Access, Not Capability: What Retrieval Buys Small Local Models on Private-Corpus QA

Patrick Keough
pskeough@gmail.com

Robert H. Tai
Australian Catholic University

Draft of July 23, 2026

Abstract

Organizations holding private document collections increasingly seek question answering over those collections with small locally hosted models, a preference grounded in privacy, cost, and access. Confronting the resulting three-way fork, the practitioner must choose: retrieve over the documents, fine-tune on them, or pay for a frontier API. Our answer is that what a frontier API sells on a private corpus is mostly access to documents the organization already owns, not model capability. Denied the documents, frontier models fail to beat even a 1 billion parameter local model with retrieval in over 90 percent of comparisons. Given identical retrieved context, a grounded judge cannot separate the local 8 billion parameter model from the frontier in roughly three of four comparisons, and the residual frontier edge remains small, bounded (net win-rate 0.600), and no larger against a strong recent 3.8 billion parameter model than against the 8 billion one. Handing the frontier the entire corpus in context changes nothing: in a 90-question-per-model pilot, a whole-corpus frontier gains no larger an edge (net win-rate 0.561 [0.508, 0.614]) than the same frontier earns from retrieval alone, and retrieval beats full context even for the frontier itself (0.433 [0.392, 0.475], the interval entirely below parity). The evidence utilizes four configurations crossed with three local model sizes and three corpora, with every comparison judged grounded and in both presentation orders under family-wise multiple-comparison control and equivalence testing, against three frontier models (GPT-5.1, Claude-Sonnet-4.6, DeepSeek-v4-pro) under byte-identical retrieval. Three supporting findings shape the recommendation. Retrieval stands out as the only lever that helps everywhere, surviving Holm correction in 18 of 18 tests, while LoRA fine-tuning yields small corpus-independent gains and fine-tuning on top of retrieval rarely adds value. Corpus homogeneity, not corpus size, predicts retrieval's value. And retrieval flips abstention behavior with corpus structure, raising correct abstention on unanswerable questions from 62 to 100 percent on the homogeneous corpus while collapsing it from 63 to 17 percent on the heterogeneous one. The recommendation follows: buy retrieval and roughly 4 to 8 billion parameters of local capacity rather than cloud access, measure corpus homogeneity before selecting an architecture, and add an abstention guard on heterogeneous corpora.

1 Introduction

A school district holds twenty years of curriculum documents, a clinic its treatment protocols, a small laboratory the papers its grant rests on, and each wants its staff to ask questions of those documents without sending them to a third party. Facing this build-versus-buy decision, the practitioner finds three paths: attaching a retrieval system to a small local model, fine-tuning that model on the documents, or paying a frontier API per query. Deployment of these systems has outpaced the evidence for choosing among them, which remains anecdotal, single-corpus, or

judged without controls for the position, verbosity, and evidence-window artifacts that pairwise LLM judging is known to carry.

This paper supplies a controlled answer: buy retrieval, not cloud access. What a frontier API sells on a private corpus is mostly access to documents the organization already owns; equalizing that access leaves a capability premium that is real but small, bounded, and unchanged by every stress test we ran. Crossing four configurations (base, retrieval, LoRA fine-tuning, and their combination) with three local model sizes and three corpora, one private and two public, we rank the levers under family-wise error control, test the mechanism that predicts when retrieval pays, and decompose the frontier advantage into a component owed to document access and a component owed to model capability. Figure 1 previews the decomposition. Closed-book, the frontier fails to beat the 8 billion parameter local model with retrieval in 95.8 percent of panel-judged comparisons [F05]; handed the same retrieved context, it ties that model in roughly 77 percent of pairs under a grounded judge [F06]; handed the entire corpus in a million-token window (a pilot arm, Section 4.6), it recovers no larger an edge than retrieval already gives it [F22].

One decomposition organizes the paper: the frontier's advantage on a private corpus splits into a component owed to document access and a component owed to model capability, and access is nearly all of it. Surviving a chain of sensitivity analyses spanning judging regime, panel aggregation rule, retrieval success, and corpus, the claim holds at all three local sizes we test and holds when the frontier is handed the entire corpus in its context window [F05, F06, F20, F22]. Three supporting findings then shape the practitioner recipe. Retrieval stands out as the only lever that helps on every corpus and every model, surviving global Holm correction in 18 of 18 tests, while LoRA fine-tuning provides small gains and the RAFT recipe of fine-tuning on top of retrieval adds value in only 2 of 9 cells [F01]. An exact decomposition of the retrieval effect shows corpus homogeneity, not corpus size, deciding whether off-target retrieval still helps [F03]. And retrieval flips abstention behavior with corpus structure, a behavioral hazard with direct deployment consequences [F04].

A tie in this paper means a grounded judging panel could not distinguish the two answers, never that the answers were equal. The frontier edge is real, we quantify it, and the question-level sign test against the 8 billion parameter model reaches p = 2.7e-19 [F06]. Every number in this paper recomputes from raw model outputs through an independent receipt script, and the claim-to-script-to-receipt map ships as the artifact; the grey bracketed tags throughout the draft name the receipts.

Our contributions are fourfold:

1. An access-versus-capability decomposition of the frontier advantage on private corpora, validated through sensitivity analyses over judging regime, aggregation rule, retrieval success, corpus, local size, and a whole-corpus long-context arm [F05, F06, F07, F11, F20, F22].

2. A family-wise-error-controlled ranking of the three practitioner levers, replicated across domain and an approximately 33 fold corpus-size difference (raw chunk basis), with TOST-certified nulls and judge-free ROUGE triangulation, directly testing the field's synergy claims (RAFT survives in 2 of 9 cells) [F01, F09].

3. A measured mechanism, an exact additive decomposition of the retrieval effect joined to an independent homogeneity metric, showing homogeneity rather than corpus size predicts whether off-target retrieval still helps [F03].

4. A retrieval-abstention hazard whose sign differs between our homogeneous and heterogeneous corpora (Section 7), established by a scripted behavioral classifier [F04, F12].

