RELEVANCE IS NOT SUFFICIENT EVIDENCE: DETECTING EVIDENCE GAPS BEFORE GENERATION IN RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) grounds large language models (LLMs) in external sources. However, retrieved passages often name the right entities while lacking the actual facts an answer requires. Generators rarely notice: even with an explicit instruction to abstain, 12 generators still answer 40.0–99.3% of insufficient-evidence questions. A common remedy trains abstention into the generator through fine-tuning or reinforcement learning. This ties the decision to one model's weights, potentially rewarding correct answers recalled from its parametric knowledge as if the evidence had supported them. The decision arrives only as the generator's own output, so every insufficient-evidence question still costs a full generator call. We ask instead whether sufficiency can be judged from the question and evidence alone, *before* any answer exists. We first demonstrate several pitfalls in constructing tests for insufficient evidence. Directly removing the relevant evidence or pairing evidence with unrelated questions can introduce *cues* that reveal the label without testing sufficiency, such as lexical overlap and evidence position. We therefore build a paired benchmark from substitution, deletion, and question-swap constructions that vary answer support while controlling selected surface features, such as word use. On this benchmark, sufficiency can be judged without generating an answer, but no single signal works on every dataset. We introduce RINSE (**R**elevance **I**s **N**ot **S**ufficient **E**vidence), which combines three signals, one for each way evidence falls short: whether every part of the question is covered, whether any passage actually offers an answer, and whether a small language model, reading the passages together, treats them as sufficient. Across six datasets, RINSE ranks sufficient above insufficient evidence with a score of 0.837 (chance 0.5), ahead of the best of 10 prior methods (0.746) and of a frontier model asked the same question through an API (0.784). Its weakest dataset scores higher than any other method's weakest (0.684 vs. 0.676). It runs locally before generation, taking only 36.5 ms per question on a single GPU. Our code is available at https://anonymous.4open.science/r/RINSE-reproduction/.

1 INTRODUCTION

A large language model (LLM) answers based on its parameters, which are fixed at training time and do not hold private information due to proprietary concerns (Carlini et al., 2021), or more recent information due to the high cost of continual learning (Jang et al., 2022). Asked beyond what they store, models often hallucinate (Huang et al., 2025), producing a fluent and confident answer that no source supports (Mallen et al., 2023). Retrieval-augmented generation (RAG) was introduced to provide external sources that are more relevant and timely (Lewis et al., 2020). However, relevant passages are not always sufficient: a passage can name the right people, places, and events but miss the one fact or connection the question depends on. This raises the question: **how can a RAG system judge whether the retrieved passages contain enough information to answer the question?**

A RAG system cannot rely on its generator to notice the gap. Given insufficient context, models often answer from possibly stale parametric knowledge instead of abstaining (Joren et al., 2025; Peng et al., 2025). Joren et al. (2025) reports that the wrong-answer rate rises from 10.2% without context to 66.1% with insufficient context. We instruct 12 generators from 7 families (135M–32B parameters) to reply *unanswerable* when the documents lack the answer, on 300 sufficient and 300 insufficient records from the six evaluation datasets. They still answer 40.0% to 99.3% of the insuf-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ficient records, and they also refuse up to 29.7% of the sufficient ones (Table 1). Each such answer costs a full generation: 7–32B models take a median of 150–855 ms for an answer of up to 24 tokens. An evidence gap should therefore be caught before the generator is called.

A natural response is to train the generator to abstain through supervised fine-tuning (Zhang et al., 2024), preference alignment (Song et al., 2025b), or reinforcement learning with rewards that separate correct answers, hallucinations, and abstentions (Wei et al., 2025; Zhao et al., 2026). However, because the supervision comes from answer outcomes, what such training teaches is the *generator's own knowledge boundary* rather than a *judgment about the evidence*: the model is rewarded for answering the questions it knows and for abstaining on those it does not, regardless of which passages sit in the prompt. The two coincide only when the model's knowledge already matches the retrieved evidence, which is the assumption RAG exists to remove. On NeoQA (Glockner et al., 2025), built from fabricated news so that answers must come from the supplied passages, the four methods that read a generated answer rank the sufficient evidence set above the insufficient one barely above chance (0.503–0.520, where chance is 0.5). Our method, which reads only the question and the evidence, reaches 0.773, and its content-coverage reader alone reaches 0.799 (Section 5). In addition, such training-based methods are bound to the model being trained, so they must be repeated for every model update, cannot be installed on a closed-source model exposed only through an API, where a user supplies training data but not the training objective, and can erode under later training with other objectives (Song et al., 2025a; Kirichenko et al., 2025). We argue, instead, that **whether a set of passages answers a question is a property of the question and the retrieved passages**, not of the model that will read them. We therefore ask **whether that property can be computed cheaply and without generating an answer**.

Given a question q and a set of passages E, the task is to decide whether E contains the information needed to derive a complete answer, including any intermediate facts, before generation begins. Related methods answer neighboring questions: relevance rankers ask what a passage is about (Sachan et al., 2022; Es et al., 2024), verification and confidence methods read an answer once it exists (Honovich et al., 2022; Kadavath et al., 2022), and retrieval controllers ask what a particular model still needs (Cheng et al., 2024). The closest existing work is Sufficient Context (Joren et al., 2025), which judges the same property with a large autorater—a frontier model at best, a fine-tuned 11B entailment model at its smallest; whether a *lightweight* local scorer can make the same decision remains open.

A benchmark for this task compares passages that support an answer with passages that do not, so anything else that differs between the two sides becomes a *cue* a scorer can use in place of sufficiency. For example, the share of the question's words that appear in the passage separates answerable from unrelated questions on SQuAD 2.0 almost perfectly (0.993, a *lexical cue*), yet falls to 0.573, near chance, once the unanswerable question is written about the passage.

We build a paired benchmark whose sufficient and insufficient sides share a question or an evidence set, and we audit the remaining construction cues. We use three constructions: 1) keep the question fixed and swap a needed passage for a related distractor; 2) pair one passage with questions written to be about it but unanswerable from it; 3) delete a needed passage from evidence sets whose sizes already vary widely, so that passage count reveals little about the label.

These constructions expose three different ways that evidence falls short. It can leave out something that the question asks for, stay on topic while offering no answer at all, or carry the answer across several passages, none of which is sufficient alone (Ferguson et al., 2020; Trivedi et al., 2022). Inspired by these failure modes, we propose RINSE, a lightweight scorer with one reader for each, checking whether every part of the question is covered, whether any passage offers an answer, and whether a small language model, reading the passages together, treats them as sufficient. Each signal is corrected for the size of the evidence set, and the three are summed with equal weights.

Averaged over six datasets, RINSE scores 0.837, ahead of all 10 prior methods, including Self-RAG (Asai et al., 2024) at 0.746 and UPR (Sachan et al., 2022) at 0.721. It is also the most consistent: its weakest dataset scores 0.684, while every other local method drops to 0.644 or lower on at least one. Asked directly whether the evidence suffices, a local 1.5B model scores 0.540 and a frontier model queried through an API scores 0.784; RINSE ranks above both. On evidence returned by real retrievers, RINSE scores 0.813, against 0.674 for the strongest prior method. RINSE needs neither a generated answer nor an API call: three lightweight components, the largest of them 1.5B

Table 1: Twelve generators instructed to reply *unanswerable* when the documents lack the answer, on 300 sufficient and 300 insufficient records drawn evenly from the six datasets of Table 4. *Answers*: share of insufficient records answered; *Refuses*: share of sufficient records refused; *ms*: median latency of a greedy answer of up to 24 tokens, with passages capped at 1,600 characters.

Model	Answers (%)	Refuses (%)	ms	Model	Answers (%)	Refuses (%)	ms
Qwen3-0.6B	53.7	29.7	131	SmolLM2-135M	99.3	0.3	358
Qwen3-1.7B	64.3	13.3	147	Llama-3.2-1B	81.7	12.7	92
Qwen3-4B	55.3	16.0	211	Mistral-7B-v0.2	68.0	7.7	448
Qwen3-8B	61.3	11.3	205	Falcon3-7B	68.0	16.3	150
Qwen3-14B	40.7	27.0	331	Zephyr-7B-beta	82.7	3.0	469
Qwen3-32B	48.7	11.7	855	Phi-4	40.0	18.3	588

parameters, take 36.5 ms per record on a single GPU (Section 5), less than the median time of any of the 12 short generations in Table 1 (92–855 ms).

