COUNTERMEM: Counterfactual Evaluation of Non-Parametric Memory Use in Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) is usually evaluated from one observed trace: the query, the retrieved context, and the generated answer. Such evaluations can tell us whether the answer is consistent with the retrieved context. They cannot tell us whether the system would still give the same answer if that context were removed, updated, or contradicted.

We introduce COUNTERMEM, an evaluation protocol for RAG systems that are expected to answer from retrieved evidence. COUNTERMEM keeps the RAG configuration fixed, changes only the retrieved context, regenerates the answer, and records whether the answer changes in the expected direction. This gives a behavioral test of evidence use: it measures whether the system follows retrieved evidence, not why the model does so internally.

Across RAGBench, ExpertQA, and DelucionQA, COUNTERMEM exposes failures that are hidden by observed-trace faithfulness: no-evidence persistence, update inertia, and conflict ignorance. Configurations that look equivalent under RAGAS faithfulness behave very differently once their evidence is removed, updated, or contradicted. In controlled gold/distractor fixtures, COUNTERMEM recovers the known retrieval-quality ranking; observed-trace faithfulness, which scores answer–context support rather than retrieval-quality, saturates on the same fixtures. Among RAGChecker-approved answers, more than half still silently miss coherent contradictions and over a third ignore coherent updates. Row-level intervention traces further help diagnose whether failures arise in retrieval, reranking, or generation. Anonymized artifact (protocol, prompts, detectors, scripts, audit traces): `https://anonymous.4open.science/r/CounterMem-F8B4/`.

1 Introduction

Retrieval-augmented generation (RAG) combines a parametric language model with non-parametric memory retrieved at inference time (Lewis et al., 2020; Gao et al., 2023b). Most RAG evaluations ask questions about the observed run: was relevant evidence retrieved, was the answer correct, and was the answer supported by the retrieved context? These are important questions, but they miss a different one: would the answer have changed if the retrieved evidence had changed?

This distinction matters for *evidence-bound* RAG systems—deployments that promise to (1) answer only when retrieved evidence is sufficient, (2) abstain when it is removed, (3) follow retrieved updates, (4) acknowledge unresolved conflicts, and (5) remain stable under irrelevant distractors. The corresponding behavioral failures that fixed-context scoring can hide are no-evidence persistence, update inertia, conflict ignorance, and noise sensitivity, each a violation of one or more of clauses (1)–(5); semantic redundancy is a measurement complication that can mask dependence under naive single-document deletion.

Observed-trace metrics are useful. Reference-free judges such as RAGAS (Es et al., 2023) and ARES (Saad-Falcon et al., 2024), and component diagnostics such as RAGChecker (Ru et al., 2024), score the realized (q, c, a) trace and can check whether the answer is compatible with the context that was actually retrieved. But two RAG systems can produce the same answer to the same query and retrieved context—one by using the passage, the other by ignoring it and answering from parametric memory—and receive identical observational scores; only an intervention on the retrieved state can tell them apart.

We introduce COUNTERMEM, a counterfactual evaluation protocol for RAG systems that are expected to answer from retrieved evidence. The protocol generalizes binary-mask context ablation to a typology of retrieved-state edits (deletions, counterfactual additions) applied at three explicit pipeline points: *post-retrieval*, *corpus-addition*, and *corpus-*

deletion. Figure 1 gives the protocol overview.

COUNTERMEM reports a profile of intervention signals, each tied to one clause of an evidence-bound criterion (Section 1). Empirically, configurations with matched RAGAS faithfulness still differ substantially in update following and conflict acknowledgment, exposing failure modes invisible to aggregate scoring (§5.1); the intervention trace localizes the failed pipeline stage—retrieval, reranking, or generator—and the matched single-stage repair outperforms wrong-stage repairs at no additional LLM-call cost beyond the behavioral test (§5.2); and behavioral failures persist among answers that aggregate diagnostics already approve, so compatibility with context is necessary but not sufficient for evidence-binding (§5.3).

2 Related Work

Observational and diagnostic RAG evaluation. Reference-free evaluators score the realized query–context–answer trace. RAGAS (Es et al., 2023) and ARES (Saad-Falcon et al., 2024) report faithfulness, relevance, and context-precision; RAG-Bench aggregates task-specific judge protocols (Friel et al., 2024); recent surveys study judge reliability and metric design (Liu et al., 2025; Ngo et al., 2024; Wang et al., 2024). Fine-grained diagnostic frameworks decompose answer quality further: RAGChecker (Ru et al., 2024) uses claim-level entailment for retriever and generator metrics, CoFE-RAG (Liu et al., 2024a) evaluates chunking and multi-stage diagnostics, RAGTrace (Cheng et al., 2025) analyzes retrieval–generation interactions, and RAGVUE (Murugaraj et al., 2026) decomposes reference-free evaluation with structured explanations. Metamorphic-testing approaches such as MetaRAG (Sok et al., 2025) probe hallucinations by perturbing factoids in the answer and rechecking entailment against the retrieved context.

Knowledge conflicts, unanswerability, and distractor robustness. A separate line of work studies how RAG systems handle natural conflicts and abstention: WikiContradict on contradictory Wikipedia passages (Hou et al., 2024), RAGTruth on conflict-driven hallucinations (Wu et al., 2023), and unanswerability benchmarks for cases without an admissible answer (Peng et al., 2025; Wallat et al., 2024). Distractor sensitivity appears in "lost in the middle" effects (Liu et al., 2024b), irrelevant-context perturbations (Shi et al., 2023), and cases where extra context unexpectedly helps (Cuconasu et al., 2024).

Grounded generation and attribution. Grounded-generation methods instruct the LLM to base its response on retrieved context with inline citations: ALCE (Gao et al., 2023a) and Self-RAG (Asai et al., 2024) provide such protocols, and self-correcting methods such as Self-Correcting RAG (Xu et al., 2026) use NLI-guided test-time search to improve grounding at generation time. Attribution benchmarks evaluate whether answers are supported by cited sources, including AttributedQA (Bohnet et al., 2023), ExpertQA (Malaviya et al., 2024), FactScore (Min et al., 2023), and the evaluation framework of Adlakha et al. (2024); faithfulness has been framed as an explanation that reflects what a model actually used (Jacovi and Goldberg, 2020).

Context-intervention probes and causal formalism. ContextCite formalizes context-attribution via binary mask ablations (Cohen-Wang et al., 2024), and entity-substitution and counter-memory studies probe knowledge conflicts via context edits (Longpre et al., 2021; Xie et al., 2024); redundant evidence further motivates cluster-level rather than single-document deletion (Chen et al., 2024; Saha Roy et al., 2025). Causal and interventional formalisms underpin the perturbation-and-effect discipline that COUNTERMEM adopts (Pearl, 2009; Feder et al., 2021). COUNTERMEM unifies single-mask deletion and counterfactual-addition interventions at three explicit pipeline points (post-retrieval, corpus-addition, corpus-deletion), reads a stage-localizable trace off the same intervention at no additional LLM-call cost, and complements rather than replaces aggregate diagnostics like RAGAS and RAGChecker.