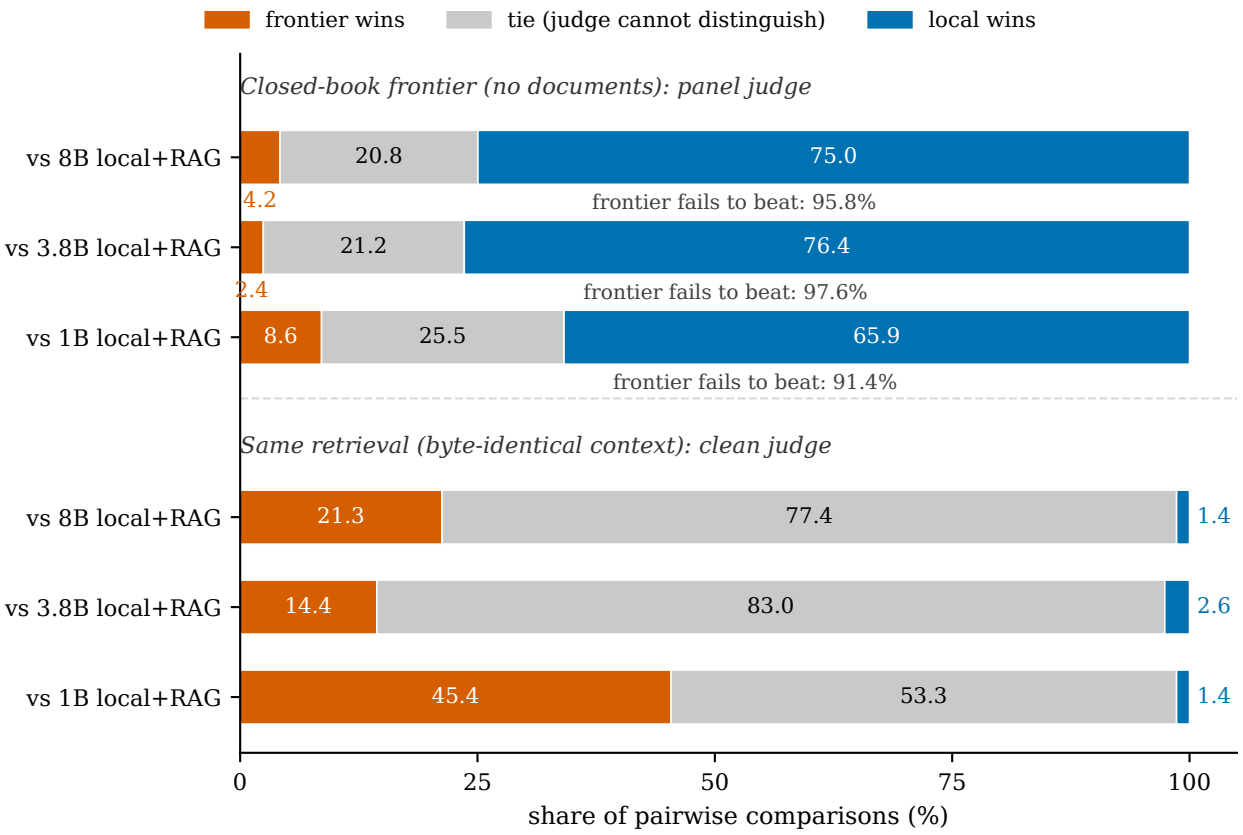

Figure 1: The access/capability decomposition as three-outcome records, at all three local sizes. Top, closed-book frontier models (GPT-5.1, Claude-Sonnet-4.6, DeepSeek-v4-pro pooled) against local models with retrieval, judged by a three-judge panel (majority of decisive verdicts, quorum of two); the frontier fails to beat any size in 91 to 98 percent of comparisons (1B and 8B rows n = 595 on the preregistered 200-question subsample; the 3.8B row n = 1,025 on the full question set). Bottom, the same frontier models given byte-identical retrieved context, scored by the clean judging protocol (n = 1,022 to 1,035 per row); ties dominate at 77.4 percent against the 8 billion parameter model, and local decisive wins fall to 1.4 percent. Win, tie, and loss shares are printed on each segment; proportions are shown without intervals here, and the clustered intervals for the same contrasts appear in Table 2 and Figure 2. [F05, F06, F20]

2 Related work

2.1 Small models with retrieval against large models

Atlas established that retrieval-augmented small models can rival much larger closed-book models on knowledge tasks (Izacard et al., 2023), and Mallen et al. (2023) mapped when parametric memory suffices against when retrieval becomes necessary, organized by entity popularity. Our delta is threefold, frontier-era contestants rather than pre-2023 baselines, a byte-identical retrieval control that gives the frontier the same evidence the local model saw, and a private corpus evaluated with panel judging rather than span-overlap metrics. A parallel line asks whether long context windows subsume retrieval outright, with position effects degrading long-context use (Liu et al., 2024), effective context lengths falling well below advertised windows (Hsieh et al., 2024), and direct comparisons on public benchmarks finding long context often stronger for capable models while retrieval stays far cheaper (Li et al., 2024). We run that contrast directly on a private corpus that fills a million-token window and find the opposite at this scale, retrieval beats full context even for the frontier models themselves (Section 4.6).

2.2 Retrieval against fine-tuning for knowledge injection

Ovadia et al. (2023), Balaguer et al. (2024), and Soudani et al. (2024) compared retrieval and fine-tuning for knowledge injection, generally finding retrieval ahead, while RAFT (Zhang et al., 2024) argued the combination outperforms either alone. Most of this evidence rests on single corpora without multiplicity control. We add family-wise error control over the full 36-test family, TOST equivalence certification of the near-null cells, and cross-corpus interaction tests, and we test the RAFT synergy claim directly, finding it in 2 of 9 cells [F01].

2.3 Irrelevant context and abstention

Shi et al. (2023) showed irrelevant context degrades reasoning, Cuconasu et al. (2024) and Yoran et al. (2024) studied retrieval noise in RAG pipelines, and the selective-QA line (Rajpurkar et al., 2018; Kamath et al., 2020) formalized answering versus abstaining. Our delta is a corpus-level moderator, we show the harm and the benefit of retrieved context on abstention are organized by a measurable corpus property, homogeneity, and that the behavioral effect flips sign across corpora [F04, F12].

2.4 Validity of LLM-as-judge

Zheng et al. (2023) introduced pairwise LLM judging along with its position and verbosity biases, Dubois et al. (2024) debiased length effects by design, and Panickssery et al. (2024) documented self-preference. Building on this line, we run every comparison in both presentation orders with strict-agreement consolidation, a design that measurably contains a judge carrying a 73 percent primacy preference, and we quantify two further artifacts, an evidence-window effect and a verbosity reward, before correcting both by protocol [F10, F15].

3 Study design

3.1 Corpora

Table 1 summarizes the three corpora. The private corpus comprises 28 published papers from a single research group's collection (13 carry OpenAlex work identifiers), divided into 593 chunks

Table 1: The three corpora. OOC counts unanswerable (out-of-corpus) questions; unique counts deduplicated question texts; homogeneity is reported as TF-IDF centroid similarity and mean pairwise cosine. Public-corpus chunk counts are the raw retriever-store counts re-derived in the validation campaign; public token counts were not re-derived and are omitted rather than approximated. [F03, F13, F17]

corpus	gold	eval n (uniq.)	OOC (uniq.)	chunks / tokens	recall@6	homogeneity
Private	LLM (Qwen-2.5-14B)	345 (315)	52 (27)	593 / 416,698	0.294	0.235 / 0.053
COVID-QA	human	400 (398)	0	8,260 / –	0.813	0.136 / 0.018
QASPER	human	400 (381)	98	19,817 / –	0.232	0.140 / 0.019

totaling 416,698 tokens; no public QA set or gold answers exist for it, so gold answers were generated by Qwen-2.5-14B and audited, a provenance we return to in Section 8 [F17]. QASPER (Dasigi et al., 2021) and COVID-QA (Möller et al., 2020) carry human gold. Because the private corpus consists of published papers, we make no contamination-immunity claim by construction; the closed-book floor of Section 4 serves as the empirical non-memorization probe. Deduplication is disclosed for all three evaluation sets, 315 of 345 unique on the private corpus (the 52 out-of-corpus questions contain 27 unique texts), 381 of 400 on QASPER, and 398 of 400 on COVID-QA [F13]. One possible duplicate source paper ("Planning Early" appearing twice) is flagged for a camera-ready diff [F17].

3.2 Models and configurations

Three local models span the consumer-hardware range, Llama-3.1-8B, Phi-4-mini (3.8 billion parameters), and MiniCPM (1 billion parameters). Each runs in four configurations, `base`, `rag` (top-6 retrieval-augmented generation (Lewis et al., 2020)), `lora`, and `lora_rag`. The retrieval stack is one fixed design choice held constant across every configuration: hybrid BM25 plus dense retrieval (Nomic Embed Text v1.5) fused by reciprocal rank, a BGE cross-encoder rerank, and the top 6 chunks served identically wherever retrieval appears. On the private corpus it operates over 1,024-token chunks (the 593 chunks of Table 1). Its parameters were locked by one retrieval sweep before the final runs and never revisited; Section 9 names the stack among the study's fixed hyperparameters. The fine-tuning recipe, QLoRA (Dettmers et al., 2023; Hu et al., 2022) with continued pretraining followed by supervised fine-tuning, was frozen once and applied unchanged to every model and corpus, so cross-corpus differences cannot be attributed to recipe tuning.

3.3 Judging protocol

All quality comparisons are pairwise and grounded, the judge receives the question, the gold answer, and the gold evidence. Every pair is judged in both presentation orders, and we consolidate with a strict agreement rule, a question is decisive for a configuration only when both orders name that same configuration, with any disagreement, any tie verdict, or any missing order counted as a tie. Win-rates count ties as one half. Within-model comparisons use a single grounded judge (Qwen3.6-Flash, temperature 0, fixed seed); a pre-registered bridge sample re-judged 75 of its within-model pairs with the frontier panel, and the single judge agrees with the panel majority at Krippendorff's $\alpha = 0.84$ (89 percent exact agreement), so the paper's two result families sit on one measurement scale [F19]. The frontier comparison adds a family-firewalled three-judge panel (Gemini, Mistral, and Kimi; no judge shares a model family with any contestant, with the model that wrote the gold answers, or with the model that wrote the fine-tuning pairs), consolidated per judge and then by majority of decisive verdicts with a quorum of two, plus a gold-anchor arena. One labeling note applies throughout, the lever we call RAG-over-LoRA denotes `lora_rag` versus `lora`, retrieval

added on top of the fine-tuned model, not `rag` versus `lora`.[1]

The frontier comparison uses a clean-judge protocol motivated by two artifacts our own audit surfaced. With a capped evidence window, expanding the window flipped ties to the frontier 20 to 0 (exact binomial p = 1.9e-6), and on independent re-reading roughly 14 to 15 of those 20 flips rewarded verbosity rather than correctness [F15]. The clean protocol therefore places the full retrieved context in the judge's evidence window and adds a de-verbosity rubric line (shipped verbatim in the artifact). Generation used max_tokens of 512 (the DeepSeek arm was regenerated at 4,096 after the token-cap incident of Section 8), and a silent temperature-drop-on-error fallback path existed in the generation stack; both are disclosed [F08, F10].