2 Related work

The closest line of work judges **evidence sufficiency** directly. Sufficient Context (Joren et al., 2025) asks whether a diligent reader could answer from the provided context alone, the same target as RINSE, and labels it with an autorater ranging from an 11B entailment model to a frontier LLM; S2G-RAG (Li et al., 2026) makes the same judgment inside an iterative retrieval loop. We compare with Sufficient Context on both a local and a frontier model, and study how negative construction shapes the evaluation. A second line performs **relevance scoring**: UPR (Sachan et al., 2022), RAGAS (Es et al., 2024) and CRAG (Yan et al., 2024) judge whether a passage is about the question, yet a relevant passage can omit the fact the answer needs, and a per-passage grade misses support spread across passages. Many methods instead read **generator-side signals**. Retrieval controllers such as UAR (Cheng et al., 2024), FLARE (Jiang et al., 2023) and Self-RAG (Asai et al., 2024), and answer-side confidence estimates such as P(True) (Kadavath et al., 2022), semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) and attribution (Honovich et al., 2022; Gao et al., 2023), all depend on one generator's hidden states or outputs, whereas RINSE reads only the question and the evidence. Methods for **trained abstention**, including R-Tuning, Trust-Align, TruthRL and GRACE (Zhang et al., 2024; Song et al., 2025b; Wei et al., 2025; Zhao et al., 2026), teach one generator to refuse, and later fine-tuning erodes that behavior (Song et al., 2025a; Kirichenko et al., 2025); a generator-independent score could supply their rewards with a sufficiency signal. Finally, work on **constructed negatives** found hypothesis-only and lexical-overlap artifacts in natural language inference (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019), and SQuAD 2.0 replaced generated unanswerable questions with authored near-misses (Rajpurkar et al., 2018); our cue audit applies the same scrutiny to evidence sufficiency (Section 3). Appendix G gives details.

3 Constructing and Auditing Sufficiency Contrasts

Evaluating evidence sufficiency requires constructing insufficient evidence, yet different constructions test different capabilities and expose different unintended cues. A score therefore reflects the construction as well as sufficiency, so we design the evaluation before the method: we construct paired sufficiency contrasts, audit which cues other than sufficiency still predict the label, and use the resulting failure regimes to motivate the three readers introduced in Section 4.

3.1 Paired Sufficiency Contrasts

Each contrast pairs a question q and a sufficient passage set E^+ with an insufficient counterpart from the same underlying example. Some counterparts are rebuilt by us and others come from the source dataset. Figure 1 illustrates the three constructions:

- **Substitution** replaces a required passage with a distractor while preserving the question and the number of passages. It is our primary construction because it holds the passage count fixed: on

q *When* was *Parasite*'s director born?

d_1 *Parasite* was directed by Bong Joon-ho.

d_2 Bong Joon-ho was born in 1969.

d_3 *Parasite* won an award in 2019.

PAIR	QUESTION	SELECTED PASSAGES	SUFFICIENCY	CUE ANALYSIS
Reference	q	$\{d_1, d_2\}$	**Sufficient** 1969	—
Deletion	q (same)	$\{d_1\}$	**Insufficient** no birth year	Passage count: $2 \rightarrow 1$
Substitution	q (same)	$\{d_1, d_3\}$	**Insufficient** no birth year	Evidence-only patterns
Question swap	q': *When* $\rightarrow$ *Where*	$\{d_1, d_2\}$	**Insufficient** no birthplace	Question wording

Figure 1: Paired evidence contrasts. Green marks supporting passages; orange marks the distractor, changed inputs, and insufficient outcomes. Deletion and substitution change passages with q fixed; question swap changes q with the passages fixed. The last column lists possible cues.

our HotpotQA rebuild (Yang et al., 2018), the passage count alone scores 0.500. Section 3.3 audits the cues it can leave, such as the distractor's low word overlap with the question.

- **Deletion** removes required support. It creates the evidence gap directly, but it also shrinks the evidence set, so a scorer could call the smaller set insufficient without reading it. We therefore take deletion pairs from NeoQA, where each question comes with many evidence sets of one to 120 passages; there, set size is only weakly tied to sufficiency, and passage count alone scores 0.563.
- **Question swap** keeps the evidence fixed and pairs it with a question that is related to the passage but cannot be answered from it. This construction tests whether a scorer can tell a question the passage answers from a related one it does not. The choice of replacement question is critical. On SQuAD 2.0 (Rajpurkar et al., 2018), counting how many of the question's words appear in the passage separates answerable from unrelated questions almost perfectly (0.993), but scores only 0.573 against the unanswerable questions the dataset's annotators wrote about each passage. Section 3.3 audits the remaining question-wording cues.

3.2 Paired Evaluation

We evaluate a scorer by whether it orders the two sides of each contrast correctly. Records built from the same example form a group; within it, each sufficient–insufficient pair contributes one point if the sufficient side receives the higher score, one half for a tie, and zero otherwise. Averaging these outcomes gives the paired concordance $\mathrm{AUC}_{\mathrm{wq}}$, where 0.5 indicates chance and 1.0 indicates perfect ordering. When a group contains multiple sufficient–insufficient pairs, $\mathrm{AUC}_{\mathrm{wq}}$ weights groups by their number of pairs. Confidence intervals use a cluster bootstrap with 2,000 draws that resamples complete groups, preserving dependence among records derived from the same question or passage. We report each construction separately. Pooling unrelated groups into one AUC would let between-group differences inflate the score of a scorer that fails within each example (Robinson, 1950; Simpson, 1951).

3.3 Contrast Sources and Residual Cues

For HotpotQA (Yang et al., 2018), we rebuild fixed-size contrasts by replacing an annotated support passage with a sampled distractor. MuSiQue (Trivedi et al., 2022) supplies its own paired answerable and unanswerable variants, which we use without selecting new distractors. NeoQA (Glockner et al., 2025) supplies native test deletion variants, which we use without sampling new deletions. SQuAD 2.0 (Rajpurkar et al., 2018) and fresh Wikipedia hold the evidence fixed and vary the question, while IIRC (Ferguson et al., 2020) groups questions by source article, and each question keeps its own supplied evidence. Appendix B specifies the pairing and sampling rules.

Residual cues. A scorer might distinguish sufficient from insufficient records using a cue that does not establish whether the passages answer the question. We test passage count, minimum

Table 2: Construction cues under shared passage additions. Columns 0, +5, +10, +15, and +20 count passages added to each side before BM25 orders the full set; the 0 condition is reordered too. Entries are $\mathrm{AUC}_{\mathrm{wq}}$ (chance 0.5; lower values point the other way). *Size* (passage count) and *Q-only* (question-only classifier) do not depend on k; minimum IDF uses the $k = 0$ weights, and evidence-only is refitted at each count.

	Any k		**Min. IDF coverage**					**Evidence-only**				
Dataset	Size	Q-only	0	+5	+10	+15	+20	0	+5	+10	+15	+20
SQuAD	0.500	0.607	0.577	0.524	0.529	0.536	0.535	0.500	0.494	0.472	0.481	0.484
IIRC	0.891	0.691	0.248	0.486	0.495	0.497	0.500	0.839	0.754	0.708	0.675	0.665
NeoQA	0.563	0.500	0.489	0.500	0.500	0.500	0.500	0.568	0.545	0.530	0.521	0.515
HPQA	0.500	0.500	0.941	0.541	0.520	0.513	0.510	0.763	0.758	0.750	0.745	0.743
MuSQ	0.501	0.500	0.524	0.507	0.504	0.502	0.502	0.759	0.763	0.766	0.767	0.768
Wiki	0.500	0.656	0.939	0.567	0.558	0.552	0.544	0.500	0.508	0.500	0.519	0.490

Table 3: Cue-controlled subsets: paired concordance $\mathrm{AUC}_{\mathrm{wq}}$ with 95% group-bootstrap intervals. HotpotQA passages use a fixed shuffle; each row restricts pairs to the stated control.