3 Counterfactual Evaluation of Non-Parametric Memory Use

COUNTERMEM evaluates whether a fixed RAG configuration's outputs are behaviorally bound to its retrieved non-parametric memory. It intervenes on the retrieved state, regenerates with the same generator and prompt, and measures the behavioral transition under the evidence-bound criteria of Section 1.

3.1 Estimand and Detectors

A reported COUNTERMEM profile is always conditional on a query distribution, RAG configuration,

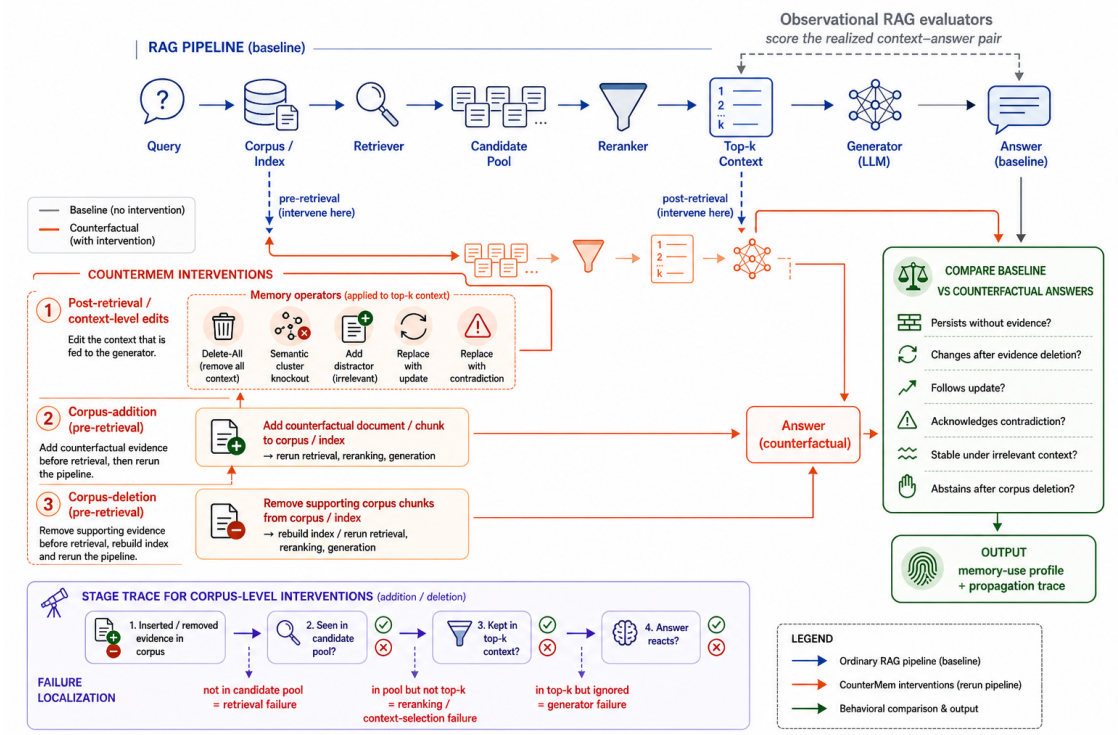

Figure 1: **CounterMem overview.** Five edit operators (Delete-All, semantic-cluster knockout, distractor injection, update injection, contradiction injection) act at the same *post-retrieval* point; corpus-addition and corpus-deletion apply earlier in the pipeline. Update and contradiction injections each append a synthesized counterfactual of the same factual type as the reference answer, framed once as superseding evidence and once as a competing source. Answers are regenerated under the same generator and prompt and compared to the baseline. The propagation trace localizes whether an intervention failed at retrieval, reranking, or generation.

prompting contract, intervention design, answer-similarity function, and behavioral detectors; under these conditions, it estimates behavioral-transition statistics under controlled retrieved-state interventions.

Let $R_c(q) = (d_1, \ldots, d_k)$ be the ordered context returned by configuration c for query q, and let $G_{\theta,c}(\cdot \mid q, K) \in \Delta(\mathcal{A})$ be the generator distribution over answers given retrieved state K under fixed parameters, prompt, and decoding. The base and intervened answer distributions are

$$A_0^c(q) = G_{\theta,c}(\cdot \mid q, R_c(q)), \quad (1)$$
$$A_I^c(q) = G_{\theta,c}(\cdot \mid q, I(R_c(q))), \quad (2)$$

where the intervention $I : \text{Context} \to \text{Context}$ generalizes the binary-mask operator of Cohen-Wang et al. (2024) to arbitrary deletions and counterfactual additions. Per-query signals below are expectations over draws $a \sim G_{\theta,c}$; with deterministic decoding, single API generations are plug-in estimates of the corresponding means.

We compare answers with an operational similarity function $\mathrm{Sim}(a, a') \in [0, 1]$ and read three behavioral labels off the answer text:

$$\mathrm{Abstain}(a) \in \{0, 1\}, \quad (3)$$
$$\mathrm{ConflictAck}(a) \in \{0, 1\}, \quad (4)$$
$$\mathrm{UpdateFollow}(a; u) \in \{0, 1\}, \quad (5)$$

indicating an insufficient-evidence response, an explicit acknowledgment of conflict between sources, and adoption of an injected update u. Each label is produced by a Claude-Sonnet judge classifier over a fixed rubric shipped with the released harness; the abstain label additionally cross-checks an explicit-marker regex against the judge to lower-bound criteria-level abstention. The similarity function, thresholds, and detector rubrics are protocol choices; we audit them and report sensitivity to similarity functions, clustering thresholds, and prompt style in Appendix G.

3.2 Interventions and Signals

Each intervention maps the retrieved state to a counterfactual state and answers a different diagnostic question. **Delete-All**, the *closed-book* intervention familiar from prior RAG analysis (Liu et al., 2024b;

Cuconasu et al., 2024), removes all retrieved evidence, $I_\emptyset(R_c(q)) = \emptyset$, and asks whether the answer survives without context:

$$\mathrm{DAP}(q) = \mathop{\mathbb{E}}_{\substack{a_0 \sim A_0^c(q) \\ a_\emptyset \sim A_{I_\emptyset}^c(q)}} [\mathrm{Sim}(a_0, a_\emptyset)]\,. \tag{6}$$

We distinguish persistence DAP, behavioral dependence $1-\mathrm{DAP}$, abstention after deletion $\mathrm{DelAll} = \mathbb{1}[\mathrm{Abstain}(a_\emptyset)]$, and the substantive violation flag

$$\mathrm{DAV}(q) = \mathbb{1}\big[\,\mathrm{DAP}(q) \geq \tau \;\wedge\; \neg\,\mathrm{Abstain}(a_0) \;\wedge\; \neg\,\mathrm{Abstain}(a_\emptyset)\,\big], \tag{7}$$

which fires only when both answers are substantive yet similar. We use $\tau = 0.5$ throughout; high DAP alone is not a failure outside evidence-bound criteria.

