

When Retrieval Similarity Hurts: A Diagnostic Framework for Document Weighting in Black-Box RAG

Dongju Seo Beakcheol Jang*

Graduate School of Information, Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
dongju_allright@yonsei.ac.kr bjang@yonsei.ac.kr

Abstract

Document weighting is an underspecified bottleneck in black-box retrieval-augmented generation. Training-free systems inherit answer-aggregation weights from retrieval similarity, yet this proxy has not been systematically compared with evidence-based alternatives. We propose a unified two-parameter framework that generalizes REPLUG with an evidence term; it is training-free, black-box, and needs no large language model (LLM) calls beyond voting. Similarity-only ($\lambda=0$) and evidence-only ($\lambda\rightarrow\infty$) weighting are its limiting cases, so one interpolation parameter separates the retrieval prior from the evidence signal. We evaluate across 3 LLMs (7B–70B), 3 open-domain question-answering datasets (31,861 queries), and 3 off-the-shelf evidence types. Cross-encoder and embedding-stability evidence improve exact match in all 9 LLM-dataset cells, whereas natural language inference (NLI) fails as a weighting signal, to our knowledge the first systematic evidence of this failure mode. The similarity prior is largely redundant on NQ-Open and PopQA but helps on TriviaQA, where retriever scores track gold documents better; the pattern replicates under the E5 retriever, with smaller gains concentrated on long-tail PopQA. One default transfers across datasets and scales with minimal loss; calibration depends on signal type, not scale. Practically: prefer cross-encoder evidence, avoid NLI, and use vote mass for selective prediction; weighting itself adds sub-millisecond overhead.

1 Introduction

Retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2023) has become the de facto approach for grounding large language model (LLM) outputs in external knowledge and mitigating hallucinations, yet a fundamental question remains underexplored: how should retrieved

Method	Train-free	Black-box, no tok. prob	Soft ans.-level wts	Prior/evid. separation
REPLUG	✓	×	×	×
Cross-encoder rerank	✓	✓	×	×
Multi-agent	✓	✓	×	×
Trained reranker	×	✓	×	×
Ours	✓	✓	✓	✓

Table 1: Method comparison (columns: training-free; black-box, no token probabilities; soft answer-level weights; prior/evidence separation). Among the compared methods, only ours combines all four properties; multi-agent methods additionally require multiple LLM calls per query (§2).

documents be weighted during answer aggregation? Not all retrieved documents contribute equally: some contain the answer while others are topically related but factually unhelpful or noisy.

The dominant approach inherits weights from the retrieval stage. REPLUG (Shi et al., 2024), a widely adopted baseline, applies softmax over retriever cosine similarity scores to mix per-document LLM token distributions, requiring access to output token probabilities. We adapt this formulation to the voting setting for a stricter black-box regime that uses only text outputs (§2). The underlying assumption is straightforward: documents ranked higher by the retriever should contribute more to the final answer. A growing line of work instead trains specialized document scorers or relies on multi-step LLM inference at test time (§2); such approaches typically require task-specific training, model updates, or extra LLM calls. Table 1 contrasts them with ours along four axes (training-free, black-box, soft answer-level weighting, and prior/evidence separation).

Two questions have received comparatively little attention. First, does retrieval similarity, a proxy for query-document relevance, serve as an effective signal for answer-level aggregation? Second, which off-the-shelf evidence signals, if

*Corresponding author

any, can improve document weighting without additional training? To the best of our knowledge, prior work has not systematically compared different off-the-shelf evidence types, namely cross-encoder relevance (CE), embedding stability (ES), and natural language inference (NLI), as document weights for RAG answer aggregation.

We propose a unified two-parameter weighting framework that generalizes REPLUG via an evidence term, with similarity weighting ($\lambda=0$) and evidence-only weighting ($\lambda\rightarrow\infty$) as limiting cases of a single λ -controlled interpolation. We conduct an empirical analysis spanning 3 LLMs from three model families (7B–70B), 3 open-domain question-answering (QA) datasets (31,861 queries), and 3 off-the-shelf evidence signal types (§4).