Dataset	**Control**	**Pairs**	**RINSE**	**Self-RAG**	**UPR**	**1.5B yes/no**	**API Suff. Ctx.**
HotpotQA	min. IDF coverage tied	771	**0.776** [0.75, 0.81]	0.665 [0.65, 0.68]	0.737 [0.71, 0.76]	0.619 [0.58, 0.65]	0.709 [0.69, 0.73]
MuSiQue	answer string on both sides	1,796	**0.835** [0.82, 0.85]	0.677 [0.66, 0.69]	0.795 [0.78, 0.81]	0.564 [0.54, 0.59]	0.730 [0.72, 0.74]
IIRC	equal $\lvert E\rvert$	414	0.773 [0.72, 0.82]	0.650 [0.60, 0.70]	0.558 [0.51, 0.61]	0.599 [0.54, 0.66]	**0.774** [0.74, 0.80]
NeoQA	equal support count	832	**0.644** [0.58, 0.71]	0.377 [0.30, 0.46]	0.438 [0.36, 0.52]	0.112 [0.08, 0.15]	0.567 [0.52, 0.61]

IDF coverage (the lowest share of the question's IDF-weighted words found in any single passage), and classifiers that see only the question or only the passages (Appendix C). Table 2 shows how often each cue gives the sufficient side of a pair the higher score. Passage count alone scores 0.891 on IIRC. A classifier reading only the question scores 0.607–0.691 on SQuAD, IIRC, and fresh Wikipedia.

Adding passages hides some cues but not others. On HotpotQA, the distractor usually covers fewer of the question's important words than the passage it replaces, so minimum IDF coverage alone scores 0.941 ($k = 0$). We then add the same 5, 10, 15, or 20 passages to both sides and order each full set with BM25 (Appendix F). With 20 additions, an added passage often has the lowest coverage on both sides. The two sides tie on this cue in 97.9% of pairs, and its score falls to 0.510. The evidence-only classifier still scores 0.743, showing that it can use other differences in the passages.

Table 3 makes a second check. It keeps only pairs that match on one cue: minimum IDF coverage for HotpotQA, passage count for IIRC, number of annotated supporting passages for NeoQA, and answer-string presence for MuSiQue. RINSE has the highest score among local methods in each subset. Each check controls one cue at a time; for example, in 831 of the 832 NeoQA pairs with equal support counts, the sufficient side still has fewer passages.

Failure regimes. The resulting contrasts show three distinct ways evidence can be insufficient. First, a set of passages can omit support for one part of the question, even while its remaining passages are relevant. Substitution and deletion make this *missing-coverage* regime visible. Second, passages can be on topic but offer no answer-bearing span; question swaps isolate this *topical-but-answerless* regime. Third, the needed facts can be spread across passages, such that no individual passage is sufficient, but their combination is. This *cross-passage* regime occurs in multi-hop settings including HotpotQA, IIRC, and MuSiQue. The regimes overlap, and together they explain

why the best single relevance, answerability, or generator-confidence signal differs from dataset to dataset. Section 4 therefore builds one answer-free reader for each regime.

4 THE RINSE SCORER

Given a question q and an evidence set E, let $h_r(q, E)$ be the raw score from reader $r \in \mathcal{R}$, where $\mathcal{R} = \{\text{content}, \text{span}, \text{latent}\}$. We define RINSE as

$$s(q, E) = \sum_{r \in \mathcal{R}} \frac{h_r(q, E) - \widehat{c}_r(|E|) - \widehat{\mu}_r}{\widehat{\sigma}_r}.$$

All hatted quantities are estimated only on the training fold. The size correction $\widehat{c}_r$ is the linear trend $\widehat{\alpha}_r + \widehat{\beta}_r \log(1 + |E|)$ fitted across training records. For the span and latent readers, sets larger than any training set instead follow a size curve $\widehat{f}_r$ learned from padded training records, $\widehat{c}_r(|E|) = \widehat{\alpha}_r + \widehat{\beta}_r \log(1 + N) + \widehat{f}_r(|E|) - \widehat{f}_r(N)$ for $|E| > N$, where N is the largest training set. $\widehat{\mu}_r$ and $\widehat{\sigma}_r$ are the mean and standard deviation of the size-corrected scores $h_r - \widehat{c}_r$. A larger $s(q, E)$ indicates more sufficient evidence. Every component scores the input without decoding text, and the largest backbone makes one forward pass.

Content coverage. The content reader measures whether every local part of the question is supported somewhere in the evidence. We split the question into overlapping four-WordPiece windows with stride two (Devlin et al., 2019). Each window keeps its best match across passages under both IDF-weighted overlap (Spärck Jones, 1972) and a 118M multilingual reranker trained on mMARCO (Wang et al., 2021; Bonifacio et al., 2021). A match must come from one source, so unrelated fragments cannot jointly satisfy a window; chunks with the same source identifier are rejoined before matching. Quantiles of the two coverage profiles feed a conditional-logistic head fitted on paired training examples (Breslow et al., 1978).

Answer-span evidence. The span reader asks whether any passage contains text that an extractive QA model prefers to its no-answer option. We train a 125M RoBERTa-base reader (Liu et al., 2019) with explicit null supervision on Natural Questions (Kwiatkowski et al., 2019): each negative replaces the answer sentences with a length-matched topical window (Appendix E). At inference, the reader scores every passage and overflow window by its best legal span logit minus its no-answer logit, and returns the maximum margin.

Cross-passage evidence. The latent reader captures evidence that becomes informative only when passages are read together. Frozen Qwen2.5-1.5B-Instruct (Qwen Team, 2024) encodes the question and the five highest-reranked passages in one forward pass, preserving their input order. We project the unit-normalized final-token state at layer 17 onto a fixed direction: the family-balanced mean of sufficient-minus-insufficient hidden-state differences estimated from the other training families. The five-passage cap bounds context length and exposure to irrelevant material.

Size correction and training. Raw reader scores differ in scale and grow with evidence-set size; the correction and standardization above put them on one scale. Because the best reader varies across datasets (Section 6), the weights stay equal and fixed rather than tuned per dataset. Training sets hold at most 22 passages while NeoQA's reach 120. To extend the correction past that range, we pad each training record with 5 to 100 passages from other questions of its dataset, which changes its size but not its label, and fit a piecewise-linear curve in $\log(1 + |E|)$ with knots at the deciles within records; its slope falls from 2.9 for small sets to about 1 between 20 and 100 passages, against 3.1 for the linear fit. Only NeoQA and one IIRC record exceed the training range. On NeoQA validation timelines, a flat tail beat linear extrapolation and the padding curve matched it (Appendix D). The padding curve and latent-reader training on the same top five passages used at inference together raise NeoQA from 0.673 to 0.773. The neural backbones are frozen and shared across datasets. RINSE outputs a ranking score; each deployment thresholds it to answer, abstain or retrieve more.

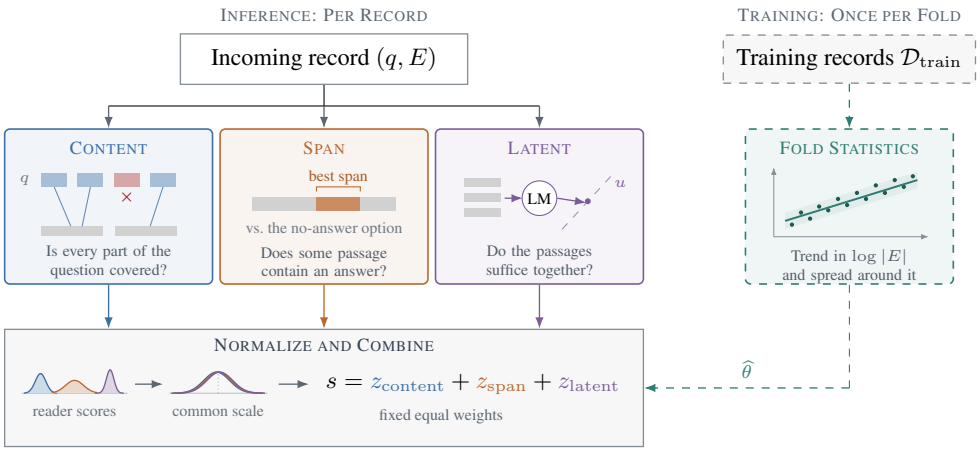

Figure 2: RINSE combines content coverage (blue), answer-span evidence (orange), and latent cross-passage evidence projected onto u (purple). Solid paths show inference; the dashed path fits training-fold statistics for size correction and standardization before an equal-weight sum.