Cluster knockout. Targeted deletion removes one retrieved item or semantic evidence unit S while leaving the rest of the context intact. In redundant corpora, single-document deletion can miss dependence because equivalent passages remain. We therefore cluster $R_c(q)$ into evidence units $\mathcal{C}(q) = \{C_1, \ldots, C_m\}$ via complete-linkage HAC and delete one whole cluster at a time. The cluster-knockout signal is

$$\mathrm{CNS}(q) = \max_i \mathbb{E}\left[1 - \mathrm{Sim}(a_0, a_{\setminus C_i})\right], \tag{8}$$

the largest answer change induced by removing any single evidence unit. High CNS indicates that the answer depends on a specific cluster; under evidence-bound criteria this dependence is desirable, while under open-domain use the same value can be benign. Appendix G compares alternative clustering rules (Saha Roy et al., 2025).

Counterfactual additions. Each counterfactual addition appends one synthesized document to the retrieved context, leaving the existing documents unchanged. *Distractor injection* adds an irrelevant chunk drawn from another query in the same dataset and reports noise robustness $\mathrm{NR} = \mathbb{E}[\mathrm{Sim}(a_0, a_{+n})]$. *Update injection* and *contradiction injection* each append a 2–3 sentence passage that asserts a synthesized counterfactual answer of the same factual type as the reference: update injection frames the counterfactual as superseding new information ("Recent sources have established that . . . "), while contradiction injection frames it as a competing claim ("one source claims u, not the reference"). The two probes test complementary clauses of the same retrieved-state perturbation: $\mathrm{Upd} = \mathbb{1}[\mathrm{UpdateFollow}(a_u; u)]$ measures adoption of the appended fact, while $\mathrm{Contra} = \mathbb{1}[\mathrm{ConflictAck}(a_{\mathrm{con}})]$ measures explicit acknowledgment of the conflict; we also report contradiction stability $\mathrm{CS} = \mathrm{Sim}(a_0, a_{\mathrm{con}})$. The probes are related to the entity-substitution and counter-memory protocols of Longpre et al. (2021) and Xie et al. (2024), but differ in operating by addition rather than in-place substitution: the original retrieved context is preserved, and the counterfactual enters alongside it.

For each query, COUNTERMEM returns the profile

$$\rho(q) = (\,\mathrm{DAP}, \mathrm{DAV}, \mathrm{DelAll}, \mathrm{CNS}, \mathrm{NR}, \mathrm{Upd}, \mathrm{Contra}, \mathrm{CS}\,). \tag{9}$$

Aggregating over queries gives a configuration-level profile with paired-bootstrap confidence intervals. The components are not collapsed into a single score: stability is desirable under noise, but undesirable under deletion or relevant updates.

Appendix A provides a per-signal reading guide (intervention, operational question, desired direction under each evidence-bound clause, per-query LLM-call cost, and main pitfall). In what follows, ↑ marks "higher is desirable under the criteria," ↓ "lower is desirable;" criteria-dependent signals (e.g., CS under contradictions) are reported with their companion signal and no arrow.

3.3 Intervention Modes and Pipeline-Stage Trace

The same intervention can be applied at three observable points. In *post-retrieval* mode, we modify only the context supplied to the generator. In *corpus-addition* mode, the intervention document is added to the indexable corpus and ranked by the retriever's own scoring function. In *corpus-deletion* mode, supporting chunks are removed from the corpus and retrieval is rerun, which tests end-to-end evidence-bound abstention. Prior context-modification work (Cohen-Wang et al., 2024; Liu et al., 2024b; Cuconasu et al., 2024) operates at the post-retrieval point only; we make the three intervention points explicit because they probe different stages of the pipeline and yield distinct propagation traces.

For each intervention and configuration, the harness logs the pre- and post-intervention retrieved

Dataset	Setting	Primary use	Size
RAGBench-HotpotQA	Passage RAG	Judge comparison; metric disagreement	424 / 150 per sweep cell
RAGBench-TechQA	Passage RAG	Metric disagreement; redundancy	304 / 150 per sweep cell
RAGBench-PubMedQA	Passage RAG	Metric disagreement; redundancy	500 / 150 per sweep cell
ExpertQA	Attributed QA	Human-support validation	500 full / 200 GPT / 187 Claude
DelucionQA conflict subset	Passage RAG	Natural conflicts	100

Table 1: Datasets and roles in the evaluation.

state, first-stage candidate pool, reranked top-k context, generated answer, and behavioral label. This trace localizes observable propagation failures: absent from the candidate pool implies retrieval failure; present in the pool but absent from top-k implies reranking or context-selection failure; present in top-k but behaviorally ignored implies generator failure. Sections 5.2–5.2 test whether this localized stage predicts the effective repair. The claim is operational and pipeline-level, not a structural causal model of the generator.

4 Experimental Setup

Datasets. We evaluate on RAGBench's HotpotQA, TechQA, and PubMedQA splits (Friel et al., 2024), ExpertQA (Malaviya et al., 2024), and the DelucionQA conflict subset (Sadat et al., 2023); per-dataset roles and sizes appear in Table 1. We run two settings: *benchmark-conditioned* uses dataset-provided contexts and intervenes post-retrieval, isolating generator-side memory use; *end-to-end* builds sparse BM25, dense, and hybrid retrievers with $k \in \{2, 4, 8\}$, so ranking disagreements between metrics are attributable to the metric rather than the system.

Generator, judge, and prompt. The generator under test is CLAUDE-4.5-HAIKU for all headline experiments (RAGBench-27 sweep, coherent-intervention probe, ExpertQA, end-to-end retrieval suite). Behavioral classification (abstain, conflict acknowledgment, silent pick, intervention validity) uses CLAUDE-SONNET-4.6 as a stronger LLM judge. Decoding is deterministic where supported; for API backends we use one generation per condition and quantify sampling sensitivity in Appendix G. All runs share a grounded QA prompt: answer from the provided context, abstain when evidence is insufficient, mention conflicts when sources disagree. Interventions leave the instruction unchanged and modify only the retrieved context.

Intervention parameters. All interventions from Section 3 are instantiated with a single fixed parameterization. Evidence units are formed by complete-linkage HAC on SBERT all-MiniLM-L6-v2 embeddings at cosine threshold 0.6; appendix ablations compare alternative clustering rules (Saha Roy et al., 2025). Answer similarity uses the same SBERT bi-encoder (SAS), with cross-encoder and Jaccard variants reported as robustness checks. Cluster knockout removes one whole cluster at a time and aggregates by $\max_i$ (Eq. 8); the DAV threshold is $\tau = 0.5$ (Section 3). Distractor injection adds one irrelevant chunk drawn from another query in the same dataset. Update and contradiction payloads are synthesized by Claude-Haiku conditioned on the question and the reference answer (an alternative answer of the same factual type) and audited per-row by Claude-Sonnet (Appendix G); only judged-valid contradictions enter the headline coherent-intervention rates.

