
Evaluation Leakage via Retrieved Context: 53 of 75 Held-Out Passages Leaked Without Detectable Effect

Dmitrii Lozhkin
AVELIN AI · October 26, 2026

Abstract

Retrieval-augmented fine-tuning sets pair each training question with retrieved passages; a split by source passage separates the passages questions were written from, but not those the retriever adds. We audited a set we built and used to fine-tune a small open-weight model: 1,778 Arabic and English passages from 24 public documents, 75 of them evaluation source passages, and four passages per prompt retrieved from the whole corpus. Fifty-three of the 75 appeared in 310 of the 1,469 training prompts, 340 times, three quarters through questions about other documents; reference answers cited 108 of these occurrences. Retraining on prompts from an index without evaluation passages gave 27.1% fully supported answers against 27.9% as built, and a difference in differences (as built minus clean, exposed minus unexposed passages) of 0.0 percentage points (pp; 95% CI -10.8 to $+11.6$): no added effect was detected, but one of about 11 pp cannot be ruled out. Matched inference and a closed-book probe agreed. A 7.4 pp advantage of exposed passages among never-trained models, seen under a lenient judge, vanished under strict review.

1 Introduction

A retrieval-augmented fine-tuning example has three parts: a question written from one source passage, a prompt in which that passage sits among other retrieved passages, and a reference answer grounded in the prompt; such distractors teach a model to ignore irrelevant passages [4]. The usual guard is a split by source passage, which the retriever ignores: if its index is the whole corpus, any evaluation passage can be retrieved into a training prompt.

In open-domain question answering, 60–70% of test answers also occur in training sets, about 30% of test questions have a near-duplicate paraphrase, and models score far lower on questions without overlap [5]. How much contamination appears to matter depends on the overlap measure [1] and on whether a model exploits the data or only memorizes it [6]; rephrased or translated items slip past n-gram filters [2]. Search agents can retrieve test items at evaluation time [3]; here retrieval leaks evaluation text into training inputs, a route we did not find quantified.

We measure the leak in a dataset we had already used for training, trace it to the index, and test its effect by retraining on prompts from a clean index.

2 Setup

Corpus and split. The corpus held 1,778 passages (1,000 Arabic, 778 English) from 24 public documents: 12 long reports of 19–506 passages and 12 short web pages of 2–11 passages, 56 in all. The split drew passages long enough for questions from every document, with a quota growing with the square root of document size: 458 training and 75 evaluation source passages. A larger model wrote up to four questions per passage, 256 for 67 evaluation passages.

Example construction. A multilingual dense retriever ranked all 1,778 passages for each question; the top four formed the prompt, with the source passage replacing the fourth when missing (52% of answerable rows). One row in ten omitted the source page, to teach the model to reply that the material does not cover the question. The larger model wrote cited reference answers; 1,469 training prompts remained after 74 went to validation. Nothing excluded evaluation passages from the index.

Audit. We matched the 75 evaluation identifiers against all 5,876 prompt slots; an evaluation passage is exposed if it filled at least one slot. We noted the document each occurrence’s row was written from and whether its reference answer cited the slot. Twins are passages of which one contains at least 80% of the other’s word 8-grams. An AI reviewer compared each evaluation question with its nearest training questions.

Retrain. We fine-tuned a small open-weight instruction model, loss on reference answers only, two seeds each, on the set as built and on a clean version in which each evaluation passage or twin in a prompt gave way to the question’s next-ranked passage (352 slots in 311 of 1,461 prompts, after dropping 8 written from twins). A second large

model rewrote the reference answers of those 311 rows in both versions alike, so the rewrite alone cannot separate them. All models answered with greedy decoding, open-book from the stored evaluation prompts and closed-book from the question alone; the original trained model and its untrained base were rerun the same way. The primary estimate is a difference in differences: as built minus clean on the 50 exposed passages with evaluation questions, minus the same on the 17 unexposed ones (8 of the 75 passages had no questions).

Review. Answers were scored in two blinded passes of AI review, with a third on disagreements (κ 0.66–0.86 between passes on the retrained and rerun models, open-book). An open-book answer is fully supported with 2 of 2 for support by the passages, citations and answering the question: no unsupported detail, correct sentence-level citations, full coverage. The reviewing model shares the rewriting model’s family; this cancels between the retrained versions, and we compare neither with other models. The same review rescored stored answers of the trained model and of eight never-trained conditions: the same model under three instructions, three other open-weight models (one of them under two instructions) and the reference-answer model. Intervals are 95% percentile bootstraps over evaluation passages within exposure groups; p values come from permuting exposure labels.