5 EXPERIMENTS

Evaluation protocol. We compare scorers on the same sufficient–insufficient pairs from six datasets using paired concordance AUC_{wq} (Section 3); the cue checks use these records too. HotpotQA substitution passages are shuffled once, so the distractor's position carries no signal (Appendix B). We assemble 13,600 training records from MuSiQue (Trivedi et al., 2022), 2WikiMultiHopQA (Ho et al., 2020), IIRC (Ferguson et al., 2020), SQuAD 2.0 (Rajpurkar et al., 2018), HotpotQA (Yang et al., 2018), and CRAG (Yang et al., 2024). For each evaluated family, we fit the content head, latent direction, and normalization without records from that family. The span reader is trained once on Natural Questions (Kwiatkowski et al., 2019) and shared across datasets. We compare with methods based on passage relevance (Asai et al., 2024; Sachan et al., 2022; Es et al., 2024), retrieval control (Cheng et al., 2024; Yan et al., 2024; Jiang et al., 2023), and confidence in a generated answer (Kadavath et al., 2022; Honovich et al., 2022; Gao et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024). The closest baseline, Sufficient Context (Joren et al., 2025), asks a model directly whether the context is sufficient; we run its official prompt on GPT-6 Luna and on Qwen2.5-1.5B-Instruct (Qwen Team, 2024). Other local methods use the same 1.5B model where possible, and the UAR, CRAG, and Self-RAG classifiers keep their own backbones (Appendix A).

Main results. RINSE has both the highest average (0.837, against 0.746 for Self-RAG, the strongest prior method) and the highest minimum (Table 4). Asked directly whether the evidence suffices, the Sufficient Context prompt scores 0.540 on the local 1.5B model and 0.784 on the frontier model, whose yes-or-no labels tie on 26–57% of pairs, each counted as one half under AUC_{wq}. RINSE leads the frontier model on five of six datasets, with a mean paired difference of $+0.053$ $[+0.047, +0.059]$. RINSE's weakest dataset, SQuAD 2.0, still scores 0.684; every other local method drops to 0.644 or lower somewhere, and eight of the eleven drop to 0.520 or lower. RINSE leads on IIRC, NeoQA, HotpotQA, MuSiQue and fresh Wikipedia, spanning all three constructions and IIRC's native contrast. Among local methods, Self-RAG leads on SQuAD 2.0 (0.752 versus 0.684 for RINSE). On NeoQA, RINSE's content reader alone reaches 0.799, above the full sum (0.773) and every other local scorer (Section 6). NeoQA's news articles still contain plausible answer spans after the needed fact is deleted, so the span reader scores near chance there (0.508) and lowers the sum.

Methods that read a generated answer fail on NeoQA: all four score 0.503–0.520. NeoQA's events are invented, so the generator cannot have memorized them, yet it stays confident in answers its evidence does not support. UAR's decision tree scores 0.497 on average, at chance.

Real retrieval output. To test RINSE beyond constructed contrasts, we compare the top five passages returned by BM25 and e5-base-v2 for each question. Across five datasets, we retain 3,149

Table 4: Paired concordance $\mathrm{AUC}_{\mathrm{wq}}$ (chance 0.5), single-record latency (ms), and saturated throughput (records/s). Datasets: SQuAD 2.0, IIRC, NeoQA, HotpotQA substitution (HPQA), MuSiQue (MuSQ), and fresh Wikipedia (Wiki). Avg/Min summarize the six datasets; timings use 57 records from each on one GPU. Shading marks the best three local methods; brackets give 95% group-bootstrap intervals (Avg combines per-dataset errors). HPQA passages use a fixed shuffle.

Method	**Size**	**SQuAD** Q-swap	**IIRC** Native	**NeoQA** Delete	**HPQA** Subst.	**MuSQ** Subst.	**Wiki** Q-swap	**Avg**	**Min**	**p50** ms	**p95** ms	**rec/s**
Frontier model through an API (reference)												
Suff. Context	API	0.799 [0.79, 0.81]	0.813 [0.80, 0.83]	0.676 [0.66, 0.69]	0.764 [0.76, 0.77]	0.788 [0.78, 0.79]	0.864 [0.85, 0.87]	0.784 [0.78, 0.79]	0.676	API call		
Local, reads the question and evidence												
Suff. Context	1.5B	0.539 [0.53, 0.55]	0.584 [0.57, 0.60]	0.514 [0.51, 0.52]	0.521 [0.51, 0.53]	0.514 [0.51, 0.52]	0.567 [0.55, 0.58]	0.540 [0.53, 0.54]	0.514	5566	5934	0.43
yes/no prompt	1.5B	0.720 [0.70, 0.74]	0.579 [0.56, 0.60]	0.500 [0.49, 0.51]	0.684 [0.67, 0.70]	0.569 [0.56, 0.58]	0.933 [0.92, 0.94]	0.664 [0.66, 0.67]	0.500	31.8	424	8.1
UPR	1.5B	0.699 [0.68, 0.72]	0.541 [0.52, 0.56]	0.699 [0.69, 0.71]	0.712 [0.70, 0.72]	0.731 [0.72, 0.74]	0.946 [0.93, 0.96]	0.721 [0.72, 0.73]	0.541	24.5	537	9.8
RAGAS-CR	1.5B	0.588 [0.57, 0.60]	0.485 [0.48, 0.49]	0.508 [0.50, 0.51]	0.549 [0.54, 0.56]	0.513 [0.51, 0.52]	0.665 [0.65, 0.68]	0.551 [0.55, 0.56]	0.485	77.9	1764	5.1
CRAG	770M	0.626 [0.61, 0.65]	0.584 [0.56, 0.61]	0.549 [0.54, 0.56]	0.609 [0.60, 0.62]	0.633 [0.63, 0.64]	0.794 [0.77, 0.81]	0.632 [0.63, 0.64]	0.549	39.4	968	5.7
UAR-tree	7B	0.527 [0.52, 0.54]	0.467 [0.45, 0.48]	0.481 [0.48, 0.48]	0.500 [0.50, 0.50]	0.502 [0.50, 0.50]	0.506 [0.49, 0.52]	0.497 [0.49, 0.50]	0.467	72.6	780	3.1
Self-RAG	7B	0.752 [0.73, 0.77]	0.688 [0.67, 0.71]	0.644 [0.63, 0.65]	0.696 [0.69, 0.70]	0.759 [0.75, 0.77]	0.939 [0.93, 0.95]	0.746 [0.74, 0.75]	0.644	44.3	2041	2.8
Local, reads a generated answer												
P(True)	1.5B	0.670 [0.65, 0.69]	0.607 [0.58, 0.63]	0.520 [0.51, 0.53]	0.590 [0.58, 0.60]	0.571 [0.56, 0.58]	0.854 [0.84, 0.87]	0.635 [0.63, 0.64]	0.520	164	587	7.4
Attrib.	1.5B	0.699 [0.68, 0.72]	0.615 [0.59, 0.64]	0.517 [0.51, 0.52]	0.570 [0.56, 0.58]	0.551 [0.54, 0.56]	0.929 [0.92, 0.94]	0.647 [0.64, 0.65]	0.517	142	630	9.3
SemEnt	1.5B	0.593 [0.57, 0.61]	0.602 [0.58, 0.62]	0.505 [0.50, 0.51]	0.539 [0.53, 0.55]	0.544 [0.53, 0.55]	0.589 [0.57, 0.61]	0.562 [0.56, 0.57]	0.505	499	1382	2.4
FLARE	1.5B	0.586 [0.57, 0.61]	0.627 [0.60, 0.65]	0.503 [0.49, 0.51]	0.539 [0.53, 0.55]	0.565 [0.55, 0.58]	0.666 [0.64, 0.69]	0.581 [0.57, 0.59]	0.503	184	778	11.9
RINSE	1.5B	0.684 [0.66, 0.70]	0.926 [0.91, 0.94]	0.773 [0.76, 0.78]	0.853 [0.84, 0.86]	0.813 [0.80, 0.82]	0.973 [0.97, 0.98]	**0.837** [0.83, 0.84]	**0.684**	36.5	603	17.4

pairs for which one retrieved set contains all annotated supporting passages and the other does not (Table 5). With the evaluated family held out during fitting, RINSE averages 0.813, ahead of Self-RAG (0.674), UPR (0.661) and the yes/no prompt (0.590), and leads the local methods on four of the five datasets; UPR leads on SQuAD 2.0 (0.871 vs. 0.813). Both sets hold five passages, so set size carries no signal, and always choosing one retriever's set scores 0.204–0.796; RINSE beats both choices on every dataset. The frontier model averages 0.800 on the same pairs, below RINSE, and leads on 2WikiMultiHopQA (0.899 vs. 0.840).

