PURETENSOR
INTELLIGENCE DISTILLED

Where 45 Retrieval Misses Were Lost

A Per-Stage Trace and the One Fix That Passed a Rule Set in Advance

Paper: PT-R-2026-064 v1.0

Author: PureTensor Ltd

Date: 2 November 2026

Status: Published

Abstract

Our agents search a store of Markdown notes before they act. The search is a hybrid pipeline: a lexical leg, a dense-vector leg, reciprocal-rank fusion of the two, a cross-encoder reranker, a provenance blend, and a per-file cap. When it misses an answer we know is there, at which stage was the answer lost, and was the miss a retrieval defect at all, or a defect in the evaluation that called it a miss?

We rebuilt the search as a traced, stage-by-stage instrument that reproduced the live function exactly on 295 of 295 diagnosis queries, and used it to follow 45 misses (27 on a golden question-to-file set, 18 on a hand-built probe set). The losses are spread across every stage, from 11 answers never retrieved by either leg to 1 dropped by a filter. Thirteen of the 45 were near-misses. In 7 of the 18 probe misses the right file came back at file rank 1 or 2 but the chunk-level scorer wanted a different chunk; we did not check whether the returned chunk held the answer. Our hand classification judged a majority of the 45 to be evaluation defects; its labels were not retained, so it is judgment only.

We tested three candidate fixes against a seven-part adoption rule written before the first run. One passed all seven: letting the reranker score up to three windows of a long passage instead of only its first 800 characters. On the untouched half of the golden set, Recall@5 rose from 0.8750 to 0.8922. On all 496 evaluable golden queries in a quiet pair of runs (a set that includes the half used for diagnosis), it rose from 0.8627 to 0.8841 for a cost of 28 ms at p95. Two other fixes failed the rule, mainly on its guards against stale answers. No single population shows a significant gain on its own. The effect is positive in sign in every population and significant in none, at a modest latency cost.

1. The Question After the Repair

As we reported in an earlier post on our memory substrate, fixing the questions in an early evaluation took Recall@5 from 0.073 to 0.629 without touching the retriever. On the repaired evaluation, some questions still miss. This study asks what that residue is made of.

A retrieval defect is something to engineer away. An evaluation defect (a golden list that names one file where several files answer, a retired target, a scorer that wants a specific chunk) is something to correct in the test, since engineering against it tunes the system to a flawed ruler.

We also wanted to know whether any generic change would win under a rule that could not be bent after the fact.

The system under test is a Postgres index of 14,285 chunks (the count recorded when the independent-set runs were taken), searched by the pipeline described above. Fusion keeps limit times three candidates, which is 30 for a limit of ten. The reranker is configured to rescore the top 50 but can only rescore what fusion passes on, so it sees at most 30, and it is given only the first 800 characters of each passage. The lexical leg normally requires every query term to match (an AND query) and falls back to matching any term (OR) only when the AND form returns nothing useful and the query is plain prose. A hyphen, slash, or a dot or colon between alphanumerics makes a query not plain prose.

2. A Traced Copy of the Search, and Four Question Sets

The trace instrument is a script that calls the real search function with in-process hooks, leaving the engine unedited. For each target chunk it records the rank in each leg (with an unbounded re-run, so that "absent" means absent and not just beyond the leg's depth), the fused rank, the rerank rank, its rank after each later stage, and the character offset of the first supporting text. Traced and plain search results were equal on all 295 diagnosis queries (292 needed one attempt, 3 needed two), with no degraded leg.

We used four populations. The golden set has 540 question-to-file pairs, of which 496 were evaluable (the evaluator dropped 44). It is scored at file level. We split it by hash into a diagnosis half of 264 and a measurement half of 232. A hand-built set of 57 paired probe questions (a current fact against a stale or opposite one; a companion study owns its design) was split into 31 for diagnosis and 26 held out, and is scored at chunk level. An incident replay set of 76 questions (42 held out) scores whether the top three hits are current rather than superseded (current@3) and whether a stale answer ranks first (stale-first). A parallel workstream built 141 further probe questions, used once, at the end, as an independent check.