3 Results

The index, not the split, decided which passages leaked. One evaluation passage appeared 40 times and 16 once; the 11 most frequent account for 62.1% of the 340 occurrences (Figure 1a). In the 12 long documents, 40 of 62 evaluation passages (64.5%) reached training prompts, close to the 60.9% reached among eligible passages that the split had assigned to neither side. The short web pages did the rest: the retriever put 54 of their 56 passages into training prompts, and because the quota grew with the square root of document size, they supplied 13 of the 75 evaluation passages. Short, general passages are retrieved for many questions, and a size-balanced split oversamples exactly them. Wider prompts expose more: with 2, 4 and 8 passages per prompt, the same retrieval exposes 40, 53 and 63.

Three quarters arrived through other documents; a third were cited by reference answers. Of the 340 occurrences, 260 came from questions about another document and 80 from the same document, so a document split with the same index would have removed at most 80. In 108 occurrences (107 rows) the reference answer cited the evaluation passage, so its content entered the training target; this reached 35 of the 53 exposed passages.

Identifiers missed copies and paraphrases. A reprinted Arabic summary gave 85 identical passage pairs under different identifiers. Of five evaluation passages with a twin, three had it in training prompts, raising exposure by text to 54 of 75, and two had it as a training source passage (8 rows). No evaluation question repeated a training one, but the AI reviewer judged 76 of 256 to ask for the same information as a training question, 56 on exposed passages.

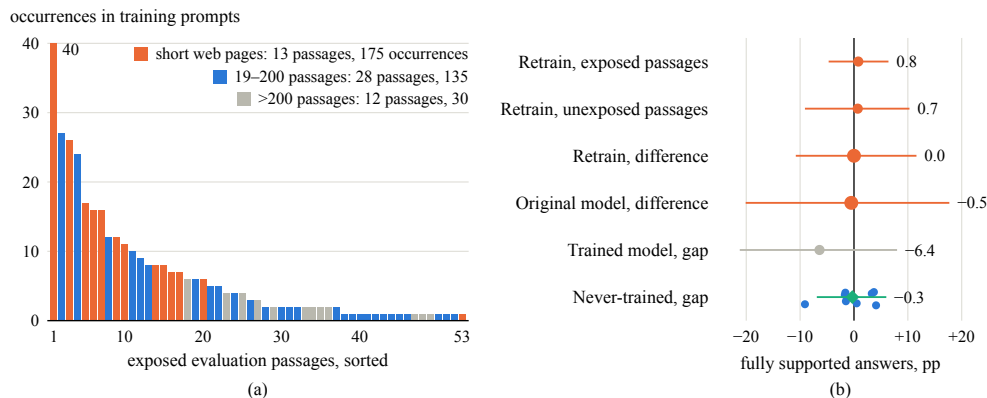


Figure 1: (a) Occurrences in training prompts of the 53 exposed evaluation passages, sorted; the 12 short web pages (orange) hold 7 of the 10 highest counts. (b) Differences in fully supported answers with 95% intervals over passages: matched comparisons (orange), the trained model’s exposure gap (grey), the same gap for eight never-trained conditions (blue); green: their mean.

Retraining with a clean index detected no exposure effect. The as-built and clean versions answered 27.9% and 27.1% of questions fully supported. The as-built version led by 0.8 pp on exposed passages (−4.7 to +6.4) and 0.7 pp on unexposed ones: a difference in differences of 0.0 pp (95% CI −10.8 to +11.6; $p = 0.94$), with −2.5 and +2.6 pp for the two seeds, a spread larger than the exposed-passage difference (Figure 1b). The 13 exposed passages whose slots the clean index refilled from their own page differed no less (+4.0 pp) than the other 37 (−0.4 pp).

The original model and a closed-book probe agree. Under identical decoding, the original trained model led its untrained base by 25.0 pp, as much on exposed as on unexposed passages (difference −0.5 pp, −20.1 to +17.7). Without passages, models stated the source passage’s answer for 4–9% of questions, and the as-built version gained

1.8 pp (−1.4 to +5.1) over the clean one on exposed relative to unexposed passages. The mechanism is plain: 224 of the 256 evaluation prompts held the source passage, so an open-book model reads the answer anyway, and leaked passages entered training mainly as prompt context, on which no loss was computed.