3.4 Statistics

Within-model effects use exact McNemar tests (McNemar, 1947) on discordant pairs, with the full 36-test family controlled by Holm (Holm, 1979) and, as a sensitivity variant, Benjamini-Hochberg (Benjamini and Hochberg, 1995). Near-null cells are tested for equivalence by Schuirmann's TOST (Schuirmann, 1987) at a margin of ±0.05. Lever-by-corpus interactions use G-tests with question-block permutation as a clustering guard. The frontier arm is estimation-first because three frontier models share each question set, we report question-clustered bootstrap intervals and question-level sign tests, with descriptive p-values and no significance stars. Triangulation beyond the judge comes from Bradley-Terry rankings (Bradley and Terry, 1952) with Krippendorff's α (Krippendorff, 2004) for panel agreement, and from judge-free ROUGE-L (Lin, 2004).

4 Access, not capability

This section establishes the title claim: closed-book, the frontier loses to a local model of any size we test once that model has retrieval; handed the identical retrieved context, it recovers only a small, bounded edge.

4.1 The access gap

Run closed-book, the three frontier models fail to beat the 8 billion parameter local model with retrieval in 95.8 percent of panel-judged comparisons, and fail against the 1 billion parameter model in 91.4 percent, with outright frontier losses at 75.0 and 65.9 percent (Figure 1; the panel judged a 200-question subsample per matchup, and single-judge figures over the full sets are 98.8 and 96.0 percent) [F05]. A model without the documents losing on document QA surprises nobody. What the floor supplies is its magnitude and the failure-mode split: GPT-5.1 attempts an answer on 49.9 percent of questions with losses dominated by hallucinated content, while Claude-Sonnet-4.6 and DeepSeek-v4-pro formally abstain on 66 to 77 percent, a threshold-sensitive band (classifier rules in Appendix E) [F05]. The scoring is fair to abstention, since across all 98 pure-abstain pairs on out-of-corpus questions the frontier model is never scored as losing [F05]. Because the private corpus consists of published papers, this floor doubles as the empirical non-memorization probe promised in Section 3.1. The news of this result sits in Section 4.2, the size of the premium once access is equalized.

Table 2: Frontier versus local under byte-identical retrieval, clean judging protocol, strict both-orders consolidation. F/L/T are frontier wins, local wins, and ties over pairs; net win-rate counts ties as half; intervals are question-clustered bootstrap (5,000 resamples). Per-cell question counts run 342 to 344 on the private corpus (attrition disclosed); the 3.8 billion parameter row runs on all 345; the QASPER frontier cells run on a 200-question subsample of the 400-question set (149 answerable, 51 out-of-corpus). [F06, F07, F20]

cell	F/L/T	net win-rate [95% CI]	local not-lose	tie share
GPT-5.1 vs 8B (private)	66/5/271	0.589	0.807	–
Claude-4.6 vs 8B (private)	87/3/254	0.622	0.747	–
DeepSeek-v4 vs 8B (private)	66/6/272	0.587	0.808	–
pooled vs 8B, private, all	219/14/797	0.600 [0.580, 0.620]	0.787	0.774
pooled vs 8B, private, answerable	219/14/641	0.618 [0.595, 0.641]	0.749	0.733
pooled vs 8B, QASPER, answerable	85/20/342	0.573 [0.543, 0.603]	0.810	0.765
pooled vs 3.8B, private, all	149/27/859	0.559 [0.541, 0.577]	0.856	0.830
pooled vs 1B, private, all	467/14/548	0.720 [0.695, 0.746]	0.546	0.532
pooled vs 1B, QASPER, answerable	250/14/183	0.764 [0.726, 0.802]	0.441	0.409

4.2 The capability premium

Table 2 gives the same-retrieval comparison. Leading with the three-outcome record, the clean judge ties the 8 billion parameter model with the frontier on 77.4 percent of pairs, the frontier wins 91 to 97 percent of the decisive minority, and local decisive wins number 14 of 1,030 pairs (1.4 percent) [F06]. The pooled net win-rate is 0.600 with a question-clustered interval of [0.580, 0.620], rising to 0.618 on answerable questions, and the question-level sign test is 98 to 10 ($p = 2.7e\text{-}19$). Independent re-reading of judge verdicts (Appendix D) shows the frontier's edge concentrating in fine-detail extraction by the local model, a misread effect size, a wrong regression type, a false abstention [F15].

4.3 Stability of the estimate

Figure 2 assembles the chain of checks that makes the claim land. Moving from the capped panel to the clean protocol shifts the pooled estimate from 0.573 to 0.600, the two judge artifacts of Section 3.3 roughly canceling; a capped single-judge cross-check (0.573 to 0.578) isolates the protocol change from the panel-to-single change [F06, F15]. Across the five main panel aggregation rules the net win-rate spans 0.540 to 0.575, with unanimous-only, the strictest variant, at 0.535, and requiring two decisive judges in agreement raises local not-lose to 0.835 [F15]. Excluding the DeepSeek arm entirely leaves the pooled estimate at 0.606 with not-lose 0.777, so the headline is not carried by any single frontier vendor [F17], and the question-clustered sign test survives even the least favorable judging regime (capped panel, 60 to 26, $p = 3.2e\text{-}4$) [F06]. The decisive-verdict provenance splits 51 unanimous, 59 two-to-one, 66 single-decisive, and 5 two-judge [F15].

The stratification by retrieval success answers the strongest attack we could construct against the headline, the hypothesis that the 77 percent tie mass merely reflects both models failing to see the gold passage at a recall of 0.294. Stratifying every clean-judged pair by whether a gold chunk sat in the shared top-6, the HIT minus MISS delta in net win-rate is +0.009 [-0.037, +0.057] on the private corpus and +0.057 [-0.029, +0.142] on QASPER, with all four pooled cells' intervals containing zero [F11]. The 8 billion parameter model ties the frontier roughly 77 percent of the time

[1] A naive reading of the label as `rag` versus `lora` produces different counts and does not reproduce any table in this paper. [F01]