Baselines. Observational baselines: RAGBench judge scores (Friel et al., 2024), RAGAS faithfulness and context precision (Es et al., 2023), RAGChecker on the stage-diagnosis fixture (Ru et al., 2024), reference-based answer quality where applicable, and Recall@k.

Validation, audit, and cost. ExpertQA human-support labels provide a human-grounded reliability check (Appendix D); we report ROC-AUC, PR-AUC, Spearman correlation, and 95% paired-bootstrap CIs over per-query pairs. A stratified audit covers Delete-All outputs, cluster knockouts, updates, and contradictions for intervention validity and detector reliability; sensitivity to clustering threshold, similarity function, prompt style, and sampling is reported in Appendix G. COUNTERMEM adds $m + 4$ extra generations per query for m clusters; we use a staged policy (Delete-All triage $\rightarrow$ cluster knockouts on flagged queries $\rightarrow$ update/contradiction probes where contractually required), with the per-condition cost numbers reported in Appendix D.

Data, licenses, and intended use. All datasets used (RAGBench, ExpertQA, DelucionQA) are publicly released benchmarks for RAG evaluation research, and our use is consistent with their stated evaluation purpose. We collect no new data; the corpora contain no personally identifying information beyond what is already public in their source documents. The released COUNTERMEM harness ships under the license specified in the artifact repository.

5 Results

We organize the results around three questions: (1) does each COUNTERMEM signal flag failure modes that observational baselines miss (§5.1)? (2) does the same intervention trace localize the pipeline stage to repair (§5.2)? and (3) what behavioral failures does COUNTERMEM reveal among answers that aggregate diagnostics already approve (§5.3)? Construct-validity, end-to-end corpus deletion, distribution-shift transfer (Appendix B), and robustness checks are deferred to the appendix. The RAG configurations throughout are evaluation fixtures, not proposed systems.

5.1 Failure-mode Coverage on RAGBench-27

COUNTERMEM flags update-inertia and contradiction-ignorance failures that observational baselines rarely co-flag. We sweep 27 RAGBench configurations (3 datasets $\times$ 3 retrievers $\times$ $k \in \{2, 4, 8\}$, n=150 queries per cell) and ask, for each evidence-bound failure mode, how many configurations COUNTERMEM flags as failing and how many of those are also flagged by RAGAS-faithfulness or RAGChecker hallucination at standard thresholds (Table 2).

No-evidence non-abstention is partially visible to observational signals. Of the 18 configurations that fail to abstain reliably under Delete-All, RAGAS-faithfulness co-flags 6 (33%) and RAGChecker hallucination co-flags 9 (50%) on the realized base answer. Configurations that rarely abstain when context is removed tend also to have lower realized-trace support, so faithfulness and hallucination scores partially correlate with abstention behavior, but neither signal sees the counterfactual directly. The persistence-substantive flag DAV reported in the protocol gives a stricter view tied to the answer text rather than to abstention markers.

Update inertia and contradiction ignorance are rarely co-flagged by observational baselines. Of the 7 configurations that fail update propagation and the 12 that fail conflict acknowledgment, RAGAS-faithfulness co-flags 0 on either axis and RAGChecker hallucination co-flags 0 and 1 respectively. This is structural: faithfulness and hallucination score the realized (q, c, a) trace, but update-following and conflict-acknowledgment are properties of the answer under counterfactual evidence. The clearest demonstration of the gap: among the 31 same-dataset configuration pairs whose RAGAS-faithfulness scores match within ± 0.01, the maximum within-pair behavioral difference is 0.31 in update following and 0.29 in conflict acknowledgment (median 0.11 and 0.12).

5.2 Pipeline Diagnosis at Constant LLM Cost

The intervention trace localizes the dominant failure stage, and the matched single-stage repair outperforms wrong-stage repairs—at no additional LLM-call cost beyond the behavioral test.

Trace localization names the engineered bottleneck. On the three engineered fixtures, the dominant trace class matches the construction: A's trace is 66% retrieval, B's is 96% rerank, C's is 53% generator (Table 3). In the natural HotpotQA fixture, generator-stage failures dominate the unrepaired residual (46%), with retrieval and rerank failures rare (3% and 4%).

Targeted repair beats wrong-stage repair with non-overlapping confidence intervals on three of four fixtures. A's retriever swap reaches 0.53 reactive vs. $\leq$ 0.12 for off-stage repairs (Δ=+0.40 [+0.29, +0.51]); B's reranker swap reaches 0.56 vs. 0.00 for the off-stage retriever swap (Δ=+0.54 [+0.45, +0.63]); on natural HotpotQA generator-class queries, prompt repair reaches 0.44 vs. 0.06 for naive pool/k expansion (Δ=+0.37 [+0.22, +0.52]). Row C is suggestive: prompt repair reaches 0.63 vs. 0.47 for the next best, but the paired delta CI includes zero (Δ=+0.09 [−0.02, +0.19]).

COUNTERMEM's stage attribution is read off the same intervention trace already used to compute Contra. RAGChecker is a strong baseline at the metric level (distinguishing A from B on 10/11 metrics), but its retriever component combines retrieval and reranking, so achieving the same stage decomposition would require rerunning RAGChecker at each pipeline stage's intermediate output and multiplying its LLM-call cost. The full repair matrix and the per-metric RAGChecker breakdown are in Appendix C.

5.3 Coherent-intervention Failure Modes Invisible to RAGChecker

Behavioral failures are common among RAGChecker-approved answers: aggregate faithfulness is necessary but not sufficient for evidence-binding.

Setup. For each of six RAGBench datasets (HotpotQA, TechQA, PubMedQA, CovidQA, EManual,

Table 2: **Detection of behavioral failure modes on RAGBench-27.** Counts of 27 configurations failing each COUNTERMEM threshold, and how many of those are also flagged by RAGAS faithfulness < 0.9 or RAGChecker hallucination > 0.15. Independent baseline flag rates over all 27 configurations: 22% (faith), 33% (hall). Update inertia and contradiction ignorance are counterfactual properties; realized-trace baselines are not designed to detect them.

Failure mode (CounterMem threshold)	Failing configs	Co-flagged by RAGAS-faith	Co-flagged by RAGChecker hall
No-evidence non-abstention (DelAll abst. < 0.5)	18/27	6/18 (33%)	9/18 (50%)
Update inertia ($\text{Upd}_{\text{changed}} < 0.5$)	7/27	**0/7** (**0**%)	**0/7** (**0**%)
Contradiction ignorance (Contra ack. < 0.3)	12/27	**0/12** (**0**%)	1/12 (8%)

Table 3: **Stage diagnosis and targeted repair.** Trace localization (%), post-repair reactive rate, and targeted Δ with 95% paired-bootstrap CIs ($*$: CI excludes zero). Bold marks the trace-dominant stage and matching diagonal repair. Natural-fixture rates and Δ are on the generator-class subset (n=46) vs naive pool/k expansion; A/B/C use all n=100 vs the best wrong-stage repair.