Causes were sought on the diagnosis populations only, 264 golden plus 31 probe, which is the 295 above. The comparisons in Sections 5 and 6 use the untouched halves and the independent set; the quiet pair in Section 7 covers all 496 golden queries.

3. Where 45 Misses Were Lost

Of the 295 diagnosis queries, 250 hit and 45 missed (a rank worse than 5): 27 golden misses (10.2% of 264) and 18 probe misses (58.1% of 31). The probe set was built to be hard.

Population	Misses	Final rank 6 to 10	Not in top 10, right file present as another chunk	Not in top 10, nothing of the file
Golden (264)	27	9	0	18
Probe (31)	18	4	7	7
Total (295)	45	13	7	25

Table 1. Shape of the 45 misses by population. The third column is nonzero only for the chunk-level probe scorer.

Thirteen of the 45 were near-misses at rank 6 to 10. In 7 of the 18 probe misses the right file came back at file rank 1 or 2, but a different chunk of the same file was returned, and the chunk-level scorer, which wants the specific supporting chunk, counted it as a failure. This is a possible evaluation defect: the system found the document and the scorer asked a narrower

question than the user would. We did not check whether the returned chunk carried the answer, and four of the seven carry a fusion-cut or rerank-demotion label in Table 2, so they are at least partly genuine chunk-level misses.

The trace also labels the stage at which each answer was lost.

Loss stage	Golden	Probe	Total
Not retrieved by either leg	6	5	11
Fused beyond rank 5, rerank kept it beyond 5	8	3	11
Cut at the 30-candidate fusion limit	4	5	9
Rerank demoted it out of the top 5	3	5	8
Provenance blend demoted it out of the top 5	5	0	5
Dropped by a source-type filter	1	0	1
Total	27	18	45

Table 2. Loss stage assigned by the trace. Variants that also involve the wrong chunk are merged.

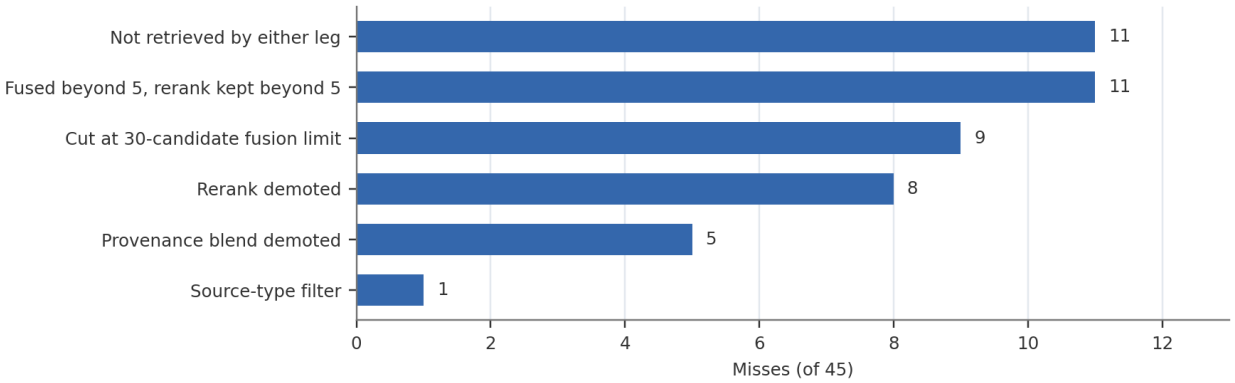


Figure 1. Where the 45 misses were lost, by pipeline stage (values in Table 2).

No stage accounts for more than 11 of the 45 labels, so no single-stage fix addresses more than a quarter of them (Figure 1).