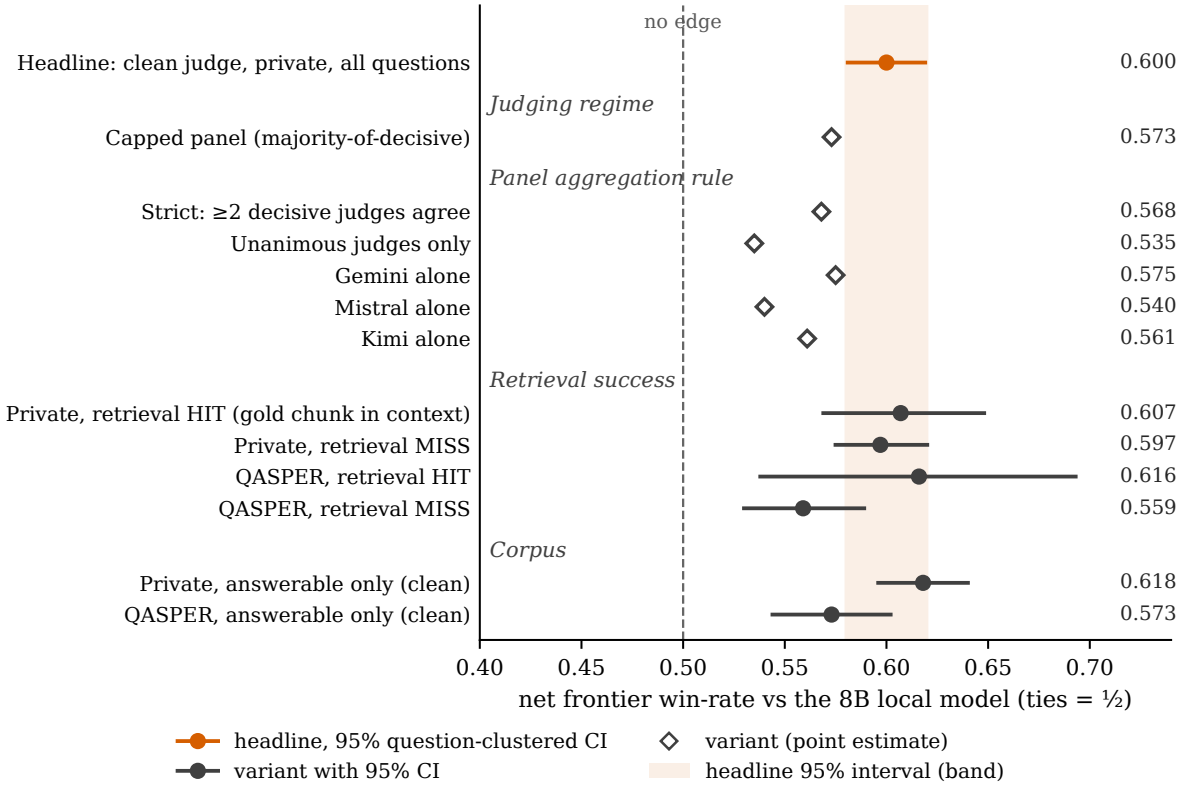

Figure 2: The estimate of the capability premium under every variation we tested (private-corpus pooled cells n = 1,030 pairs; HIT/MISS strata and QASPER cells as in Appendix G). Judging regime moves the pooled net win-rate from 0.573 (capped panel) to 0.600 (clean protocol); the five main panel aggregation rules span 0.540 to 0.575, with unanimous-only, a sixth and strictest variant, at 0.535; stratifying by retrieval success leaves the premium statistically unchanged (HIT minus MISS delta of +0.009 [-0.037, +0.057] on the private corpus); the public-corpus replication lands at 0.573 [0.543, 0.603]. The shaded band is the headline 95 percent interval (legend). [F06, F11, F15]

even when the gold chunk is in context, so the tie mass reflects capability, not mutual blindness. Finally, the result replicates on public QASPER at 0.573 [0.543, 0.603] with answerable not-lose of 0.810; under the original capped single judge the same arm gave 0.524 with not-lose 0.906 [F06, F07].

4.4 Contamination control: the QASPER inversion

A naive reading of the closed-book results suggests the frontier looks stronger on public QASPER (frontier win-rate 0.446 on the 200-question QASPER frontier subsample against 0.157 on the private corpus), inviting a memorization explanation. Stratification dissolves it. The entire lift sits in the out-of-corpus stratum, where abstention is the correct response and earns a frontier win-rate of 0.833; on answerable public questions the closed-book frontier loses just as it does on the private corpus, a frontier win-rate of 0.313 with local not-lose of 0.982 [F07]. Frontier models do not recall public QASPER answers either, so the private-corpus access gap is not a memorization artifact.

4.5 The size floor

The capability premium is no larger against the 3.8 billion parameter model than against the 8 billion one; only the 1 billion configuration leaves the frontier a real edge. Against the 1 billion parameter model under identical retrieval, the frontier's net win-rate rises to 0.720 [0.695, 0.746] on the private corpus and 0.764 [0.726, 0.802] on QASPER [F06]; against the intermediate 3.8 billion parameter model, judged by the same clean protocol on the same full question set, it falls to 0.559 [0.541, 0.577], at or slightly below its 0.600 against the 8 billion parameter model (Table 2) [F20]. Closed-book, the story is the same at every size, the frontier fails to beat the 3.8 billion parameter model in 97.6 percent of panel-judged comparisons, level with the 91.4 and 95.8 percent it manages against the 1 and 8 billion parameter models, so the access gap is size-independent [F20]. The ladder is therefore not a graded size response. The frontier's edge is large only against the 1 billion configuration, and the strong recent 3.8 billion parameter model matches or edges the older 8 billion one, a two-model contrast on one corpus that reads as model quality mattering more than raw size once capacity clears roughly 4 billion parameters. The capped-evidence panel reproduces the ordering [F20], and one caveat holds on the closed-book figures, the 1 and 8 billion arms used the preregistered 200-question subsample while the 3.8 billion arm ran on the full 345.

4.6 The procurement arm: full context does not rescue the frontier

The same-retrieval comparison hands the frontier the top-6 chunks the local model used. A procurement-minded reader asks the harder question, what if the frontier is given the whole corpus, the 416,698-token collection that fits a modern million-token window. We ran that arm as a pilot, the full corpus in context for every question, on the two frontier models whose windows hold it. GPT-5.1's does not, its 400,000-token window is smaller than the 416,698-token corpus and the endpoint rejects the request, so full context is not even an option for one of the three frontier models, a first result in its own right. Against the local 8 billion parameter model with retrieval, a frontier handed the entire corpus scores a net win-rate of 0.561 [0.508, 0.614] pooled over the two models, and the pooling hides real heterogeneity we report rather than average away: Claude gains a modest edge (0.622) while DeepSeek sits at exact parity (0.500); pooled, the whole-corpus arm's 0.561 does not exceed the 0.600 the same frontier earns with retrieval alone, and the local model is not beaten in 89 percent of comparisons [F22]. Handing the frontier everything does not widen the access gap. And for the frontier itself, retrieval beats full context in both models, its own top-6 retrieval outscores its whole-corpus version at a net 0.433 [0.392, 0.475] pooled (Claude 0.478, DeepSeek 0.389), the interval falling entirely below parity, direct evidence at full-corpus scale

Table 3: Frontier API generation cost per 1,000 questions (USD), measured from logged token usage at current list prices. Local figures are marginal API cost only; hardware is assumed owned, and no latency or total-cost-of-ownership claim is made. [F14]

setting	GPT-5.1	Claude-Sonnet-4.6	DeepSeek-v4-pro
closed-book	$1.44	$1.70	$0.29
with retrieval (∼4.5k-token context)	$6.31	$16.34	$2.20
local 8B (any setting)	$0 marginal API cost		

of the degradation the long-context literature documents, positional loss (Liu et al., 2024) and effective context lengths well below advertised windows (Hsieh et al., 2024) [F22]. The corpus enters the window in its natural file order, one ordering per question, so position effects are folded into the estimate rather than tuned away, the realistic naive paste a practitioner would run. The pilot covers 90 questions per model under the clean judge, and it points one way, buying the frontier at full strength buys no more than retrieval already provides, and retrieval remains the better way to deliver the documents even to the frontier.

4.7 A tested null

Corpus heterogeneity does not grow the frontier premium. The QASPER estimate of 0.573 sits at or below the private-corpus values of 0.600 to 0.618 under clean judging, so the homogeneity variable of Section 6 organizes the lever, mechanism, and abstention results but not this one [F07].

4.8 Cost

Table 3 prices the alternatives. Serving 1,000 retrieval-augmented questions through the frontier APIs costs between $2.20 and $16.34 in generation alone, against zero marginal API cost for the local 8 billion parameter model on owned hardware. Combined with the access gap and the bounded capability premium, the cost table is what earns the democratization framing of Section 10, the gap between a small organization and a frontier lab on private-document QA is document access plus one consumer GPU, not API budget. We make no latency or throughput claim, neither was measured.

5 The lever ranking: buy retrieval, not fine-tuning

If the frontier's advantage is access, the practitioner's question becomes which lever buys that access best. Table 4 and Figure 3 present all 36 within-model cells. Retrieval is the first lever. It wins 59.3 to 87.0 percent of comparisons across the 18 retrieval cells and survives global Holm correction in all 18 (RAG-over-base and RAG-over-LoRA combined) [F01]. Fine-tuning is the second, with LoRA-over-base gains of roughly 2 to 6 points, 7 of 9 surviving Holm, and several near-null cells certified TOST-equivalent to no effect, indicating low returns on effort from fine-tuning alone. Across three training seeds on two cells, the seed standard deviation is 0.004 on the TOST-null private Llama cell and 0.009 on the Holm-surviving private Phi cell, where the roughly six-point effect is about six times its own seed noise (Section 9) [F18]. RAFT is the final and weakest lever, surviving in only 2 of 9 cells, COVID-QA with MiniCPM at 58.0 percent ($p = 3.7e\text{-}6$) and QASPER with Llama at 54.6 percent ($p = 7.6e\text{-}4$). Under the Benjamini-Hochberg variant the tally moves to 29 of 36 (LoRA 8 of 9, RAFT 3 of 9), leaving the ordering unchanged.