Overlap and added passages. We remove exact matches to evaluation questions and passages from the span reader's Natural Questions training data, then audit residual shared sentences; overlap is highest in HotpotQA and MuSiQue (Appendix E). Adding the same 5–20 passages to both sides of each group and BM25-ordering the full sets lowers RINSE's mean $\mathrm{AUC}_{\mathrm{wq}}$ from 0.847 at $k = 0$ to 0.804 at $k = 20$ (Appendix F).

Cost. On one NVIDIA A100 GPU, RINSE takes 36.5 ms per record at the median, versus 24.5 ms for UPR and 31.8 ms for the 1.5B yes/no prompt. It has the highest local throughput, 17.4 records/s versus FLARE's 11.9. Its 603 ms 95th-percentile latency comes mostly from the 118M reranker (465 ms); the 1.5B pass takes 29.0 ms (Table 4; Appendix D).

Because RINSE runs before the generator, it can also decide whether to call it. As a gate in front of the 12 generators of Table 1, thresholded on training folds and applied to the same 600 records, it lowers the mean rate of answering on insufficient evidence from 0.636 to 0.511 and saves 13.3% of generation time, while the rate of answering sufficient records falls from 0.861 to 0.783.

Table 5: Paired concordance $\mathrm{AUC}_{\mathrm{wq}}$ on 3,149 BM25-versus-dense retrieval pairs; 2Wiki is 2WikiMultiHopQA and the other abbreviations follow Table 4. The last two rows always choose the named retriever's set. *Avg*: unweighted mean over datasets. **Bold**: best local method.

	SQuAD	**IIRC**	**MuSQ**	**HPQA**	**2Wiki**	**Avg**
Pairs	481	556	608	1,147	357	
Suff. Context (API)	0.810	0.725	0.785	0.783	0.899	0.800
RINSE	0.813	**0.766**	**0.808**	**0.837**	**0.840**	**0.813**
UPR	**0.871**	0.583	0.664	0.588	0.601	0.661
Self-RAG	0.798	0.589	0.677	0.678	0.627	0.674
yes/no prompt	0.636	0.662	0.572	0.616	0.462	0.590
dense side	0.536	0.426	0.340	0.796	0.742	0.568
BM25 side	0.464	0.574	0.660	0.204	0.258	0.432

6 Component Analysis

Table 6: Reader ablation, $\mathrm{AUC}_{\mathrm{wq}}$. Every subset uses the same size correction and equal-weight sum. Abbreviations follow Table 4; *Avg* and *Min* are over the six datasets. **Bold**: best per column.

Readers	**SQuAD**	**IIRC**	**NeoQA**	**HPQA**	**MuSQ**	**Wiki**	**Avg**	**Min**
RINSE (all three)	0.684	**0.926**	0.773	0.853	**0.813**	**0.973**	**0.837**	**0.684**
content + span	0.618	0.912	0.740	**0.887**	0.785	0.958	0.817	0.618
content + latent	0.664	0.772	0.764	0.835	0.796	0.962	0.799	0.664
span + latent	**0.716**	0.925	0.607	0.756	0.731	0.940	0.779	0.607
content	0.579	0.607	**0.799**	0.861	0.762	0.941	0.758	0.579
span	0.635	0.904	0.508	0.644	0.625	0.855	0.695	0.508
latent	0.712	0.827	0.583	0.714	0.702	0.927	0.744	0.583

Table 6 evaluates every subset of the three readers. We observe that each reader leads where its failure regime dominates. For example, content coverage is the strongest single reader on the evidence-removal contrasts, NeoQA (0.799), HotpotQA (0.861) and MuSiQue (0.762), where the remaining passages stay on topic and still hold answer-like spans. The span reader leads on IIRC (0.904), and the latent reader on SQuAD 2.0 (0.712), whose authored near-misses reuse the passage's vocabulary. A deployed scorer cannot tell which contrast a record comes from, so RINSE sums all three readers. Removing any reader lowers the average: to 0.817 without latent, 0.799 without span, and 0.779 without content. The best single reader's weakest dataset scores 0.583, against 0.684 for the sum. The full sum has the best average and minimum of all seven subsets.

7 Conclusion

Retrieved passages can be relevant without being sufficient, and generators rarely notice the gap. Instead of relying on the generator, we asked whether sufficiency can be judged from the question and evidence alone, before generation. Testing this requires care, because direct ways of constructing insufficient evidence can leave shortcuts that solve the test without judging sufficiency; we therefore built paired contrasts and audited their residual cues. The contrasts show three ways evidence falls short: the question's content is missing, on-topic passages offer no answer, or support is spread across passages. No single reader covers all three, so we propose RINSE, which sums three complementary readers, one per failure. In our evaluation, RINSE has the highest average and the highest weakest-dataset score of all compared methods, including a frontier model asked the same question. The advantage carries over to real retriever output, and RINSE takes only 36.5 ms per question on one GPU. Sufficiency can thus be judged cheaply and locally, before any generator is called.

Statement on the use of AI

Generative AI tools were used during the preparation of this manuscript to assist with language and formatting, literature retrieval and discovery, research ideation and execution, including artifact implementation and experimental support, and the generation of synthetic questions for the fresh Wikipedia evaluation set described in Appendix B. The authors reviewed the resulting text, equations, tables, figures, and citations, and take responsibility for the final content.

Ethics statement

This work uses publicly available QA datasets and Wikipedia content and involves no human subjects or collection of private data. The synthetic questions may inherit biases from the generating model; their construction and filtering are documented in Appendix B.

Reproducibility statement

The paper and appendices specify RINSE's algorithm, reader architectures, training procedures, dataset-construction rules, baseline adaptations, overlap audits, timing methodology, and statistical analysis. All train/test splits use disjoint question groups, and confidence intervals are computed using the group-level bootstrap procedure described in Section 3. The anonymous reproduction repository at https://anonymous.4open.science/r/RINSE-reproduction/ provides the code and anonymized numerical inputs to refit the final RINSE fusion and recompute the reported six-dataset main results, frontier-model comparison, and added-passage analysis. It also includes scripts for verifying paired concordance and group-bootstrap intervals.

References

Amazon Web Services. OpenAI GPT-6 Sol and GPT-6 Luna are now generally available on Amazon Bedrock. AWS What's New, September 22, 2026, 2026. URL https://aws.amazon.com/about-aws/whats-new/2026/09/openai-gpt-6-sol-luna-on-amazon-bedrock/.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Ingeol Baek, Hwan Chang, ByeongJeong Kim, Jimin Lee, and Hwanhee Lee. Probing-RAG: Self-probing to guide language models in selective document retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3287–3304. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.181. URL https://aclanthology.org/2025.findings-naacl.181/.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. mMARCO: A multilingual version of the MS MARCO passage ranking dataset. *CoRR*, abs/2108.13897, 2021. doi: 10.48550/arXiv.2108.13897. URL https://arxiv.org/abs/2108.13897.

N. E. Breslow, N. E. Day, K. T. Halvorsen, R. L. Prentice, and C. Sabai. Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology*, 108 (4):299–307, 1978. doi: 10.1093/oxfordjournals.aje.a112623. URL https://doi.org/10.1093/oxfordjournals.aje.a112623.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. Unified active retrieval for retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 17153–17166. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.999. URL https://aclanthology.org/2024.findings-emnlp.999/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246/.

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

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

James Ferguson, Matt Gardner, Hannaneh Hajishirzi, Tushar Khot, and Pradeep Dasigi. IIRC: A dataset of incomplete information reading comprehension questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1137–1147. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.86. URL https://aclanthology.org/2020.emnlp-main.86/.