		Trace localization (%)				Repair reactive rate			
Fixture	n	Retr	Rerank	Gen	React	Retr swap	Rerank swap	Prompt	Targeted Δ [95% CI]
Natural HotpotQA	100	3	4	46	47	0.17	0.33	**0.44**	$+0.37^*$ $[+0.22, +0.52]$
A: retrieval-broken	100	**66**	1	24	9	**0.53**	0.08	0.12	$+0.40^*$ $[+0.29, +0.51]$
B: rerank-broken	100	2	**96**	0	2	0.00	**0.56**	0.02	$+0.54^*$ $[+0.45, +0.63]$
C: generator-broken	100	2	9	**53**	36	0.45	0.47	**0.63**	$+0.09$ $[-0.02, +0.19]$

Table 4: **Behavioral failures among RAGChecker-approved answers** (rch_hall ≤ 0.10) **with judged-valid contradiction payloads.** Share of approved answers that ignore a coherent update, miss a coherent contradiction, or fail both. Total-row brackets: Wilson 95% CIs (n=446); per-dataset cells (especially TechQA, n=31) carry wider intervals.

Dataset	n_{good}	Upd ignored	Contra fail	Both fail
TechQA	31	71.0%	80.6%	67.7%
EManual	90	62.2%	80.0%	52.2%
FinQA	84	58.3%	45.2%	26.2%
CovidQA	72	29.2%	50.0%	20.8%
HotpotQA	86	23.3%	41.9%	14.0%
PubMedQA	83	6.0%	49.4%	0.0%
Total	446	38.8% [34.4, 43.4]	55.6% [51.0, 60.1]	26.2% [22.4, 30.5]

FinQA), we synthesize *coherent* update and contradiction documents using Claude Haiku conditioned on the original retrieved context, then run the standard COUNTERMEM update- and contradiction-addition interventions. A per-row audit by Claude-Sonnet finds 93.6% of contradiction payloads well-formed; we restrict the headline subset to RAGChecker-approved answers (rch_hall ≤ 0.10) with judged-valid contradictions (n=446, out of n=593 generated). Each baseline answer is scored by RAGChecker; we measure update following on the update intervention and conflict acknowledgment on the contradiction intervention.

One in three approved answers ignores a coherent update; more than half fail a coherent contradiction; one in four fails both. Aggregated across the six datasets, 38.8% [34.4, 43.4] silently ignore the coherent update, 55.6% [51.0, 60.1] fail to acknowledge the coherent contradiction, and 26.2% [22.4, 30.5] exhibit both failures simultaneously (Wilson 95% CIs over the judged-valid subset; Table 4). RAGChecker scored every one of these answers as faithful to the original retrieved context.

The failure rate varies sharply by domain. TechQA models exhibit both failures on the majority of approved answers (67.7%, n=31); EManual is similar (52.2%); PubMedQA approved answers follow updates well (6.0% ignored) and never fail both (0.0%) but still miss conflicts almost half the time (49.4%). The cross-domain gradient indicates that compatibility-with-context (RAGChecker's signal) and behavioral following-of-context (COUNTERMEM's signal) are correlated but not equivalent.

Worked example. On the HotpotQA query *"What city is the women's Vans Triple Crown of Surfing event hosted at?"* the baseline answers *Honolulu* and RAGChecker reports rch_hall $= 0.000$. After appending a coherent update specifying *Sunset Beach*, the model continues to answer *Honolulu* (update ignored). After appending a coherent contradiction stating *Sunset Beach*, the model still answers *Honolulu* with no conflict marker (silent

confident match). The aggregate diagnostic has no signal here; the intervention does.

The pattern reproduces on natural conflicts and on a human-grounded endpoint. On 100 DelucionQA naturally-conflicting passages, silent-pick failures rise from 4% at baseline to 24% under one-sided context removal (Appendix E). On the ExpertQA human-support endpoint, adding the COUNTERMEM profile to RAGAS lifts ROC-AUC under both judges (+0.150 under a weak GPT-4o-mini judge, +0.066 under a strong Claude-Sonnet judge; both $p<0.001$, Table 9). The lift is carried primarily by $1 - \mathrm{DAP}$; CNS independently correlates with the human label (Spearman $\rho \approx 0.2$) but adds little incremental signal on top of $1 - \mathrm{DAP}$ in the ensemble, and NR is uncorrelated with the human-support endpoint as expected, since clause (5) and clauses (1)–(2) target different evidence-bound criteria (Appendix D).

End-to-end corpus deletion confirms the evidence boundary, and exposes a residual-corpus regime. Removing gold-supporting chunks from the index and rerunning retrieval moves abstention from baseline to 0.99 on HotpotQA and 1.00 on PubMedQA: the system follows the evidence boundary end-to-end. On TechQA, abstention only moves $0.51 \rightarrow 0.57$ because large technical-manual gold passages leave many related residual chunks in the corpus; the probe makes that corpus property operationally testable (Appendix F).

6 Conclusion

For RAG systems that promise to answer from retrieved evidence, observational diagnostics ask the wrong question. Whether an answer is consistent with the realized context tells us nothing about whether the system would still produce that answer if the evidence were removed, updated, or contradicted. COUNTERMEM operationalizes the counterfactual: it intervenes on the retrieved state, regenerates with the same generator and prompt, and measures the behavioral transition.

The protocol exposes failures invisible to aggregate scoring. On RAGBench's 27-configuration sweep, same-dataset matched-faithfulness pairs differ by up to 0.31 in update following and 0.29 in conflict acknowledgment (n=31 matched-faith pairs). On six datasets, 38.8% [34.4, 43.4] of RAGChecker-approved answers silently ignore a coherent retrieved update and 55.6% [51.0, 60.1] miss a coherent contradiction (Wilson 95% CIs over the judged-valid subset, n=446). The same intervention trace localizes the pipeline stage to repair: matched repairs produce large gains on retrieval- and rerank-broken fixtures (+0.40 and +0.54 reactive rate), a significant gain on the natural HotpotQA generator class (+0.37 vs. naive pool/k expansion), and a smaller non-significant gain on the engineered generator fixture. This lets practitioners select, debug, and audit RAG configurations against requirements that aggregate evaluators cannot enforce alone.