We focus on the per-document voting regime, whose motivating scenarios include multi-agent RAG and edge/on-device deployment (not evaluated here), and which natively supports uncertainty-aware selective prediction (evaluated directly). While concat-prompt attains higher exact match (EM) with larger LLMs (§6.6), voting’s vote-mass distribution yields a native confidence signal that concat-prompt lacks.

Our contribution is a diagnostic framework supported by systematic empirical validation, rather than a new weighting algorithm:

- **A diagnostic, not a new algorithm.** The unified framework nests REPLUG ($\lambda=0$) and evidence-only weighting ($\lambda\rightarrow\infty$) as limiting cases of a single interpolation, so λ cleanly attributes performance to the retrieval prior versus the evidence signal.
- **Signal hierarchy.** CE evidence improves EM in all 9 (LLM, dataset) cells (8/9 Bonferroni-significant) and ES in all 9, while off-the-shelf NLI yields no significant gains, to our knowledge the first systematic evidence of this failure mode (§5).
- **When the similarity prior is redundant.** Evidence-only weighting matches the combined formula on all 6 NQ-Open and PopQA cells; the prior helps only on TriviaQA (+0.70–0.95%p), the dataset with the highest retriever–gold correlation (§6.1).
- **Practical robustness.** A single default ($\beta=0.5, \lambda=30$) stays within 0.52%p of per-dataset optima and transfers across datasets

and scales (7B–70B); calibration depends on the signal type, not scale (§6).

2 Related Work

Scope. We study the evidence signal used to reweight documents within a fixed voting pipeline; our formula (Eq. 2, §3) targets the weighting step, not a new full RAG system.

Trained document scoring for RAG. Cross-encoder passage rerankers are the canonical trained document scorers in retrieval pipelines (Nogueira and Cho, 2019; Nogueira et al., 2020). We use such a model off-the-shelf and frozen as one evidence signal (§3); the CE-Rerank/CE-Top1 baselines (§5, App. T) quantify what rank-based use of the same model achieves. Specialized trained scoring includes InfoGain-RAG (Wang et al., 2025a) (+17.9%p EM on Natural Questions), RE-RAG (Kim and Lee, 2024), REAR (Wang et al., 2024), CRAG (Yan et al., 2024), Self-RAG (Asai et al., 2024), RankRAG (Yu et al., 2024), and FIT-RAG (Mao et al., 2025). Trained rerankers yield larger gains (+17.9%p vs. our +0.5–4.1%p) but require task-specific labels and model updates unavailable in black-box or low-resource deployments. We therefore do not position the framework as a competitor to these systems. Where labels and model updates are available, a trained scorer such as InfoGain-RAG reports gains several times larger than our best cell, and trained pipelines such as RankRAG and Self-RAG are the appropriate choice for raw EM. We instead characterize what a strictly training-free, off-the-shelf signal delivers at low cost (no extra LLM calls; about 20 ms per query for the CE pass and under 1 ms for the weighting). We also provide a diagnostic of when the retrieval prior itself is the problem, which trained pipelines do not offer.

Training-free RAG improvements. REPLUG (Shi et al., 2024) uses retriever similarity as softmax weights for token-level probability mixing: our $\lambda=0$ special case, adapted to voting. TokRAG (Xu et al., 2025), DVD (Jin et al., 2024), and Entropy-Based Decoding (Qiu et al., 2025) perform token-level harmonization or per-document logit mixing, all requiring white-box access. SuRe (Kim et al., 2024), MAIN-RAG (Chang et al., 2025), and Speculative RAG (Wang et al., 2025b) require multiple LLM calls or training. Concurrent work VOTE-RAG (Xie and Sun, 2026) and

LCR (Song et al., 2026) also rely on multiple LLM calls. eRAG (Salemi and Zamani, 2024) is an offline retrieval-evaluation metric (with labels), whereas we use per-document answers at inference without labels. Our approach combines four properties (Table 1): no training; black-box compatibility, using no token probabilities, attention, or hidden states (unlike DVD, CrAM (Deng et al., 2025), Tok-RAG); no LLM calls beyond the per-document voting pipeline (unlike SuRe, MAIN-RAG, Speculative RAG, VOTE-RAG, LCR); and a single λ -parameterized formula generalizing REPLUG-style weighting. To our knowledge, it is the first training-free diagnostic disentangling retrieval-prior from evidence-signal contributions across three off-the-shelf signal types.