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

Max Glockner, Xiang Jiang, Leonardo F. R. Ribeiro, Iryna Gurevych, and Markus Dreyer. NeoQA: Evidence-based question answering with generated news events. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11842–11926. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.616. URL https://aclanthology.org/2025.findings-acl.616/.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th*

International Conference on Computational Linguistics, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3905–3920. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287/.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://dl.acm.org/doi/10.1145/3703155.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=vfsRB5MImo9.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Sahib Julka. When confidence takes the wrong path: Diagnosing retrieval-state lock-in in RAG. *CoRR*, abs/2606.22728, 2026. doi: 10.48550/arXiv.2606.22728. URL https://arxiv.org/abs/2606.22728.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022. doi: 10.48550/arXiv.2207.05221. URL https://arxiv.org/abs/2207.05221.

Adrian Kieback, Uyiosa Philip Amadasun, Aman Chadha, and Aaron Elkins. TASR: Training-free adaptive stopping for iterative retrieval. *CoRR*, abs/2606.13814, 2026. doi: 10.48550/arXiv.2606.13814. URL https://arxiv.org/abs/2606.13814. Accepted at the Agent4IR Workshop, KDD 2026.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *Advances in Neural Information Processing Systems 38, NeurIPS 2025*, 2025. URL https://arxiv.org/abs/2506.09038.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion

Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang, and Guodong Zhou. S2G-RAG: Structured sufficiency and gap judging for iterative retrieval-augmented QA. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25846–25862. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1185. URL https://aclanthology.org/2026.acl-long.1185/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. doi: 10.48550/arXiv.1907.11692. URL https://arxiv.org/abs/1907.11692.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 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)*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334/.

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. Adaptive retrieval without self-knowledge? bringing uncertainty back home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6355–6384. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.319. URL https://aclanthology.org/2025.acl-long.319/.

Jaewan Park, Solbee Cho, and Jay-Yoon Lee. Stop-RAG: Value-based retrieval control for iterative RAG. *CoRR*, abs/2510.14337, 2025. doi: 10.48550/arXiv.2510.14337. URL https://arxiv.org/abs/2510.14337. Presented at the NeurIPS 2025 MTI-LLM Workshop.

Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. Unanswerability evaluation for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8452–8472. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.415. URL https://aclanthology.org/2025.acl-long.415/.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191. Association for Computational Linguistics, 2018. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023/.

Jingxi Qiu, Zeyu Han, and Cheng Huang. SURE-RAG: Sufficiency and uncertainty-aware evidence verification for selective retrieval-augmented generation. *CoRR*, abs/2605.03534, 2026. doi: 10.48550/arXiv.2605.03534. URL https://arxiv.org/abs/2605.03534.

Qwen Team. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950. doi: 10.2307/2087176. URL https://doi.org/10.2307/2087176.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3781–3797. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.249. URL https://aclanthology.org/2022.emnlp-main.249/.

Yohanes Andre Setiawan. Prompt-based abstention fails under misleading context: A controlled study of small frozen RAG models. *CoRR*, abs/2608.22228, 2026. doi: 10.48550/arXiv.2608.22228. URL https://arxiv.org/abs/2608.22228.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951. doi: 10.1111/j.2517-6161.1951.tb00088.x. URL https://doi.org/10.1111/j.2517-6161.1951.tb00088.x.

Linxin Song, Taiwei Shi, and Jieyu Zhao. The hallucination tax of reinforcement finetuning. *arXiv preprint arXiv:2505.13988*, 2025a. URL https://arxiv.org/abs/2505.13988.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=Iyrtb9EJBp.

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. doi: 10.1108/eb026526. URL https://doi.org/10.1108/eb026526.

Fangzheng Tian, Debasis Ganguly, and Craig Macdonald. Predicting retrieval utility and answer quality in retrieval-augmented generation. In *Advances in Information Retrieval: 48th European Conference on Information Retrieval, ECIR 2026*, volume 16483 of *Lecture Notes in Computer Science*, pp. 368–385. Springer, 2026. doi: 10.1007/978-3-032-21289-4_24. URL https://doi.org/10.1007/978-3-032-21289-4_24.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez,

Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/arXiv.2307.09288. URL https://arxiv.org/abs/2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2140–2151. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.188. URL https://aclanthology.org/2021.findings-acl.188/.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen-tau Yih, and Xin Luna Dong. TruthRL: Incentivizing truthful LLMs via reinforcement learning. *arXiv preprint arXiv:2509.25760*, 2025. URL https://arxiv.org/abs/2509.25760.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *CoRR*, abs/2401.15884, 2024. doi: 10.48550/arXiv.2401.15884. URL https://arxiv.org/abs/2401.15884.

Diji Yang, Linda Zeng, Jinmeng Rao, and Yi Zhang. Knowing you don't know: Learning when to continue search in multi-round RAG through self-practicing. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1305–1315. ACM, 2025. doi: 10.1145/3726302.3730018. URL https://doi.org/10.1145/3726302.3730018.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. Crag - comprehensive RAG benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10470–10490. Curran Associates, Inc., 2024. doi: 10.52202/079017-0335. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1435d2d0fca85a84d83ddcb754f58c29-Abstract-Datasets_and_Benchmarks_Track.html.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-Tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.394. URL https://aclanthology.org/2024.naacl-long.394/.

Hengran Zhang, Keping Bi, Jiafeng Guo, Jiaming Zhang, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. LLM-specific utility for retrieval-augmented generation. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management, CIKM 2026*, 2026. URL https://arxiv.org/abs/2510.11358. To appear.

Yibo Zhao, Jiapeng Zhu, Zichen Ding, and Xiang Li. GRACE: Reinforcement learning for grounded response and abstention under contextual evidence. *arXiv preprint arXiv:2601.04525*, 2026. URL https://arxiv.org/abs/2601.04525.

Youchao Zhou, Heyan Huang, Yicheng Liu, Rui Dai, Xinglin Wang, Xingchen Zhang, Shumin Shi, and Yang Deng. Do retrieval augmented language models know when they don't know? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35158–35166. Association for the Advancement of Artificial Intelligence (AAAI), 2026. doi: 10.1609/aaai.v40i41.40822. URL https://doi.org/10.1609/aaai.v40i41.40822.

A ADAPTATION OF COMPARISON METHODS

We adapt the comparison methods in Table 4 as follows. Passage-level scores use the maximum over evidence items; answer-aware methods start from one greedy answer to the full set, with 32,768-token prompts that are not truncated. Adapted local prompts use Qwen2.5-1.5B-Instruct (Qwen Team, 2024), entailment uses mDeBERTa-v3-base (279M; He et al., 2023), and released classifiers retain their own weights.

Sufficient Context (frontier and local). Joren et al. (2025) ask whether a context contains enough information to answer. We use their published autorater instructions on local Qwen2.5-1.5B-Instruct (Qwen Team, 2024) and on GPT-6 Luna through Amazon Bedrock's `us.openai.gpt-6-luna` inference profile (Amazon Web Services, 2026). For the Table 4 frontier row, we fill the released prompt's timestamp with 2026-09-02, cap each passage at 1,600 characters, and make one API call for each of the 74,841 records. We set a 4,096-token completion limit and leave other inference options at their defaults. The parsed JSON sufficiency label is a binary score, so two records with equal labels count as a tie under $\mathrm{AUC}_{\mathrm{wq}}$ (26–57% of pairs across datasets); two responses without a parseable label receive 0.

Yes/no prompt. This control uses the same local Qwen2.5-1.5B-Instruct, but replaces the published Sufficient Context prompt with the short template below and scores token probabilities instead of a generated JSON label. It lists every evidence item in its evaluation input order as `[1]`, `[2]`, ..., cutting each to 800 characters. The text is one user turn in the model's chat template, with the assistant turn opened; the full prompt is not truncated and no answer is decoded. Its score is the last-position log-odds $\log \sum_{t \in Y} e^{\ell_t} - \log \sum_{t \in N} e^{\ell_t}$, where Y and N contain first tokens of *yes* and *no* in upper and lower case, with and without a leading space.

```
Judge whether the documents below contain enough information
for a diligent reader to derive a definitive answer to the
question.  Combining facts across documents is allowed;
guessing beyond the documents is not.
Question:  {q}
Documents:
{docs}
Do the documents contain enough information to answer the
question?  Answer with one word, yes or no.
```

UPR. Sachan et al. (2022) rerank passages by the likelihood of the query given each passage. We use their prompt with the local 1.5B model, take mean $\log P(q \mid p)$ for each passage, and keep the maximum.