The lexical gate. Of the 295 diagnosis queries, 106 contain a hyphen, a slash, or a dot or colon between alphanumerics, and so fail the plain-prose test. Every one of those 106 took the AND form and never the OR fallback. Of the other 189, 152 took the OR fallback; the gate is a block, not a guarantee, since the remaining 37 also stayed on AND. Twenty-two of the 45 misses are gated queries, a miss rate of 20.8% (22 of 106) against 12.2% (23 of 189) for the rest.

Gated queries contain identifiers (a hyphenated node name, a dotted version string) and may be harder anyway. We do not claim the gate caused those 22 misses; the fix below recovered one query on its measurement half.

Answers beyond the reranker's window. In the probe set, 12 of the 57 questions have their supporting sentence beyond character 800 of its chunk. For 4 of the 18 probe misses the supporting text starts beyond character 800 of the truth chunk (at offsets of 971 to 1,452 characters, in chunks of 1,360 to 2,048), and in a second chunk of a fifth. Golden misses have no such offset, since that set records files and not sentences. A cross-encoder cannot score text it was not shown, so these answers can reach the reranker's head and still be scored on a passage that omits them.

4. Real Misses and Evaluation Defects

We also classified the 45 by hand, with two adversarial critics reviewing the labels, into retrieval defects and evaluation defects. The evaluation-side families were golden lists naming one file where several answer, stale or unanswerable golden targets, and the probe scorer rejecting the right file when a sibling chunk carries the answer. The retrieval side held answers past the 800-character window, candidates lost to the lexical gate, and candidates lost at the fusion cut.

We did not retain the per-miss labels or the critics' notes. The judgment was that a majority of the 45 were evaluation defects. It is the weakest claim in the paper: our reading of the cases, produced by the team that wrote the fixes, and not re-derivable from the artefacts we kept. The retained trace supports only the 7 file-found, chunk-not-returned probe cases (of uncertain status) and the 13 near-misses. The golden set has not yet been repaired, so its absolute figures are a floor.

5. Three Fixes and a Rule Written First

We implemented three changes behind default-off flags; with the flags off, results are byte-identical to the base.

The first, F1, gives the reranker up to three windows of any passage longer than 800 characters. Window 0 is the first 800 characters. Later windows overlap the previous one by 150 characters, the window with the densest overlap with distinctive query terms is preferred, and the maximum score per chunk is kept. The second, F2, stops punctuation from disabling the lexical OR fallback. The third, F3, fills the fusion to the rerank depth, fusing to 50 candidates instead of 30.

The adoption rule was coded into the A/B script before the first arm ran. The script was last modified at 22:02:48 UTC and the first arm's output is stamped 22:03:52; no external registration exists, so the ordering rests on file timestamps. A fix is adopted only if all seven hold on the measurement halves against the base arm: golden Recall@5 wins exceed losses; no significant golden loss (exact sign test, p below 0.05); incident current@3 not lower; incident stale-first not higher; probe truth@5 not lower; probe lure-first (the rate at which the stale or opposite item ranks first) not higher; and golden p95 latency cost at most 150 ms. We evaluated F1, F2 and F3 singly, then F2 with F3, then all together, back to back on the same index state.

Metric	Base	F1	F2	F3	F2+F3	All
Golden R@5	0.8750	0.8922	0.8793	0.8621	0.8664	0.8836
Golden R@10	0.8966	0.9138	0.9009	0.8966	0.9009	0.9095
Golden MRR	0.7521	0.7756	0.7524	0.7397	0.7450	0.7669
Golden wins / losses vs base	-	7 / 3	1 / 0	1 / 4	2 / 4	8 / 6
Golden sign-test p	-	0.34	1.00	0.38	0.69	0.79
Probe truth@5	0.8182	0.8696	0.7391	0.7727	0.7727	0.8750
Probe lure-first	0.1818	0.0000	0.1818	0.0909	0.0909	0.0000
Incident current@3	0.8553	0.8816	0.8553	0.8289	0.8289	0.8947
Incident stale-first	0.0395	0.0395	0.0658	0.0526	0.0658	0.0658
Golden p95 (ms, loaded host)	324	355	351	373	388	393
Rule verdict	-	Adopt	Reject	Reject	Reject	Reject

Table 3. Each arm against the pre-set rule on the measurement halves (232 golden, 26 held-out probe, 76 incident). Single run per arm; latency is from a loaded host.