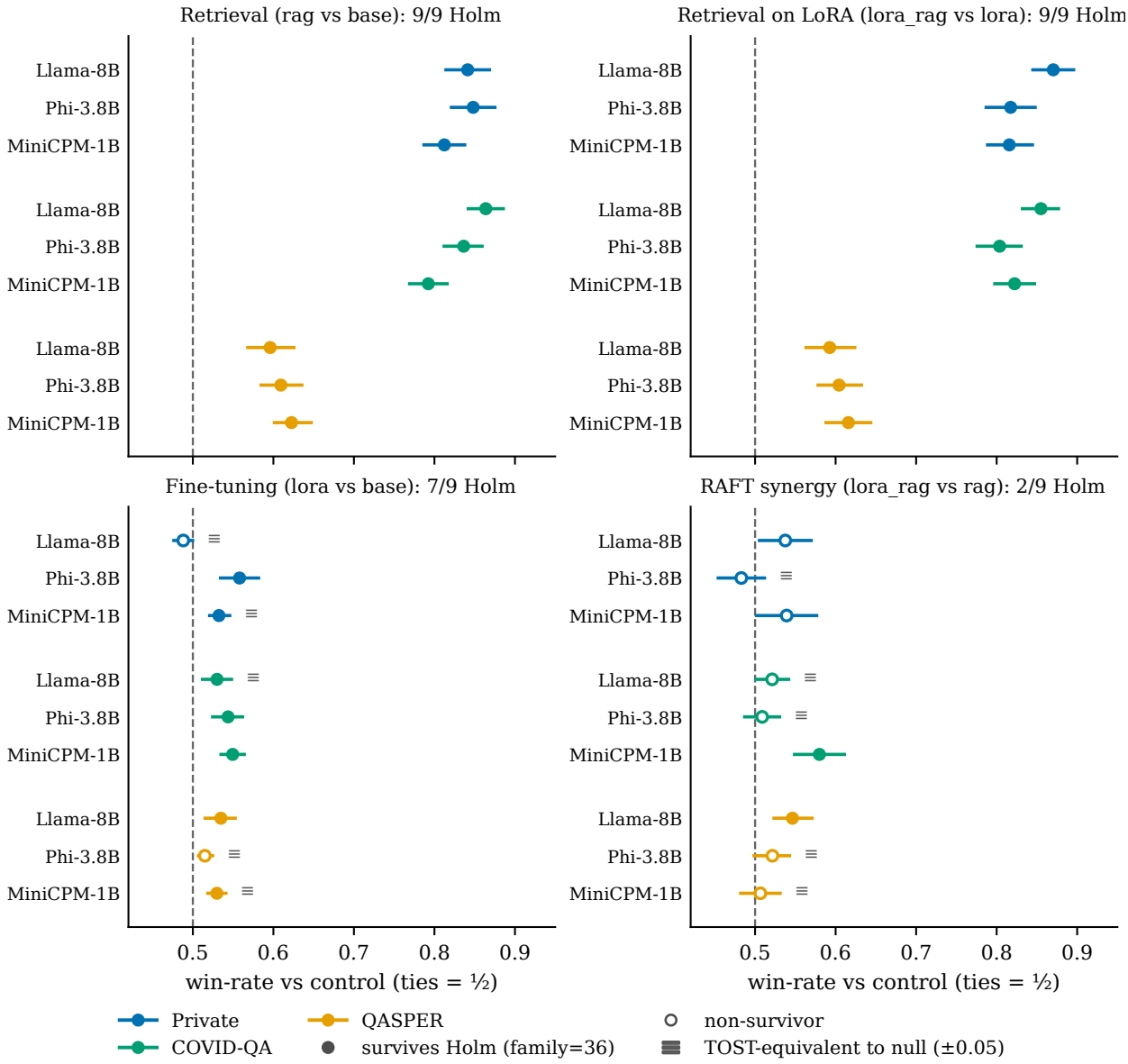

Figure 3: All 36 within-model lever cells (answerable questions, strict both-orders consolidation, ties counted as half). Horizontal bars are 95 percent confidence intervals; per-cell n is 293 (private), 393 to 400 (COVID-QA), and 302 (QASPER), as in Table 4. Filled markers survive global Holm correction over the 36-test family; open markers do not; ≡ marks cells TOST-equivalent to the null at ±0.05. Retrieval helps everywhere; fine-tuning helps a little everywhere it helps at all; the RAFT synergy appears in 2 of 9 cells. [F01]

Table 4: Win-rates (percent, ties as half) for the four levers across all 36 cells, answerable questions only. ✓ survives global Holm (family of 36); ≡ is TOST-equivalent to the null at ±0.05. Per-cell n: private 293, COVID-QA 400 (MiniCPM 393–395 from 76 missing raw rows, disclosed), QASPER 302. RAG-over-LoRA denotes `lora_rag` versus `lora`. MiniCPM QASPER cells are affected by the generation truncation disclosed in the text (136, 78, and 23 of 400 answers on `lora_rag`, `lora`, and `rag`); read those cells with that caveat. [F01]

corpus	model	RAG-over-base	RAG-over-LoRA	LoRA-over-base	RAFT-over-RAG
Private	Llama-8B	84.1 ✓	87.0 ✓	48.8 ≡	53.8
Private	Phi-3.8B	84.8 ✓	81.7 ✓	55.8 ✓	48.3 ≡
Private	MiniCPM-1B	81.2 ✓	81.6 ✓	53.2 ✓ ≡	53.9
COVID-QA	Llama-8B	86.4 ✓	85.5 ✓	53.0 ✓ ≡	52.1 ≡
COVID-QA	Phi-3.8B	83.6 ✓	80.4 ✓	54.4 ✓	50.9 ≡
COVID-QA	MiniCPM-1B	79.2 ✓	82.2 ✓	54.9 ✓	58.0 ✓
QASPER	Llama-8B	59.6 ✓	59.3 ✓	53.5 ✓	54.6 ✓
QASPER	Phi-3.8B	60.9 ✓	60.4 ✓	51.5 ≡	52.2 ≡
QASPER	MiniCPM-1B	62.3 ✓	61.6 ✓	52.9 ✓ ≡	50.7 ≡

Examining the interaction structure sharpens the picture. Retrieval's effect shrinks on the heterogeneous corpus, an interaction decisive for both retrieval levers (G = 26.5 and 39.8, both under 10^{-6} and confirmed by question-block permutation), while LoRA and RAFT show no corpus interaction at all (minimum p = 0.091) [F02]. One scope qualifier applies, the retrieval interaction is carried by Llama and Phi while MiniCPM sits saturated, so it should not be read as uniform across sizes [F02].

A caveat box belongs here. On QASPER, MiniCPM exhibits generation truncation in 136 of 400 `lora_rag` answers, 78 `lora` answers, and 23 `rag` answers, a symmetric collapse at judge time that we disclose so the affected cells are read with appropriate caution [F10]. The practitioner reading of this section is short, spending belongs on retrieval rather than fine-tuning, and RAFT rarely adds value on top of an existing RAG pipeline. Judge-free ROUGE-L reproduces the configuration ordering on all 9 corpus-model rows (Appendix F) [F09].

6 Corpus structure, not size, predicts retrieval's value

Retrieval wins everywhere, but by margins that range from 59.6 to 86.4 percent, and this section supplies the variable that predicts the difference. Writing r for retrieval recall@6, w_{hit} for the win-rate when a gold chunk is retrieved, and w_{miss} for the win-rate when none is, the aggregate retrieval effect obeys the identity

$$\text{aggregate} \;=\; r \cdot w_{\text{hit}} \;+\; (1 - r) \cdot w_{\text{miss}}, \tag{1}$$

which holds exactly on all three corpora [F03]. Two regimes emerge from the decomposition. COVID-QA is recall-driven, a recall of 0.813 multiplied by a hit-win of 0.931 produces nearly the entire aggregate of 0.864. The private corpus and QASPER instead separate on the miss side, sitting in a similar low-recall band (0.294 against 0.232) while their miss-win rates diverge, 0.826 against 0.532.