Limitations

The protocol depends on the chosen interventions and detectors. Update and contradiction payloads are LLM-synthesized rather than human-curated; a per-row Claude-Sonnet audit of all 593 coherent contradictions finds 93.6% well-formed, and a separate stratified audit on the larger sweep (n=50 per class; Appendix G) finds Delete-All and update payloads 100% and 96% well-formed. Headline behavioral rates are computed on the judged-valid contradiction subset; behavioral classification uses the LLM judge throughout. Answer-similarity detectors can miss small but consequential factual edits, and cluster-deletion granularity defines the unit of evidence removal. Sample sizes vary by cell: pooled headline rates are computed over 446 RAGChecker-approved answers with valid contradictions, but per-dataset cells (especially TechQA, n=31) carry wider intervals. COUNTERMEM is behavioral and does not identify the generator's internal causal mechanism: the protocol observes whether outputs change under counterfactual retrieved states, not which internal computation produced them. Finally, COUNTERMEM reports a *profile* rather than a single score: high DAP is a leakage risk only for systems that contractually promise to answer from retrieved evidence, and DAV is the stricter flag when the concern is unsupported substantive answering rather than benign stability.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-following models for question answering. *TACL*, 12.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learn-

ing to retrieve, generate, and critique through self-reflection.

Bernd Bohnet, Vinh Q. Phu, Sandra Meister, Siva Reddy, Seid Muhie Yimam, Sebastian Riedel, and Omer Levy. 2023. Attributed question answering: Evaluation and modeling for attributed large language models.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Sizhe Cheng, Jiaping Li, Huanchen Wang, and Yuxin Ma. 2025. Ragtrace: Understanding and refining retrieval-generation dynamics in retrieval-augmented generation. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, pages 1–20.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*.

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.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. RAGAS: Automated evaluation of retrieval augmented generation.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *TACL*.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. RAGBench: Explainable benchmark for retrieval-augmented generation systems.

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

Yunfan Gao, Yunqing Gao, Wenhui He, Kaixuan Chen, Muhao Wang, Aohan Zhou, and Zhiyuan Li. 2023b. Retrieval-augmented generation for large language models: A survey.

Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. WikiContradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from wikipedia.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4198–4208.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Jintao Liu, Ruixue Ding, Linhao Zhang, Pengjun Xie, and Fie Huang. 2024a. Cofe-rag: A comprehensive full-chain evaluation framework for retrieval-augmented generation with enhanced data diversity. *arXiv preprint arXiv:2410.12248*.

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

Shuliang Liu, Xinze Li, Zhenghao Liu, Yukun Yan, Cheng Yang, Zheni Zeng, Zhiyuan Liu, Maosong Sun, and Ge Yu. 2025. Judge as a judge: Improving the evaluation of retrieval-augmented generation through the judge-consistency of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5788–5807.

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)*, pages 7052–7063.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Kathleen McKeown, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers.

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.

Keerthana Murugaraj, Salima Lamsiyah, and Martin Theobald. 2026. RAGVUE: A diagnostic view for explainable and automated evaluation of retrieval-augmented generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 512–526.

Nghia Trung Ngo, Chien Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Comprehensive and practical evaluation of retrieval-augmented generation systems for medical question answering. *arXiv preprint arXiv:2411.09213*.

Judea Pearl. 2009. Causality: Models, reasoning and inference. In *Cambridge University Press*.

Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. 2025. Unanswerability evaluation for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8452–8472.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, and 1 others. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027.

Jon Saad-Falcon, Rohan Jha, Hamish Ivison, and Nick Craswell. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R. Menon, Md Rizwan Parvez, and Zhe Feng. 2023. DelucionQA: Detecting hallucinations in domain-specific question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 822–835.

Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech. 2025. Evidence contextualization and counterfactual attribution for conversational qa over heterogeneous data with rag systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 1040–1043.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context.

Channdeth Sok, David Luz, and Yacine Haddam. 2025. MetaRAG: Metamorphic testing for hallucination detection in RAG systems.

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

Yang Wang, Alberto Garcia Hernandez, Roman Kyslyi, and Nicholas Kersting. 2024. Evaluating quality of answers for retrieval-augmented generation: A strong llm is all you need. *arXiv preprint arXiv:2406.18064*.

Yuqing Wu, Nan Hu, Sheng Bi, Guilin Qiu, Huiwen Cao, and Tong Zhao. 2023. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of LLMs in knowledge conflicts. In *International Conference on Learning Representations (ICLR)*.

Shijia Xu, Zhou Wu, Xiaolong Jia, Yu Wang, Kai Liu, and April Xiaowen Dong. 2026. Self-correcting RAG: Enhancing faithfulness via MMKP context selection and NLI-guided MCTS.

A Signal Card

Table 5 is the per-signal reading guide referenced in Section 3, with the desired direction under each evidence-bound clause (Section 1).

Table 5: **COUNTERMEM signal card.** Signals grouped by intervention, with the operational question and main pitfall under the evidence-bound criteria (Section 1). Group header reports marginal LLM-call cost.

Signal	Operational question	Main pitfall
Delete-All *+1 call*		
DAP ↓	Does the answer survive without retrieved evidence?	Persistence can be valid zero-context knowledge; it is a leakage risk only under evidence-bound criteria.
DelAll ↑	Does the system abstain when evidence is removed?	Counts explicit abstention markers, so it lower-bounds criteria-level abstention.
DAV ↓	Is the no-evidence answer both substantive and similar?	Use as the leakage flag instead of raw DAP $\geq \tau$.
Cluster knockout *+m calls*		
CNS ↑	Does removing any single semantic evidence unit change the answer?	Clustering rule matters; low CNS is the criteria-relevant warning (evidence-independent answer or redundant clusters), high CNS is desirable under evidence-bound use and benign elsewhere.
Counterfactual addition *+1 call per probe*		
NR ↑ (distractor)	Is the answer stable under an irrelevant added chunk?	Stability is desirable only for irrelevant-context additions.
Upd ↑ (update)	Does the answer adopt a counterfactual appended as superseding evidence?	Fixture detectors may miss out-of-distribution updates.
Contra ↑ (contradiction)	Does the answer acknowledge conflict when the same counterfactual is framed as a competing source?	Use an LLM judge; keyword matchers undercount criteria-relevant conflict handling.
CS (contradiction)	How similar is the answer to the pre-injection baseline?	Direction is criteria-dependent; report alongside Contra.

B Distribution-shift Transfer of Metric Rankings

We test whether per-configuration metric rankings on dataset D_1 predict reference-based answer quality on dataset D_2, over the six ordered pairs in the RAGBench 3-dataset subset (HotpotQA, PubMedQA, TechQA). Interventional signals transfer in the expected direction on every pair; observational faithfulness and context-precision flip sign on half (Table 6; faithfulness saturation analyzed in Appendix G).

C Pipeline-Diagnosis Details

Table 7 reports the full 3×4 baseline$\times$repair reactive-rate matrix on the engineered HotpotQA bottlenecks summarized in Section 5.2.

Table 8 gives RAGChecker's full 11-metric breakdown across the same three engineered baselines.

D Multi-metric ExpertQA Validation

We fit logistic-regression ensembles over COUNTERMEM features (1−DAP, CNS, NR) and RAGAS faithfulness on the ExpertQA human-support endpoint (Table 9), supporting the framing in Section 5.3.