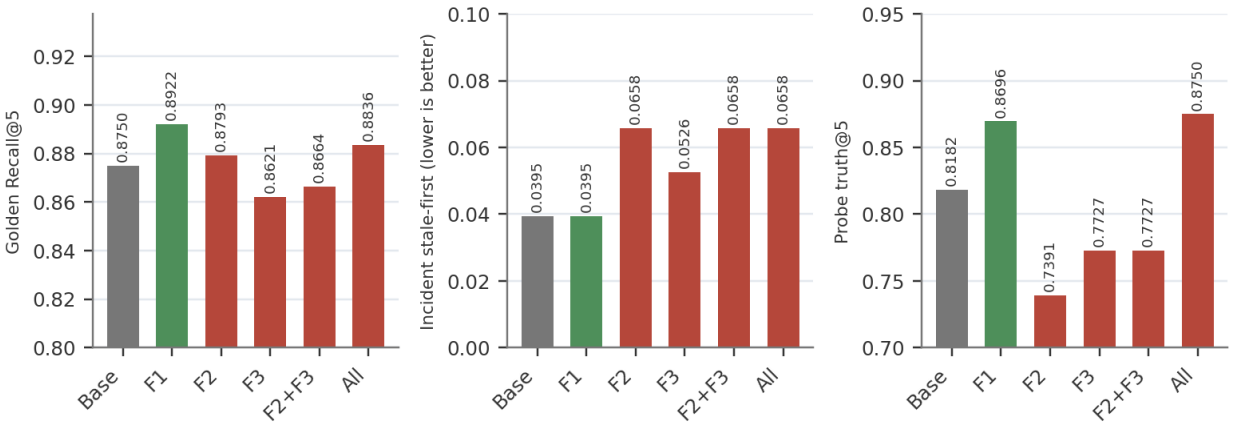


Figure 2. The three guarded metrics per arm: golden Recall@5, incident stale-first, and probe truth@5 (values in Table 3). Only F1 (green) passed the rule. The Recall@5 and truth@5 axes do not start at zero.

F1 passed all seven criteria. In raw counts, 203 of 232 golden queries hit in the top 5 for the base and 207 for F1 (7 gained, 3 lost); incident current@3 went from 65 to 67 of 76, and probe lure-first from 2 of 11 lure-bearing items to 0. Probe truth@5 is computed over items where the right file was returned: 18 of 22 for the base and 20 of 23 for F1.

F2 failed twice: incident stale-first rose from 0.0395 to 0.0658 (3 and 5 of 76 questions), and probe truth@5 fell by 7.9 points. Widening the OR leg pulls in stale chunks sharing an identifier. It gained one golden query. F3 failed four of seven: golden losses outnumbered wins, 4 to 1; current@3 fell from 0.8553 to 0.8289; stale-first rose; and probe truth@5 fell. F2 with F3 failed the same four.

The all-fixes arm passed six of seven criteria and posted the highest incident current@3, 0.8947, but failed on stale-first; the stale-answer guards did the rejecting.

6. Confirming the Adopted Fix

A single run per arm shows little, so we repeated base against F1 three times, interleaved, over the whole battery. Quality was deterministic: F1 gave identical figures on every repeat, and so did the base except for one golden MRR of 0.7514 instead of 0.7521, a rank difference below the top 5. No golden query flipped between repeats within an arm.