RAG calibration and robustness. CalibRAG (Jang et al., 2025), Reliability-Aware RAG (RA-RAG) (Hwang et al., 2025), CrAM (Deng et al., 2025), and Bayesian RAG (Ngartera et al., 2026) target related calibration and robustness goals. Their mechanisms differ from ours: trained hidden-state forecasting (CalibRAG), per-source reliability estimation (RA-RAG), white-box attention modification (CrAM), and a Monte Carlo dropout variance penalty (Bayesian RAG) whose subtractive form does not recover REPLUG.

3 Background and Method

3.1 Document Weighting in RAG

In per-document RAG with weighted majority voting, each retrieved document d_i independently conditions the LLM to generate an answer $a_i = \text{LLM}(x, d_i)$, and the final answer is selected by weighted vote:

$$\hat{y} = \arg \max_y \sum_{i: a_i=y} w_i \quad (1)$$

where w_i is the weight assigned to document d_i . The choice of w_i determines how much each document influences the final answer.

3.2 Unified Weighting Framework

To compare retrieval-similarity and evidence signals on equal footing, we propose a two-parameter family of weights that nests both regimes:

$$w_i = \frac{\exp(\beta \cdot s_i) + \lambda \cdot e_i}{\sum_{j=1}^k (\exp(\beta \cdot s_j) + \lambda \cdot e_j)} \quad (2)$$

where s_i is the retriever similarity score (cosine similarity), $e_i \in [0, 1]$ is the evidence score, β controls the prior strength, and λ controls the evidence

strength. Unlike the hard document selection used in reranking, the framework assigns graded weights, interpolating continuously between REPLUG and evidence-only weighting; this lets us attribute EM and calibration changes to the similarity prior, the evidence signal, or their interaction.

The endpoints recover established approaches. At $\lambda=0$ the weights reduce to $\text{softmax}(\beta \cdot s)$, equivalent to REPLUG-style similarity weighting (the max-shift in the softmax cancels in the ratio). As $\lambda \rightarrow \infty$ the weights approach evidence-only weighting $w_i \propto e_i$. Intermediate λ lets both signals contribute, enabling controlled ablation of the prior.

Notation. We use short labels for configurations of Eq. 2: Naive ($w_i=1/k$), SimW ($\lambda=0$ at $\beta=0.5$), Dir-X (default $\beta=0.5, \lambda=30$ with evidence X), and EO-X (evidence-only, $\lambda \rightarrow \infty$), where $X \in \{\text{CE}, \text{ES}, \text{NLI}\}$. The Dir label comes from reading Eq. 2 as a Dirichlet posterior mean ($\exp(\beta s_i)$ as prior, λe_i as evidence pseudo-counts); we use it as a name, not as a claim of calibrated posterior semantics.

3.3 Evidence Signals

We compare three off-the-shelf evidence signals:

Cross-Encoder (CE). We use ms-marco-MiniLM-L-6-v2 (Reimers and Gurevych, 2019) to compute query-document relevance, applying sigmoid to the logit: $e_i = \sigma(\text{CE}(x, d_i))$.

Embedding Stability (ES). We measure how robust the query-document cosine similarity is to small Gaussian perturbations of the embeddings, combining stability with relevance. Let $\mathbf{q}$ and $\mathbf{d}_i$ be the sentence-transformer embeddings of query x and document d_i (all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), 384-dim, document truncated to 512 characters). We sample $n=10$ i.i.d. perturbations $\epsilon_q^{(j)}, \epsilon_d^{(j)} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ with $\sigma=0.01$, compute similarities $c_i^{(j)} = \cos(\mathbf{q} + \epsilon_q^{(j)}, \mathbf{d}_i + \epsilon_d^{(j)})$, and define

$$e_i^{\text{ES}} = \frac{1}{1 + 10 \text{std}_j(c_i^{(j)})} \cdot \max\left(0, \text{mean}_j(c_i^{(j)})\right), \quad (3)$$

clipped to $[0, 1]$. In Eq. 3, the first factor rewards low variance under perturbation (stability); the second factor retains only documents with positive mean similarity (relevance). ES is a heuristic

score that weights relevance by embedding stability, not a calibrated Bayesian posterior.