E Naturalistic Conflict Reproduction on DelucionQA

We replicate the conflict-handling probe of Section 5.3 on naturally occurring DelucionQA contradictions to check that the failure mode is not a synthetic-template artifact (Table 10).

F End-to-End Corpus Deletion Details

Table 11 reports the per-dataset corpus-deletion numbers behind the end-to-end paragraph in Section 5.3.

G Sanity Checks: Construct Validity and Robustness

These checks address the sanity-check pointer in Section 5 and the validity audit referenced in Section 5.3: we (i) verify that each retrieval-quality-targeted metric correctly tracks a known retrieval-quality gradient on a synthetic gold/distractor fixture, and (ii) vary the answer-similarity function, the cluster-deletion threshold, the prompt style, and the generation-sampling regime, and report the per-class intervention-validity numbers.

Construct validity on mock gold/distractor systems. For each query, we build a system g that receives g gold passages and $K-g$ distractors, and

Table 6: Cross-distribution transfer of metric rankings on the 27-configuration RAGBench sweep. Each cell is Spearman ρ between the per-configuration ranking under the column metric measured on dataset D_1 and the per-configuration ranking by reference-based answer quality (BaseQuality) on dataset D_2. The bottom row averages the six $D_1 \to D_2$ pairs. Interventional metrics (1−DAP, NR, CS) transfer in the expected direction on every pair; observational metrics (RAGAS-faith, RAGAS-cp) flip sign on half the pairs.

$D_1 \to D_2$	1−DAP	Recall@k	NR	CS	RAGAS-faith	RAGAS-cp
HotpotQA → PubMedQA	+0.81	+0.58	+0.61	+0.55	+0.64	−0.50
HotpotQA → TechQA	+0.69	+0.41	+0.40	+0.40	+0.38	−0.21
PubMedQA → HotpotQA	+0.56	+0.62	+0.25	+0.51	+0.15	−0.10
PubMedQA → TechQA	+0.70	+0.74	+0.17	+0.65	−0.17	+0.00
TechQA → HotpotQA	+0.46	+0.45	+0.62	+0.62	−0.25	−0.34
TechQA → PubMedQA	+0.85	+0.47	+0.53	+0.52	−0.31	−0.36
mean	**+0.68**	+0.55	+0.42	+0.54	**+0.07**	−0.25

Table 7: **Reactive-rate matrix: 3 broken baselines × 4 conditions on HotpotQA,** $n = 100$**.** Cells show reactive rate (acknowledge or abstain) with 95% bootstrap CI. **Bold** marks the targeted repair for each row. Right column reports paired bootstrap delta (targeted minus best wrong-stage repair).

Baseline	none (broken)	R_{ret}	R_{rer}	R_{gen}	Δ targeted − best wrong
A: retrieval-broken	0.07 [0.02, 0.12]	0.53 [0.43, 0.63]	0.08 [0.03, 0.14]	0.12 [0.06, 0.19]	+0.40 [+0.29, +0.51]
B: rerank-broken	0.02 [0.00, 0.05]	0.00 [0.00, 0.00]	0.56 [0.47, 0.65]	0.02 [0.00, 0.05]	+0.54 [+0.45, +0.63]
C: generator-broken	0.34 [0.24, 0.44]	0.45 [0.36, 0.55]	0.47 [0.37, 0.56]	0.63 [0.53, 0.72]	+0.09 [−0.02, +0.19]

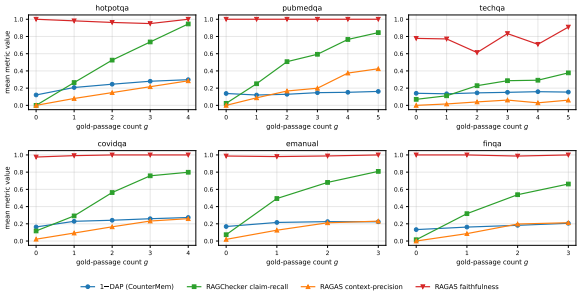

Figure 2: **Construct validity on mock gold/distractor systems.** Mean metric value as a function of gold-passage count g for six datasets. RAGAS faithfulness saturates near 1.0 regardless of g. 1−DAP (CounterMem), RAGChecker claim-recall, and RAGAS context-precision all rise monotonically with g.

ask whether each retrieval-quality-targeted metric rises monotonically in g. 1−DAP (mean Kendall $\tau = +0.86$), RAGChecker claim-recall, and RAGAS context-precision rise with g; RAGAS faithfulness saturates near 1.0 because a configuration can write a faithful answer to an all-distractor context (Figure 2).

Intervention validity by class. Table 12 reports the stratified audit on the larger sweep (Claude-Sonnet, n=50 per class). For the coherent-intervention probe of Section 5.3, a separate per-row audit of all 593 contradictions finds 93.6% well-formed; the headline rates in Table 4 are computed on the judged-valid subset.

Answer-similarity function. Pairwise Spearman correlation between DAP estimates under three answer-similarity functions on the ExpertQA n=200 subsample (Table 13).

Cluster-deletion threshold. Fraction of queries with at least one multi-document cluster under semantic SBERT clustering at four thresholds. Conclusions about redundancy ordering are stable around τ=0.6 (the main-text default).

Prompt style. Same RAGBench rows under *strict* (main-paper default) and *helpful* prompts (Table 15); stricter prompts are a partial—not complete—fix for the failure modes COUNTERMEM measures.

Repeated generation under sampling. We sampled n=49 ExpertQA queries and called Claude-Haiku with temperature 0.7 three independent times per condition (Table 16); aggregate DAP standard deviation across reps is 0.003, supporting single-generation estimates for the main results.

Table 8: **RAGChecker metrics on the three broken baselines,** $n = 100$**.** RAGChecker (Amazon Research, 2024) provides component-level diagnostic metrics: retriever (claim_recall, context_precision) and generator (faithfulness, hallucination, etc.). Bottom row reports paired bootstrap delta $\mu_A - \mu_B$; $\circ$ marks CIs containing zero. Most metrics distinguish A from B at the metric level, but RAGChecker does not separate retrieval from rerank within the retriever component—obtaining stage attribution would require running RAGChecker at each pipeline stage's output, multiplying its LLM-call cost.