RAGAS-CR. The reference-free context-relevance measure of Es et al. (2024) extracts sentences needed to answer. We use its official prompt and score the fraction of context sentences extracted.

CRAG. The retrieval evaluator of Yan et al. (2024) grades a passage as correct, incorrect, or ambiguous. We apply its released T5-large classifier (770M; Raffel et al., 2020) to `query [SEP] passage` and take the maximum correct-class logit.

UAR-tree. Cheng et al. (2024) use a four-head decision tree to decide whether an instruction needs retrieval. We use the released Llama-2-7B backbone (Touvron et al., 2023) and the official DROP template (Dua et al., 2019), reading every passage in its evaluation input order; the score is 1 for "no retrieval". The 4,096-token limit prefix-truncates 13,080 of 17,572 NeoQA records and 309 of 16,000 MuSiQue records, and none elsewhere.

Self-RAG. Asai et al. (2024) use reflection tokens during generation. We use only the released 7B critic, without decoding, and take the maximum passage-wise difference $\mathrm{logit}(\texttt{[Relevant]}) - \mathrm{logit}(\texttt{[Irrelevant]})$.

P(True). Kadavath et al. (2022) judge whether a proposed answer is correct. Given the shared greedy answer, we read the local model's last-position probability of *True* against *False*.

Attrib. Attribution checks whether a source supports a generated claim (Honovich et al., 2022; Gao et al., 2023). We turn the shared answer a into "the answer to q is a" and take the maximum passage-level entailment score.

SemEnt. Semantic entropy groups answers by meaning to measure uncertainty (Kuhn et al., 2023; Farquhar et al., 2024). We cluster the greedy answer and five samples by bidirectional entailment, then use negative cluster entropy.

FLARE. Jiang et al. (2023) trigger retrieval when a forthcoming sentence has low probability. We score the shared greedy answer by its minimum token log-probability, without running another retrieval step.

B Evaluation dataset construction

The six evaluation families in Table 4 use the following group and sampling rules. Paired concordance includes every sufficient–insufficient pair in a group containing both labels.

HotpotQA substitution. We start with HotpotQA (Yang et al., 2018) records that have at least two annotated supporting passages, at least four passages in the original set, and a nonempty question. A pseudorandom generator with seed 17 selects one supporting passage to replace. One distractor is drawn from a different record; a second is drawn from a different record whose passage title shares a token longer than three characters with the removed passage's title. Both insufficient sets retain the other original passages and thus have the same size as the sufficient set. We keep a group only when both distractors can be found, yielding 4,117 groups and 8,234 pairs. Passage order is then shuffled with a fixed seed for evaluation.

MuSiQue substitution. We use the answerable and unanswerable variants supplied by MuSiQue-Full (Trivedi et al., 2022); we do not select replacement passages. The dataset authors assemble a 20-passage context from supporting passages and retrieved distractors, using component questions with intermediate answers hidden for retrieval. For an unanswerable variant, they choose a constituent hop and require its answer to be absent from every context passage. We pair variants sharing the dataset's group identifier and question.

NeoQA deletion. We use the native test variants from the main, optimal-evidence, and context-ablation subsets (Glockner et al., 2025), retaining their *answerable-sufficient* and *answerable-insufficient* labels and excluding native unanswerable questions. We group variants by timeline and question family; no new test deletion is sampled. In optimal-evidence variants, an insufficient set omits a required article from a minimal sufficient coalition. In the main variants, it removes the articles carrying a selected required outline fact, which may be more than one article. Context-ablation variants add nested same-timeline nonsupporting articles around two required articles and omit each required article in turn. We compare every supplied sufficient and insufficient variant within a mixed-label group.

SQuAD 2.0 and IIRC. SQuAD 2.0 (Rajpurkar et al., 2018) groups the dataset's authored answerable and unanswerable questions by source paragraph, keeping that paragraph fixed. IIRC (Ferguson et al., 2020) groups records by source article and retains the 1,327 groups with both native answerability labels. Every answerable question is paired with every unanswerable question in its group, yielding 3,653 pairs. Each IIRC record keeps its own supplied evidence set, so both question and evidence may change within a pair.

Fresh Wikipedia. We use passages from English Wikipedia articles created after July 2026. For each group, `gpt-oss-120b` writes one answerable question and up to two related unanswerable questions about a fixed passage. Unanswerable edits include replacing an entity or adding a premise whose key entity is absent; string and alias checks require the positive answer span to appear and the

substituted or premise entity to be absent. We pool 1,294 mechanically verified candidate groups, then run three rounds of question-text-only filtering, removing the most predictable 8% of groups in each round. The retained set has 1,006 groups, 2,918 records, and 1,912 pairs.

C CONSTRUCTION-CUE MEASURES

Passage count. We score each record by the number of passages in its evidence set, $|E|$.

Minimum IDF coverage. For each passage, we compute the share of the question's IDF weight carried by words in that passage and take the minimum across the evidence set. In 7,966 one-for-one HotpotQA substitutions, the inserted distractor covers a median 4.6% of that weight, versus 44.8% for the support passage it replaces. The distractor is the least-covered passage of the insufficient set in 94.0% of these pairs. Shared added passages can make the two sets' minima equal while the original replacement remains.

Question-only classifier. We predict sufficiency from the question text alone. This classifier and the evidence-only classifier use word unigrams and bigrams hashed into 16,384 dimensions with a four-byte BLAKE2b digest. Features are weighted by $\log(1+\mathrm{tf})$ times training-fold IDF, $\log((n+1)/(\mathrm{df}+1))+1$, and normalized to unit length. We fit an L2-regularized logistic regression with an intercept using 300 full-batch gradient steps (learning rate 0.5, L2 coefficient 1). We assign complete question groups to five folds using seed 17 and score each record with the model trained on the other four folds. Models are fit separately for each dataset.

Evidence-only classifier. We apply the same feature map and training procedure to all passages joined with blank lines in their BM25 order. For the passage-addition analysis, we refit this classifier at each $k \in \{0, 5, 10, 15, 20\}$ after BM25 ordering. The out-of-fold sufficiency probabilities from both classifiers are evaluated with the paired concordance of Section 3.

D RINSE TRAINING AND RUNTIME

Train and test question groups are disjoint. Backbones and preprocessing are shared across datasets; each fold fits the content head, latent direction, and normalization statistics.

RINSE. The RINSE row of Table 4 sums the three size-corrected reader scores with equal weights (Section 4); it reads the question and evidence without generating an answer.

Size correction beyond the training range. Training sets hold at most 22 passages, while NeoQA reaches 120. On held-out validation timelines (3,159 records from 117 question groups), a flat tail raised $\mathrm{AUC}_{\mathrm{wq}}$ from 0.664 to 0.730 over linear extrapolation. The padding curve, specified before test evaluation, scored 0.729, within the preset 0.01 tolerance. We therefore use the linear correction within the training range and the learned curve beyond it. Training the latent reader on the same top five passages it reads at inference then raised validation $\mathrm{AUC}_{\mathrm{wq}}$ to 0.757; together the changes raised NeoQA test performance from 0.673 to 0.773 (paired difference $+0.100$ $_{[+0.091, +0.108]}$).

Timing. We sample 57 records from each of the six Table 4 datasets and time single-record latency on one GPU with one resident model copy. For throughput, independent warmed replicas process the same 342 records concurrently; we double the replica count from one to 16 and report the first trial reaching 90% mean GPU utilization or 90% peak memory use, dividing completed records by the slowest replica's wall-clock time.

E SPAN-READER TRAINING AND OVERLAP AUDIT

We train pretrained RoBERTa-base once on Natural Questions (Kwiatkowski et al., 2019), without earlier QA fine-tuning. Each positive uses an answer-bearing passage; its negative replaces answer

sentences with length-matched topical text, or uses an answer-free passage if replacement is impossible. We discard negatives in which a gold answer string survives, and exact matches to evaluation questions or passages.

The data contain 90,732 training and 16,013 validation pairs. We train for one epoch with AdamW (batch 32, learning rate 2×10^{-5}, 6% warm-up, weight decay 0.01, bfloat16), 384-token inputs, and spans of at most 30 tokens; this takes 12 minutes on one GPU. The score is the best span logit sum minus the no-answer logit sum at `CLS`, maximized over windows and passages. External validation reaches 0.994 paired accuracy and 0.995 ROC–AUC, above the preset thresholds of 0.80 and 0.85.