Metric	Base (3 repeats)	F1 (3 repeats)
Golden R@5	0.8750 x3	0.8922 x3
Golden R@10	0.8966 x3	0.9138 x3
Golden MRR	0.7521 / 0.7521 / 0.7514	0.7756 x3
Probe truth@5 (over items where the file was returned; 26 held-out)	0.8182 x3	0.8696 x3
Probe lure-first	0.1818 x3	0 x3
Incident current@3	0.8553 x3	0.8816 x3
Incident stale-first	0.0395 x3	0.0395 x3
Incident held-out current@3	0.7857 x3	0.8333 x3
Golden p95 ms	281.5 / 282.1 / 335.3	307.7 / 353.9 / 305.8

Table 4. Three interleaved repeats of the battery. The host was shared, so latency is noisy; mean p95 was 299.6 ms for the base and 322.5 ms for F1.

For variance across questions we used paired item flips (majority vote over the three repeats) with exact two-sided sign tests. On golden hit@5, 7 queries were gained and 3 lost (p = 0.34). On probe truth@5, 4 gained and 2 lost (p = 0.69). On incident current@3, 2 gained and 0 lost (p = 0.50). Pooled over those, 13 gained and 5 lost (p = 0.096).

On the independent set of 141 questions, run once per arm, truth@5 over the 105 questions where the right file was returned in both arms moved from 0.695 to 0.724 (5 gained, 2 lost, p = 0.45). Over all 141, hits went from 75 to 78 (6 gained, 3 lost), and lure-first from 43 to 41, nearly flat. Across the 57 probe questions plus these 141 (198, including the 31 diagnosis probes), hits went from 106 to 115, with 14 gained and 5 lost (p = 0.064).

Pooling all four populations with the independent set in its restricted 105-question form gives 18 gained against 7 lost (p = 0.043); counting all 141 gives 19 against 8 (p = 0.052). The pool mixes majority-vote flips with a single run, and its side of 0.05 depends on a denominator choice made after the fact, so we do not claim it as significant.

7. The Quiet Full-Set Pair

With the host quiet, we ran all 496 evaluable golden queries with the flag off, then on, with no search errors and the reranker applied on every query.

Metric	F1 off	F1 on	Change
Recall@5	0.8627 (0.833 to 0.894)	0.8841 (0.858 to 0.911)	+0.0214
Recall@10	0.8945	0.9106	+0.0161
MRR	0.7625 (0.732 to 0.796)	0.7796 (0.749 to 0.810)	+0.0171
nDCG@5	0.7786 (0.750 to 0.811)	0.7983 (0.770 to 0.828)	+0.0197
p50 latency (ms)	234	251	+17
p95 latency (ms)	282	310	+28

Table 5. All 496 evaluable golden queries, flag off then on, with 95% intervals as emitted by the evaluator.