The controlled contrast carries the inference. Between the private corpus and QASPER lies an approximately 33 fold corpus size gap, yet their recall values nearly coincide, and what tracks the miss-win divergence is the homogeneity gap, 1.68 fold on TF-IDF centroid similarity and 2.8 fold on mean pairwise cosine (Table 1). On a homogeneous corpus, retrieval that misses the gold

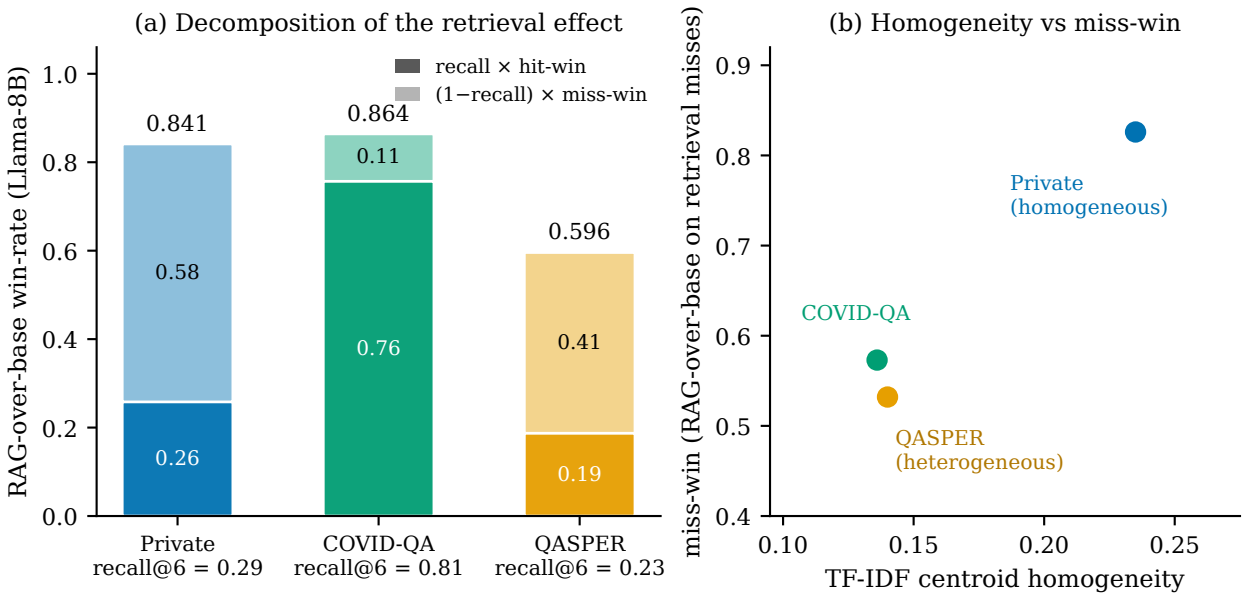

Figure 4: The mechanism behind Section 6 (Llama-8B, answerable questions; $n = 293$ private, 400 COVID-QA, 302 QASPER). (a) The decomposition aggregate $= r \cdot w_{\text{hit}} + (1-r) \cdot w_{\text{miss}}$ holds exactly on every corpus; lighter segments show the miss-side contribution. (b) Miss-win plotted against TF-IDF centroid homogeneity for the three corpora, three points shown without a fitted line and without intervals; the private corpus and QASPER sit in a similar low-recall band (0.294 against 0.232) yet differ in miss-win by 0.826 against 0.532, a pattern consistent with homogeneity rather than corpus size as the governing variable. [F03]

chunk still surfaces topically adjacent text that helps the model; on a heterogeneous corpus the same miss surfaces text from an unrelated paper. With three corpora supplying three homogeneity points, we state the relationship as consistent with homogeneity governing miss-win rather than proven by it, and a fourth corpus at intermediate homogeneity stands as the natural extension. As a practitioner diagnostic the implication is direct and cheap, TF-IDF homogeneity of the corpus can be measured before any architecture decision, predicting both the expected benefit of retrieval here and the abstention hazard of Section 7.

7 Retrieval flips abstention behavior with corpus structure

Add retrieval, and the same 8 billion parameter model's correct abstention on unanswerable questions moves from 61.5 to 100 percent on the homogeneous corpus (n = 52) and collapses from 63.3 to 17.3 percent on the heterogeneous one (n = 98), the model engaging with plausible-looking but irrelevant retrieved text; every behavioral number here comes from a deterministic scripted rule set rather than judge impressions [F12].

The pairwise endpoints separate just as cleanly, the RAG-over-base win-rate on unanswerables pooled across the three models is 0.705 on the private corpus (n = 156 pairs over 27 unique texts; text-clustered interval [0.59, 0.83]) against 0.221 on QASPER (n = 294; [0.19, 0.26]), disjoint intervals that hold per-model 3 of 3 [F04]. The direction is classifier-robust, under the stricter pure-abstain-only rule the same collapse runs 58 to 6 of 98 [F04, F12]. On answerable private-corpus questions the same mechanism is a benefit, false abstention drops from 85.7 to 11.3 percent. The beneficial direction generalizes across local size, with retrieval turning answerable-versus-unanswerable discrimination from negative to strongly positive Youden's J on the homogeneous corpus for all three

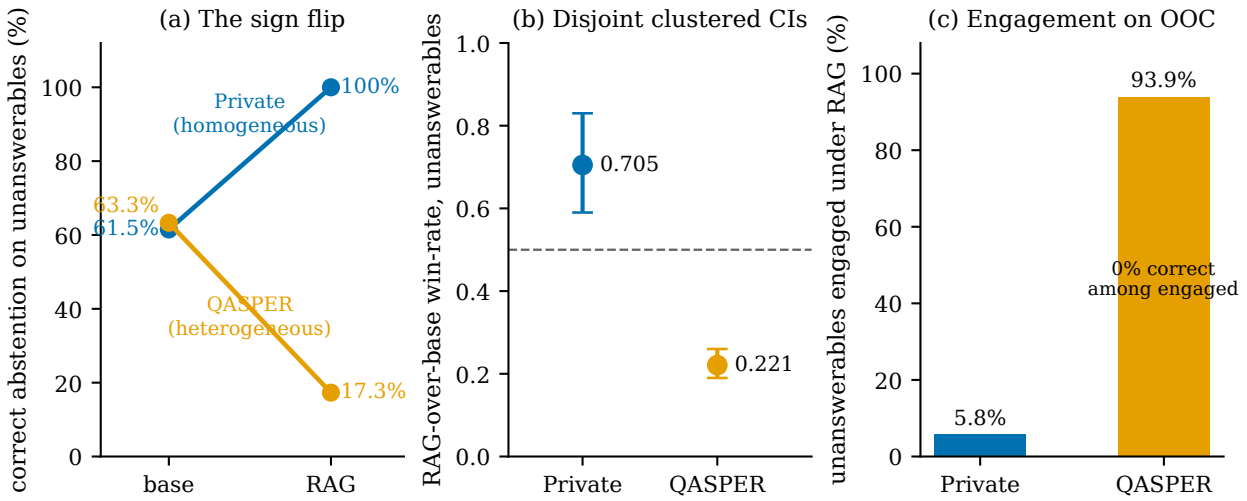

Figure 5: The abstention sign-flip (scripted classifier, rules in Appendix E; panels a and c: Llama-8B; panel b: pooled over the three local models). (a) Correct abstention on unanswerable questions moves from 61.5 to 100 percent under retrieval on the homogeneous private corpus (n = 52) and from 63.3 to 17.3 percent on heterogeneous QASPER (n = 98). (b) RAG-over-base win-rate on unanswerables with text-clustered 95 percent intervals, 0.705 (n = 156 pairs over 27 unique texts) against 0.221 (n = 294), disjoint. (c) Under retrieval, QASPER engages 93.9 percent of unanswerables, and every engaged answer is an error. [F04, F12]

models (base and LoRA J from −0.43 to −0.01, every retrieval configuration +0.68 to +0.91) [F21]. A risk-coverage reading makes the deployment hazard concrete, QASPER under retrieval engages 93.9 percent of unanswerables (self-rejecting only 6 percent) in a setting where every engaged answer is an error, while the engaged-answer quality gain from retrieval is negligible at +0.007 ROUGE-L, locating the harm in behavior rather than in answer quality [F12].

Two scoping notes apply. With two corpora the contrast supports two-corpus language only, the flip holds on the homogeneous corpus against the heterogeneous one, and gold provenance differs between them (LLM against human), a confound we disclose. The organizing variable is the same homogeneity measure as Section 6, and we read the connection as consistent with the mechanism rather than demonstrated.[2] The practitioner consequence is one sentence, a heterogeneous corpus joined to an unanswerable-heavy workload requires an abstention guard before deployment.

8 Threats to validity

We treat this section as a result in its own right, every threat below was measured rather than assumed [F10, F13, F15].

Judge biases. Position bias measures 0.490 to 0.497 across runs, and the both-orders design demonstrably contains the worst case, the mistral judge prefers the first-presented answer on 73.1 percent of its decisive raw verdicts, yet strict both-orders consolidation fully absorbs the bias. Verbosity preference is clean within-model (0.501, 0.429, 0.484, with COVID-QA favoring shorter answers). The two frontier-judging artifacts, the evidence-window effect (20 to 0 tie-flips, p = 1.9e-6) and the verbosity reward (roughly 14 to 15 of those 20 flips), were found by our own audit and corrected by the clean protocol of Section 3.3.