Passage-overlap audit. We remove exact question and source-passage matches, then check for shared sentences of at least eight words. Passage overlap is absent in NeoQA and fresh Wikipedia, low in SQuAD 2.0 and IIRC, and reaches 29.5% of HotpotQA and 97.8% of MuSiQue question groups (26.6% of MuSiQue passages). These are passage matches, not question matches.

F RINSE AS EVIDENCE SETS GROW

Adding passages from other questions weakens RINSE's paired ranking even when both sides receive the same new material. At $k \in \{0, 5, 10, 15, 20\}$, we give every record in a question group the same first k added passages, order the complete set by BM25 without filtering, and apply the unchanged scorer. Across six datasets, paired concordance falls monotonically from 0.847 to 0.804 ($\Delta = -0.043$ [-0.048, -0.039]). BM25 also reorders the original passages and recomputes reader features at $k = 0$, so Table 7 can differ from Table 4 before any passages are added.

The loss is uneven. IIRC drops from 0.925 to 0.790 ($\Delta = -0.135$), and NeoQA from 0.777 to 0.708 ($\Delta = -0.069$) despite starting with a median of 41 passages. SQuAD 2.0 and fresh Wikipedia lose 0.033 and 0.019, while HotpotQA and MuSiQue change by less than 0.005. On IIRC the max-over-passages span reader falls from 0.904 to 0.686; on fresh Wikipedia the latent reader falls from 0.927 to 0.849. On NeoQA the combined score loses more paired concordance than any reader alone.

The added passages keep their original labels. A screen for answer strings, repeated titles and same-question sufficient passages flags at most 8.3% of insufficient records in paired groups at $k = 20$. Excluding every flagged pair changes no estimate in Table 7 by more than 0.002.

Table 7: Sensitivity to shared passage additions. Values are paired concordance $\mathrm{AUC}_{\mathrm{wq}}$ with 95% group-bootstrap intervals; Δ compares $k = 20$ with $k = 0$.

Dataset	**Pairs**	k=0	k=5	k=10	k=15	k=20	$\mathbf{\Delta AUC_{wq}}$, k=20
SQuAD 2.0	2,803	0.684 [0.66, 0.70]	0.666 [0.65, 0.69]	0.661 [0.64, 0.68]	0.655 [0.63, 0.67]	0.651 [0.63, 0.67]	−0.033 [-0.049, -0.019]
IIRC	3,653	0.925 [0.91, 0.94]	0.860 [0.84, 0.88]	0.823 [0.81, 0.84]	0.801 [0.78, 0.82]	0.790 [0.77, 0.81]	−0.135 [-0.152, -0.118]
NeoQA	102,876	0.777 [0.77, 0.79]	0.745 [0.73, 0.76]	0.731 [0.72, 0.74]	0.718 [0.71, 0.73]	0.708 [0.70, 0.72]	−0.069 [-0.076, -0.062]
HotpotQA	8,234	0.896 [0.89, 0.90]	0.897 [0.89, 0.91]	0.895 [0.89, 0.90]	0.893 [0.88, 0.90]	0.893 [0.88, 0.90]	−0.003 [-0.007, +0.002]
MuSiQue	8,000	0.828 [0.82, 0.84]	0.828 [0.82, 0.84]	0.828 [0.82, 0.84]	0.829 [0.82, 0.84]	0.827 [0.82, 0.83]	−0.001 [-0.006, +0.004]
fresh Wikipedia	1,912	0.973 [0.97, 0.98]	0.970 [0.96, 0.98]	0.963 [0.95, 0.97]	0.962 [0.95, 0.97]	0.954 [0.94, 0.96]	−0.019 [-0.028, -0.010]
Average		0.847 [0.84, 0.85]	0.828 [0.82, 0.83]	0.817 [0.81, 0.82]	0.810 [0.80, 0.82]	0.804 [0.80, 0.81]	−0.043 [-0.048, -0.039]

G EXTENDED RELATED WORK

A RAG system can respond to insufficient evidence at multiple points in its pipeline. Before generation, sufficiency can be assessed from the question and retrieved evidence alone; this is the decision

made by RINSE. Relevance scoring operates at the same stage but asks a different question. During or after generation, retrieval controllers, confidence estimators, and verification methods instead rely on the generator's internal states or outputs. Finally, abstention can be trained directly into the generator.

Evidence sufficiency. Sufficient Context defines the same target as RINSE—whether a diligent reader could answer from the provided context alone—and labels it with an autorater ranging from an 11B entailment model to a frontier LLM, reporting that many instances whose context is annotated as gold are in fact insufficient (Joren et al., 2025). S2G-RAG judges sufficiency together with the remaining gap inside an iterative retrieval loop (Li et al., 2026). RINSE builds on these foundations and addresses two complementary questions: how negative construction and residual cues affect the validity of sufficiency evaluation, and whether sufficiency can be estimated accurately, locally, and at substantially lower cost than a large autorater.

Relevance estimation and passage grading. Relevance is a closely related and more extensively studied problem that asks whether an individual passage is pertinent to a question. UPR ranks by the likelihood of the query given the passage (Sachan et al., 2022), RAGAS elicits a graded judgment from a model (Es et al., 2024), and CRAG grades each retrieved document to decide whether the retrieval must be corrected (Yan et al., 2024). Relevance differs from sufficiency in two respects: a passage can be about the question and still omit the fact that the answer requires, and a per-document grade cannot capture cases in which the evidence becomes sufficient only when multiple passages are considered together.

Generator-side signals. Unlike sufficiency and relevance scorers that operate on the question and evidence alone, another line of work derives its signal from the answer generator's internal states or outputs. Among methods based on internal states, UAR classifies a frozen model's hidden state to decide whether retrieval is needed at all (Cheng et al., 2024), while Probing-RAG probes hidden states to predict answer correctness (Baek et al., 2025); related work separates what a model knows from what retrieval supplies (Zhou et al., 2026; Julka, 2026). These studies motivate RINSE 's latent reader (Section 4), although their prediction targets are tied to a particular generator.

Other methods rely on generated outputs. P(True) asks a model how much it trusts its own answer (Kadavath et al., 2022); semantic entropy compares sampled answers (Kuhn et al., 2023; Farquhar et al., 2024); and attribution methods test whether the evidence entails a generated claim (Honovich et al., 2022; Gao et al., 2023). Adaptive-retrieval controllers act earlier but on the same information: FLARE retrieves again when an upcoming token is improbable (Jiang et al., 2023), Self-RAG trains reflection tokens into the generator (Asai et al., 2024), and later systems predict whether another round will help or whether an answer has become stable (Park et al., 2025; Qiu et al., 2026; Yang et al., 2025; Kieback et al., 2026; Moskvoretskii et al., 2025; Tian et al., 2026; Zhang et al., 2026; Setiawan, 2026). In contrast, RINSE targets whether the supplied evidence is sufficient independently of the downstream generator.

Training abstention into the generator. R-Tuning fine-tunes it to say *I don't know* on questions it cannot answer (Zhang et al., 2024), Trust-Align aligns it for grounded refusal and citation (Song et al., 2025b), and TruthRL and GRACE optimize rewards that separate correct answers from hallucinations and abstentions, GRACE adding an explicit sufficiency term to the reward (Wei et al., 2025; Zhao et al., 2026). The resulting behavior belongs to one model and does not survive further training: reinforcement finetuning for reasoning reduces refusal on unanswerable questions (Song et al., 2025a), and reasoning fine-tuning lowers abstention across twenty datasets (Kirichenko et al., 2025). A generator-independent score is complementary to this line, since it supplies the sufficiency signal such rewards must otherwise construct, and it outlives any particular generator.

Evaluation with constructed negatives. Work on constructed negatives asks whether they test the intended capability. Natural language inference labels are partly predictable from the hypothesis alone (Gururangan et al., 2018; Poliak et al., 2018), models exploit lexical overlap in place of entailment (McCoy et al., 2019), and SQuAD 2.0 replaced automatically generated unanswerable questions with authored near-misses for precisely this reason (Rajpurkar et al., 2018). We apply the same discipline to evidence sufficiency: Section 3 specifies the contrast construction and audits what remains predictable from cues other than sufficiency.