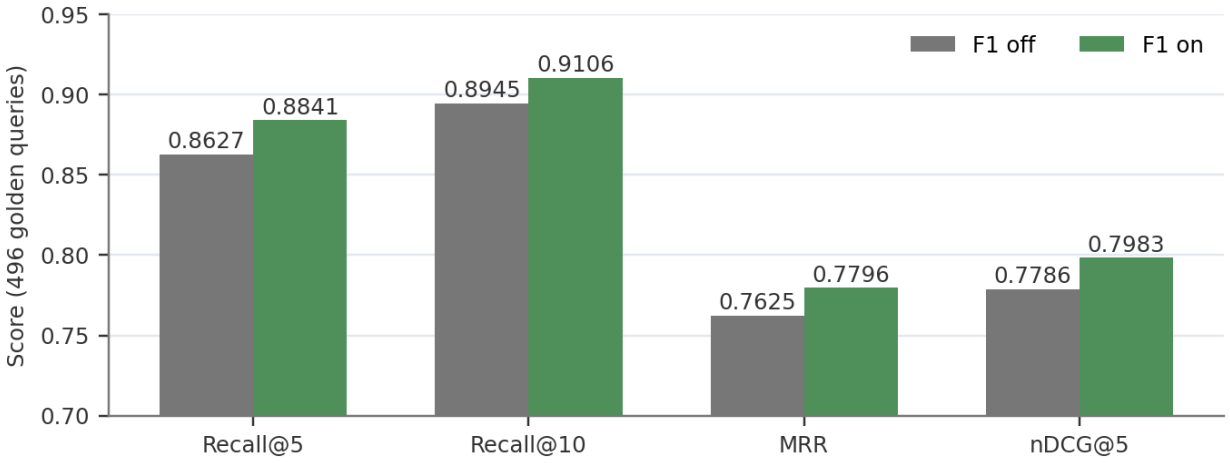


Figure 3. Quality metrics on the 496-query quiet pair, flag off against flag on (values in Table 5). The axis does not start at zero. Latency is in the table.

The intervals overlap, so the claim is the paired direction. This pair includes the 264-query diagnosis half, where the causes were found and F1 was designed, so it is not held out. The

held-out figure is the 232-query half: Recall@5 from 0.8750 to 0.8922 (+0.0172, 7 gained, 3 lost, p = 0.34). Only the quiet pair supports a latency claim; a loaded run minutes earlier gave identical quality and a p95 of 351 ms.

8. What These Numbers Will Not Carry

The fix arms on the measurement halves are single runs, and the three-repeat confirmation shows determinism, not variance across data. Samples are small where it matters: 26 held-out probe questions and 76 incident questions. One query moves golden Recall@5 by 0.43 points and incident current@3 by 1.3 points.

No population is significant alone (golden p = 0.34, probe 0.69, incident 0.50), and the pooled 18 to 7 (p = 0.043) depends on pooling heterogeneous sets. The rule's "no significant loss" test has little power on a handful of discordant pairs; the stale-answer and truth@5 guards are what rejected F2 and F3.

Arm identity for the independent set and the quiet pair is not recorded in the artefacts: both independent-set run records carry a default label, and the quiet pair's flag state appears only in file names. We inferred which run was which from run order and recomputation, and the 0.043 and 0.064 results rest on that inference. The golden scorer, which matches on path suffix, is the artefact we criticise in Section 4, so gains measured against it carry its bias; the independent probe set uses a different scorer and showed the same sign. F1 picks windows by overlap with query terms, so an answer window containing none is never chosen and some past-800 misses stay unrescued.

Three things survive: the F1 effect is positive in sign in every population, deterministic, and judged by a rule fixed beforehand; the stage trace is exact against the live function; and every query that failed the plain-prose test went down the AND-only path. The proportion of misses that were evaluation defects does not survive in checkable form, and we offer it as judgment, not measurement.

An open question is which of the 22 gated misses are recoverable; a stale-aware guard on the OR leg is untested.

9. What We Changed

F1 alone was shipped, first behind its flag and then enabled by default the same evening, with a kill switch that restores the old passage texts byte for byte. F2 and F3 were not deployed. The next piece of work is on the ruler: repairing the golden relevant-file lists and stale targets, and letting the probe scorer accept a correct file when a sibling chunk carries the answer.

10. Methods and Artefacts

All measurement ran in process against the production index, with the embedder (Qwen3-Embedding-4B) and the cross-encoder reranker (bge-reranker-v2-m3) served from GPU seats on the fleet. The store is PostgreSQL with pgvector. Statistics are exact two-sided sign tests on discordant pairs, with majority vote over three repeats for item flips. Measurements ran on 2 November 2026, from about 21:20 UTC to 22:57 UTC. We retain internally the per-query trace of all 295 diagnosis queries, the frozen sets with their split labels, per-arm outputs for the single-run and three-repeat A/B, the quiet pair, the independent-set outputs, the scripts, and the F2 and F3 patches; we do not retain the per-miss labels.

PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. This paper is

PT-R-2026-064. All measurements were taken on 2 November 2026; raw artefacts other than the per-miss classification labels are retained internally.