[2] An earlier informal spot-check of abstention frequency had the direction backwards; every behavioral number here derives from the scripted classifier, whose rules ship verbatim in the artifact. [F12]

Panel. Krippendorff's α runs 0.57 (0.565) to 0.64 across panel runs, moderate agreement, rising to 0.75 with the position-biased mistral judge excluded, direct evidence the both-orders consolidation contains the one biased judge [F15]; the rule-sensitivity table of Figure 2 spans 0.540 to 0.575 across the five main rules; kimi judge attrition (122 of 125 two-judge units) is outcome-neutral, with the frontier-local gap preserved in the attrited units.

Data integrity. Evaluation sets are hash-locked (SHA-256 verified across run snapshots), corpus routing is exact (1.0 on all 18 cells, zero empty retrievals), per-cell counts are disclosed everywhere they deviate (capped single-judge frontier arm 334 to 345, clean-protocol cells 342 to 344; COVID-QA MiniCPM 393 to 395), and deduplication is disclosed for all three corpora.

Gold provenance. Private-corpus gold is LLM-generated (Qwen-2.5-14B). Mitigations are layered, two of three corpora carry human gold, the frontier judging panel contains no Qwen-family model, judge-free ROUGE-L reproduces the configuration ordering, and a human-gold re-annotation of the private corpus (a blind, stratified, seeded pairwise protocol shipped in the artifact) is in progress and will fold into the camera-ready version. The within-model grid (the lever ranking, the mechanism, and the abstention result) is scored by a Qwen-family judge against that Qwen-written gold. The design mitigates this structurally rather than by assumption, both configurations in every pair are scored against the same gold by the same judge in both orders (any judge-gold affinity applies symmetrically to the pair), judge-free ROUGE-L reproduces the ordering, and the cross-judge bridge of Section 3.3 shows the Qwen judge agreeing with the non-Qwen panel majority at $\alpha = 0.84$ [F19].

Process evidence. A token-cap incident in the DeepSeek arm was found, fixed, and verified at the answer level, 16 of 24 first-run decisive local wins dissolved on re-judging, 14 of them provable exact-512-token truncations, with corrected deltas of +0.011 to +0.036, and the first run carried 41 exact-512 answers against zero after the fix [F08]. A second incident, a wrong-corpus retrieval defect in one configuration arm, was caught by an exact retrieved-chunk-ID namespace audit and the arm re-run, which is why corpus routing is verified at 1.0 on all 18 cells above; both detections (an exact-length histogram spike at the generation cap, and chunk-ID namespace membership) are judge-free integrity checks that generalize beyond this study. Reproducibility is structural, every number re-derives from raw outputs via the receipt scripts, and the claim-to-script-to-receipt map ships with the paper.

9 Limitations

Three corpora supply three homogeneity points, so the mechanism of Section 6 is stated as consistent-with, upgradeable by a fourth corpus at intermediate homogeneity. The frontier comparison covers two corpora and three local sizes, all three placed on one clean-protocol row on the private corpus, and it answers both the controlled capability question (same retrieval) and the procurement question (the frontier handed the entire corpus, Section 4.6). The long-context arm is a pilot at 90 questions per model on the two frontier windows that hold the 416,698-token corpus; scaling it to the full question set and to more long-context models, and a frontier-native long-context arm on the public corpora, are the natural extensions. Gold on the headline corpus is LLM-generated pending the human re-annotation. Judge agreement is moderate. Each training configuration in the main grid ran once under a fixed seed, so the replication axis of the study is models and corpora rather than seeds; three-seed replicates of two cells (LoRA-over-base on the private corpus, seeds 42, 43, and 44, identical grounded judging) bound the seed noise at standard deviations of 0.004 on the TOST-null Llama cell and 0.009 on the Holm-surviving Phi cell, against effects of 2 to 6 points, so the lever ranking sits well clear of training-seed noise on both replicated

cells [F18]. The retrieval stack of Section 3 was locked by one pre-study sweep and then held fixed, and a private-corpus recall@6 of 0.294 leaves retriever quality a live variable this study does not explore, so the reported retrieval effects describe one reasonable stack rather than retrieval at its best. No latency or total-cost-of-ownership measurement was taken, and calibration analysis in the ECE style was infeasible without a logprob-capturing rerun. Absolute quality leaves headroom for every model, the gold anchor dominates all six Bradley-Terry rankings, and three never-beaten anchors are non-convergent maximum-likelihood estimates we report as lower bounds [F09].

10 Conclusion

In summary, the frontier's advantage on private corpora is access, not capability, and the recipe that follows has two steps. Retrieval comes first, necessary and dominant: denied document access, even frontier models fail on unseen content. Capacity matters up to a point, with the 3.8 billion parameter configuration already closing most of the residual gap the 1 billion one leaves open and sitting level with the 8 billion one under a matched judge, the remaining premium small, bounded, concentrated in fine detail, and statistically unchanged by retrieval success. Corpus homogeneity, measurable for pennies with TF-IDF before any training run, predicts both the size of the retrieval benefit and the abstention hazard, with low homogeneity joined to unanswerable-heavy workloads calling for an explicit abstention guard. Fine-tuning demotes to an optional refinement, two to six points where it helps at all (7 of 9 cells), corpus-independent, worth the effort only when the last points matter. The equity reading follows from the cost table rather than from assertion: on private-document QA the distance between a school district and a frontier laboratory is document access plus one consumer GPU, not an API budget. One asymmetry deserves final statement. This is an access result and never a capability reversal; the frontier remains the more capable model and simply gets out-competed wherever it lacks the documents the local model can retrieve.

Reproducibility statement

Every number in this paper re-derives from raw model outputs, offline, through one of twenty-two receipt scripts (F01 through F22) shipped with the artifact, and a 186-claim ledger, extended by the dated 2026-07 addenda for the seed replicates, the cross-judge bridge, the 3.8 billion parameter frontier arm, the cross-size abstention discrimination, and the whole-corpus long-context procurement arm, walks each sentence of prose to its script and its receipt (Appendix H). Raw outputs for the two public corpora ship in full; private-corpus raw files stay with the corpus under the data-availability terms, so re-derivation of the headline-corpus numbers is available to auditors rather than to the public. The training seed is 42 everywhere, the per-model, per-corpus configurations were frozen before the final runs, and the evaluation sets are SHA-256 hash-locked, every downstream phase checking the hash before it will run. Each training configuration in the main grid ran once, so replication in this study runs across models and corpora rather than seeds; three-seed replicates of two cells bound training-seed variance well below the reported effects (Section 9).

Data availability

The artifact carries everything we can lawfully hand over, the full pipeline and analysis code, the recompute scripts and their claim-to-receipt map, the judge prompts and rubrics, the homogeneity metric, and the complete QASPER and COVID-QA evaluation sets with their human gold. The private corpus does not travel. Its 28 papers are copyright-protected published articles we have

no right to redistribute, so the collection stays local along with the LLM-generated gold derived from it, and we document provenance instead, 13 of 28 papers carrying OpenAlex work identifiers through which a reader with lawful access can rebuild part of the corpus.

Broader impact

We read the results as an equity finding first. A school district, a clinic, or a small laboratory holding its own documents and one consumer GPU closes most of the distance to a frontier laboratory on private-corpus QA, and the documents never leave the building, which is the point for organizations bound by privacy obligations. The hazard of Section 7 deserves equal billing, on a heterogeneous corpus retrieval drove engagement with 93.9 percent of unanswerable questions in a setting where every engaged answer was wrong, so a deployment without an abstention guard converts missing knowledge into confident misinformation. No human subjects were involved, no personal data was processed, and the corpus consists of published research papers.

References

Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*, 2024.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling (COLM)*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, 2004.

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*, 2020.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 881–893, 2024.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

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*, 2023.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*, 2023.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 2024.

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*, 2018.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15:657–680, 1987.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Fine tuning vs. retrieval augmented generation for less popular knowledge. *arXiv preprint arXiv:2403.01432*, 2024.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.

A Per-cell tables

Table 4 reports all 36 within-model cells with Holm and TOST markings; Table 2 reports the frontier cells with per-cell pair counts. Table 5 prints the underlying statistics the body figures show only as markers, the exact McNemar p-value, Holm survival, and TOST verdict for every cell. Raw win/loss/tie triples and bootstrap intervals re-derive from the shipped receipt scripts (Appendix H).