Exposed passages were not detectably easier. The trained model’s stored answers scored 26.5% on exposed and 32.8% on unexposed passages (gap −6.4 pp, −21.2 to +8.0); the eight never-trained conditions ranged from −9.1 to +4.1 pp, mean −0.3 pp (−6.9 to +6.0; $p = 0.94$). Training lifted cited, context-only and unexposed passages alike, by 21 to 27 pp; overall from 4.7% to 28.1%. An earlier single-pass judge, lenient on unsupported details and misplaced citations, gave the same never-trained answers a mean gap of +7.4 pp (+0.7 to +14.6), which we had read as easier exposed passages. A gap across unmatched models can belong to the judge; the matched retrain is the evidence.

4 Limitations

With 17 unexposed passages the interval spans about ± 11 pp, so an effect of up to about 11 pp in either direction, large against a 27% rate, remains possible. The clean index refilled slots with the next-ranked passage, 162 of 352 times from a page holding an evaluation passage and 46 times from the page of the passage replaced, so it measures the index change, not the removal of all related text. The retrained versions replicate the recipe, not the original model: another model rewrote 311 of their 1,461 targets. Absolute rates depend on the judge: strict labels give the trained model 28.1% where the earlier lenient labels gave 77.3%, so the two are not comparable, and under strict labels seven of the eight never-trained conditions score below 22%, which compresses any exposure gap; the untrained base’s low rate also reflects answers 4.6 times longer. Without the 76 paraphrased questions the difference in differences is +2.5 pp (−12.4 to +19.1). One partial cross-language counterpart, a boilerplate footnote, would make 55 of 75 exposed. This is one dataset, retriever, prompt width and base model; per-question records are not released.

5 Conclusions

A split by source passage did not keep evaluation text out of retrieval-augmented fine-tuning: a whole-corpus index carried 53 of 75 evaluation passages into training prompts, mostly via other documents, and reprints and paraphrases added routes that identifiers miss. Retraining on prompts from a clean index gave a difference in differences of 0.0 pp: in this open-book setting no inflation was detected, though a substantial one is not ruled out. An earlier sign of easier exposed passages did not survive a strict judge.

6 Recommendations

1. Build the training-time retrieval index from training passages only, then audit the final prompts. A whole-corpus index put 53 of 75 evaluation passages into 21.1% of our prompts, 63 at eight passages per prompt; a document split alone would have left 260 of 340 occurrences.
2. Deduplicate by text, not by identifier, before splitting: excluding evaluation identifiers alone still leaves 5 evaluation passages in prompts as reprinted twins.
3. Count occurrences per evaluation passage and screen evaluation questions for paraphrases: 12 short web pages, 3.1% of the corpus, held 51.5% of occurrences, and 76 of 256 evaluation questions paraphrased training ones.
4. Measure a leak’s effect with a matched retrain under one strict judge, not with an exposed-versus-unexposed gap across models, which here was +7.4 pp under a lenient judge and −0.3 pp under a strict one. Two seeds per version, 7.6 GPU-hours of training, left an interval of about ± 11 pp; a tighter bound needs more evaluation passages.

Disclosure. Dataset, training, evaluation and analysis were produced with AI coding agents under the author’s direction. All numbers were recomputed from per-question records and frozen review labels.

References

- [1] A. K. Singh, M. Y. Kocyigit, A. Poulton, et al. Evaluation data contamination in LLMs: How do we measure it and (when) does it matter? arXiv:2411.03923, 2024.
- [2] S. Yang, W.-L. Chiang, L. Zheng, et al. Rethinking benchmark and contamination for language models with rephrased samples. arXiv:2311.04850, 2023.
- [3] Z. Han, M. Mankikar, J. Michael, and Z. Wang. Search-time data contamination. arXiv:2508.13180, 2025.
- [4] T. Zhang, S. G. Patil, N. Jain, et al. RAFT: Adapting language model to domain specific RAG. arXiv:2403.10131, 2024.
- [5] P. Lewis, P. Stenetorp, and S. Riedel. Question and answer test-train overlap in open-domain question answering datasets. In EACL, 2021. arXiv:2008.02637.
- [6] I. Magar and R. Schwartz. Data contamination: From memorization to exploitation. In ACL, 2022. arXiv:2203.08242.