Baseline	precision	recall	F1	claim_recall	context_precision	ctx_utilization	noise_sens. (rel)	noise_sens. (irrel)	hallucination	self_knowledge	faithfulness
A	0.24 [0.18, 0.30]	0.27 [0.20, 0.35]	0.19 [0.14, 0.25]	0.30 [0.22, 0.38]	0.15 [0.11, 0.20]	0.34 [0.26, 0.44]	0.16 [0.11, 0.22]	0.33 [0.25, 0.41]	0.26 [0.20, 0.33]	0.06 [0.03, 0.09]	0.67 [0.59, 0.75]
B	0.16 [0.11, 0.21]	0.09 [0.04, 0.14]	0.07 [0.03, 0.11]	0.08 [0.04, 0.12]	0.04 [0.02, 0.06]	0.11 [0.05, 0.17]	0.02 [0.01, 0.04]	0.41 [0.33, 0.49]	0.42 [0.34, 0.49]	0.10 [0.06, 0.14]	0.48 [0.41, 0.56]
C	0.41 [0.36, 0.46]	0.70 [0.63, 0.77]	0.46 [0.41, 0.52]	0.75 [0.69, 0.80]	0.46 [0.41, 0.50]	0.82 [0.75, 0.88]	0.36 [0.30, 0.42]	0.16 [0.11, 0.22]	0.07 [0.04, 0.09]	0.03 [0.01, 0.04]	0.91 [0.88, 0.93]
Δ A−B	+0.08 [+0.01, +0.16]	+0.18 [+0.09, +0.29]	+0.12 [+0.05, +0.20]	+0.22 [+0.12, +0.31]	+0.12 [+0.07, +0.17]	+0.23 [+0.12, +0.35]	+0.14 [+0.08, +0.20]	−0.08 [−0.17, +0.01] ∘	−0.15 [−0.25, −0.06]	−0.04 [−0.08, −0.01]	+0.19 [+0.09, +0.28]

Table 9: **Multi-metric ExpertQA validation.** Logistic-regression ensembles over COUNTERMEM features (1−DAP, CNS, NR) and RAGAS faithfulness on the ExpertQA human-support endpoint (n=200 under GPT-4o-mini, n=187 under Claude-Sonnet). Bootstrap delta of *RAGAS* + COUNTERMEM *full* vs *RAGAS combined*: under GPT-4o-mini, ROC-AUC $+0.150$ $[+0.070, +0.233]$ (p<0.001), PR-AUC $+0.115$ $[+0.047, +0.184]$ (p<0.001); under Claude-Sonnet, ROC-AUC $+0.066$ $[+0.030, +0.104]$ (p<0.001), PR-AUC $+0.032$ $[+0.004, +0.062]$ (p=0.012). The COUNTERMEM profile complements RAGAS under both judges, with the larger lift under the weaker judge.

	GPT-4o-mini judge		Claude-Sonnet judge	
Features	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
RAGAS combined	0.464	0.624	0.610	0.763
$1 - \mathrm{DAP}$	0.654	0.769	0.663	0.781
COUNTERMEM full	0.650	0.779	0.651	0.787
RAGAS + $1 - \mathrm{DAP}$	0.615	0.743	0.676	0.795
RAGAS + COUNTERMEM full	0.615	0.743	0.676	0.797

Table 10: Naturalistic conflict on DelucionQA (low-adherence subset, n=100), classified by a Claude-Sonnet conflict judge applied to the same model outputs in two conditions: full retrieved context (*base*) and one-side context removal (*half-removed*). Silent-pick is rare in the baseline but becomes the dominant non-abstain failure mode under the intervention.

Condition	Silent-pick↓	Conflict-ack.↑	Abstain	Confident-match
Base (full context)	0.04	0.00	0.15	0.76
Half-removed (intervention)	0.24	0.00	0.24	0.51

Table 11: End-to-end corpus deletion on RAGBench. For each query, gold-supporting chunks are removed from the indexed corpus; BM25 and FAISS are rebuilt; retrieval is rerun (k=4, BM25). Abstention rises to near-saturation on HotpotQA and PubMedQA, while TechQA remains a residual-corpus regime with plausible non-gold support. *Gold removed* is the average number of corpus chunks removed per query. *Abstain (base/del)* is the fraction of queries where the explicit-abstention regex fires. 95% bootstrap CI in brackets.

Dataset	n	Gold removed	Answer change↑	Abstain (base)	Abstain (del)↑
HOTPOTQA	100	4.08	0.241 [0.224, 0.258]	0.39 [0.29, 0.48]	0.99 [0.97, 1.00]
TECHQA	100	74.39	0.156 [0.137, 0.176]	0.51 [0.41, 0.61]	0.57 [0.48, 0.67]
PUBMEDQA	100	7.89	0.170 [0.157, 0.184]	0.39 [0.30, 0.49]	1.00 [1.00, 1.00]

Table 12: **Intervention-validity audit on the RAGBench sweep.** Stratified sample of n=50 rows per intervention class drawn from the 27-configuration sweep, judged by Claude-Sonnet (a model stronger than the system under test). Validity is the fraction of intervention payloads judged well-formed. The coherent-intervention probe of Section 5.3 uses a separate per-row audit (n=593, 93.6% contradictions well-formed); headline behavioral rates there are computed on the judged-valid contradiction subset (Limitations).

Probe	n	Validity	Headline detector
Base (no intervention)	50	100%	LLM judge
Delete-All	50	100%	LLM judge
Update injected	50	96%	content match
Contradiction injected	50	82%	LLM judge

Table 13: Metric sensitivity: DAP computed with three answer-similarity functions on the ExpertQA n=200 subsample. Pairwise Spearman correlations are high, indicating that the qualitative conclusions about no-evidence persistence are robust to the choice of similarity function.

	SAS (SBERT)	Cross-encoder	Jaccard
sas	–	0.70	0.85
crossencoder	0.70	–	0.82
jaccard	0.85	0.82	–

Table 14: Threshold sensitivity for semantic redundancy detection. % rows with at least one multi-document cluster, semantic SBERT clustering. Conclusions are robust around τ=0.6.

Dataset	τ=0.5	τ=0.6	τ=0.7	τ=0.8
HotpotQA	60.1	37.7	18.9	7.5
TechQA	90.8	67.8	46.7	28.3
PubMedQA	99.2	92.6	74.2	33.6

Table 15: Prompt sensitivity. Same RAGBench rows under two prompts: *strict* (answer ONLY from context, abstain if missing/conflicting; main-paper default) and *helpful* (helpful, no abstention rule). The strict prompt produces lower base quality (more abstentions) but substantially higher CONTRA acknowledgment rates.

Prompt	Base SAS↑	DAP↓	CS	DelAll↑	Contra↑
helpful	0.834	0.790	0.948	0.000	0.000
strict	0.844	0.717	0.948	1.000	0.010

Table 16: Repeated-generation sensitivity subset announced in Section 4. We sampled n=49 ExpertQA queries and called Claude-Haiku with temperature 0.7 three independent times per condition. Per-query DAP estimates are stable across the three samples, and the dataset-level aggregate moves only marginally across reps, supporting the use of single-generation DAP estimates for the main results.

Quantity	Value
Queries / repeats / temperature	49 / 3 / 0.70
Aggregate DAP across all reps	0.619 $[0.605, 0.631]$
Std of per-rep aggregate DAP	0.0029
Per-query DAP std (mean / median / P90)	0.037 / 0.035 / 0.062
Per-query DAP range max−min (mean / P90)	0.071 / 0.113