilingual-MiniLM-L12-v2. Paraphraser is GPT-4o-mini with a Spanish system prompt; manual audit of 50 rewrites finds a 6% rate of numerical-detail loss (vs. 4% on English). Table 19 reports the resulting flagged rate alongside the matched NQ-RAG English configuration.

Config.	Raw SI%	FPR-adj. SI%	95% CI
MLQA-Spanish / Llama / BM25	30.2	22.4	[27.8, 32.6]
NQ-RAG / Llama / BM25 (ref.)	27.3	19.2	[25.1, 29.5]

Table 19: Preliminary multilingual configuration: MLQA-Spanish vs. matched English (NQ-RAG/Llama-3.1-8B/BM25). FPR-adjusted gap: 3.2 points; CIs partially overlap.

Reading. The MLQA-Spanish FPR-adjusted rate (22.4%) is within the band where we describe the result as broadly comparable to its closest English counterpart (NQ-RAG, FPR-adjusted 19.2%); the 3.2-point gap is on the order of the per-configuration cluster-bootstrap noise and the two intervals partially overlap. This is one configuration, on one language; we report it as a preliminary check of RAG-CF’s portability to a non-English benchmark, not as a multilingual study.