B The homogeneity diagnostic

The corpus-homogeneity covariate of Section 6 is a judge-free, offline TF-IDF measure a practitioner can run before any training. On a seed-1337 sample of up to 1,200 chunks per corpus (chunks under 20 words dropped as title or stub fragments so the measure reflects text, not headers), a single TF-IDF vectorizer with identical settings for every corpus (`max_features`=5000, English stop-words, sublinear term frequency, minimum document frequency 2, L2-normalized rows) produces three quantities: the mean cosine similarity over 40,000 random chunk pairs, the mean cosine to the corpus centroid (topical concentration), and the ratio of across-paper to within-paper similarity (a homogeneous corpus scores across $\approx$ within). The private corpus and QASPER separate cleanly on all three (Table 1 reports the centroid and pairwise figures), and it is this gap, not the roughly 33-fold chunk-count gap between them, that tracks the miss-win divergence of Section 6. The metric ships in the artifact [F03].

Table 5: Per-cell within-model statistics for all 36 lever cells (answerable questions, strict both-orders consolidation). Win-rate counts ties as half; p is the exact McNemar test on discordant pairs; ✓ marks cells surviving global Holm correction over the 36-test family; ≡ marks cells TOST-equivalent to the null at ±0.05. This appendix prints the statistics Table 4 and Figure 3 show only as markers. [F01]

corpus	model	lever	win-rate	McNemar p	Holm	TOST
Private	Llama-8B	RAG-over-base	84.1	4.0×10^{-54}	✓	
		RAG-over-LoRA	87.0	3.9×10^{-60}	✓	
		LoRA-over-base	48.8	0.143		≡
		RAFT-over-RAG	53.8	0.035		
	Phi-3.8B	RAG-over-base	84.8	2.6×10^{-54}	✓	
		RAG-over-LoRA	81.7	9.9×10^{-43}	✓	
		LoRA-over-base	55.8	2.4×10^{-5}	✓	
		RAFT-over-RAG	48.3	0.320		≡
	MiniCPM-1B	RAG-over-base	81.2	7.6×10^{-54}	✓	
		RAG-over-LoRA	81.6	5.7×10^{-48}	✓	
		LoRA-over-base	53.2	2.1×10^{-5}	✓	≡
		RAFT-over-RAG	53.9	0.050		
COVID-QA	Llama-8B	RAG-over-base	86.4	1.2×10^{-76}	✓	
		RAG-over-LoRA	85.5	1.6×10^{-76}	✓	
		LoRA-over-base	53.0	0.005	✓	≡
		RAFT-over-RAG	52.1	0.086		≡
	Phi-3.8B	RAG-over-base	83.6	5.2×10^{-67}	✓	
		RAG-over-LoRA	80.4	4.0×10^{-54}	✓	
		LoRA-over-base	54.4	8.2×10^{-5}	✓	
		RAFT-over-RAG	50.9	0.538		≡
	MiniCPM-1B	RAG-over-base	79.2	3.9×10^{-62}	✓	
		RAG-over-LoRA	82.2	2.5×10^{-64}	✓	
		LoRA-over-base	54.9	7.6×10^{-9}	✓	
		RAFT-over-RAG	58.0	3.7×10^{-6}	✓	
QASPER	Llama-8B	RAG-over-base	59.6	1.9×10^{-9}	✓	
		RAG-over-LoRA	59.3	3.2×10^{-8}	✓	
		LoRA-over-base	53.5	0.002	✓	
		RAFT-over-RAG	54.6	0.001	✓	
	Phi-3.8B	RAG-over-base	60.9	5.8×10^{-15}	✓	
		RAG-over-LoRA	60.4	2.7×10^{-12}	✓	
		LoRA-over-base	51.5	0.012		≡
		RAFT-over-RAG	52.1	0.105		≡
	MiniCPM-1B	RAG-over-base	62.3	1.1×10^{-22}	✓	
		RAG-over-LoRA	61.6	5.6×10^{-14}	✓	
		LoRA-over-base	53.0	7.6×10^{-6}	✓	≡
		RAFT-over-RAG	50.7	0.712		≡

C Judging protocol details

Consolidation is strict both-orders as defined in Section 3.3, a pair is decisive only when both presentation orders name the same side, and incomplete pairs are excluded from n (five single-order private-corpus pairs were dropped, reproducing the published counts exactly). The panel rule is majority of decisive per-judge verdicts with a quorum of two judges; the decisive-verdict provenance over the pooled frontier-versus-8B cell splits 51 unanimous, 59 two-to-one, 66 single-decisive, and 5 two-judge. The clean-judge protocol places the full retrieved context in the evidence window (clean judge: Gemini-3-Flash, temperature 0.2) and appends the following de-verbosity rubric line, quoted verbatim from the artifact: "IMPORTANT - judge on correctness and grounding, not elaboration. A longer or more detailed answer is NOT better unless the extra content is (a) asked for by the question and (b) correct and supported by the evidence. If both candidates correctly answer what the question asks and are supported by the evidence, return TIE even if one is longer or adds unrequested detail. Prefer a candidate only when it is more correct, covers more of what the question specifically asks, or is better grounded in the evidence - never for verbosity alone." Decoding used max_tokens of 512 (DeepSeek arm regenerated at 4,096; Section 8) with a temperature-drop-on-error fallback, disclosed in Section 3.3.

D Judge case studies

The mistral judge prefers the first-presented answer in 73.1 percent of decisive raw verdicts (pure position-flip rate 17.8 percent against 1.2 percent for gemini and 9.4 percent for kimi), and strict both-orders consolidation contains the bias entirely, its panel contribution moving the pooled estimate by under 0.035. The evidence-window experiment re-judged 120 sampled pairs with the full context, producing the flip matrix with rows capped and columns full-evidence of (26, 1, 2), (0, 1, 0), and (20, 0, 70), agreement 0.808, with the 20 to 0 tie-to-frontier flips significant at p = 1.9e-6 and 14 to 15 of the 20 classified as verbosity rewards on independent re-reading [F15].

E Abstention classifier

The classifier is a deterministic regex rule set over answer text with no model calls, assigning each answer to exactly one of pure-abstain, hedged, or substantive. Rule 1 detects abstention sentences (refusal and missing-context patterns), Rule 2 detects offer-to-help filler, and Rule 3 detects hedge bridges into general knowledge; the decision uses a priori thresholds of 12 residual words with a hedge bridge or 30 without [F05]. Threshold sensitivity is reported as bands, Claude's pure-abstain rate spans 67.8 to 80.3 percent over the threshold sweep, and the full rule text ships verbatim in the receipt. The within-model application of Section 7 extends the abstention-sentence patterns with local-model phrasings the frontier-tuned rules missed; the extended rule set ships in the artifact alongside the original [F12].

F ROUGE triangulation and arena

Judge-free ROUGE-L reproduces the configuration ordering, retrieval configurations above non-retrieval, on all 9 corpus-model rows (private Llama row 0.223, 0.480, 0.172, 0.559 for base, rag, lora, lora_rag), with repetition collapse load-bearing for the MiniCPM rows [F09]. In the gold-anchored Bradley-Terry arena the gold answer dominates all six rankings, and three never-beaten anchors are non-convergent maximum-likelihood estimates reported as lower bounds at 400 iterations.

G Hit/miss stratification

On the private corpus the pooled frontier-versus-8B cell splits into a HIT stratum (86 questions, 258 pairs, net 0.607 [0.568, 0.649]) and a MISS stratum (258 questions, 772 pairs, net 0.597 [0.574, 0.621]); on QASPER the strata give 0.616 [0.537, 0.694] and 0.559 [0.529, 0.590]. All four HIT-minus-MISS deltas have intervals containing zero [F11].

H Reproducibility and cost ledger

Every quantitative claim in this paper maps to a receipt document and an idempotent recompute script that re-derives the claim from raw model outputs offline; the validation campaign re-verified 186 claims, finding 132 exact matches, 25 corrections (none overturning a headline conclusion; the one directional error, an informal abstention spot-check, is disclosed in Section 7), 22 new results, and 7 retired claims. The campaign's design is worth stating, each claim carries a verdict (match, corrected, new, or retired), a coverage matrix sweeps every results document so no quantitative claim goes unaudited, and intermediate JSONs are treated as comparison targets, never as sources; the same discipline caught the sign-backwards abstention spot-check, two stale decomposition rows, and the cost-logging artifact below. The honest API cost ledger totals approximately $45 through the 2026-06 validation campaign, with the 2026-07 addenda adding $16.94 for the 3.8 billion parameter frontier arm [F20] plus the whole-corpus long-context pilot; the first frontier pilot logged $0.95 nominally but cost approximately $14.7 once resumed input tokens and correct judge pricing are counted, a logging artifact we disclose [F14].