NLI. We use `nli-deberta-v3-base` (He et al., 2023; Williams et al., 2018) with $e_i = P(\text{entailment}) + 0.5 \cdot P(\text{neutral})$, pairing the query as premise and the document as hypothesis. Both premise–hypothesis directions yield a low gold-vs. non-gold area under the receiver operating characteristic (ROC) curve (AUC; 0.45–0.59), so our headline NLI-unreliability finding holds regardless of direction (App. C).

4 Experimental Setup

4.1 Datasets

We evaluate on three open-domain QA benchmarks differing in domain, difficulty, and retriever. **NQ-Open (NQ)** (Kwiatkowski et al., 2019): 8,757 queries from real Google searches (Wikipedia answers), retrieved via dense passage retrieval (DPR) (Karpukhin et al., 2020) with pre-computed cosine scores; gold-document ratio 21.9% (fraction of top-10 retrieved documents containing the gold answer). **TriviaQA (TQA)** (Joshi et al., 2017): 8,837 trivia-style queries (Wikipedia, DPR), with a higher gold ratio (35.3%) and stronger retriever–gold correlation ($\rho=0.30$). **PopQA** (Mallen et al., 2023): 14,267 long-tail-entity queries retrieved via Contriever (Izacard et al., 2022); at $k=10$, 608 queries retrieve <10 docs and fall back to $k=8-9$ (strict- k details in App. F, P). Total: 31,861 queries $\times$ top-10 documents ($>310,000$ pairs).

4.2 Language Models

We use three LLMs spanning a 10-fold parameter range and three families (open-source and proprietary): **Qwen2.5-7B-Instruct-Turbo** (Yang et al., 2024) (Alibaba), **GPT-4.1-mini** (OpenAI, undisclosed size), and **Llama-3.3-70B-Instruct-Turbo** (Grattafiori et al., 2024) (Meta); the two open models were run using the Together AI platform. Each LLM answers each document independently with the same prompt template and greedy decoding (temperature=0, max_tokens=50).

4.3 Baselines

We compare against six baselines spanning the full performance range. **Naive** uses uniform weights $w_i=1/k$. **SimW** is REPLUG-style $\text{softmax}(\beta \cdot s)$ at $\beta=0.5$ ($\lambda=0$ in Eq. 2; β is varied over $\{0.5, 1.0, 2.0, 4.0\}$ only in §6). **Evidence-only**

uses $w_i = e_i / \sum_j e_j$, the $\lambda \rightarrow \infty$ limit. **CE-Rerank** hard-selects the top-5 documents by CE score with uniform weighting (approximately halving the input, as in common reranking pipelines). **Generation Oracle** marks a query solved if at least one of the 10 per-document answers matches gold. It is a generation (not retrieval) upper bound for any voting-based aggregation. Since a per-document answer may draw on the LLM’s parametric knowledge rather than the document itself, this bound is aggregation headroom, not retrieval quality. **Random** draws Dirichlet(1, ..., 1) weights averaged over 100 trials with fixed seed (performance floor).

4.4 Evaluation

We report EM as the primary metric and F1 as the secondary accuracy metric. For selective prediction, we report the area under the ROC curve of confidence against per-query correctness (AUROC), the area under the risk–coverage curve (AURC), and Risk@Coverage at 0.8 and 0.9. For calibration, we report Expected Calibration Error (ECE) and SmoothECE (Błasiok and Nakkiran, 2024) (kernel-based). Statistical significance is assessed via McNemar’s test (McNemar, 1947) with per-claim-family Bonferroni correction ($\alpha/27$, covering the 9 (LLM, dataset) cells $\times$ 3 signal methods = 27 comparisons per headline claim; the cross-claim family-wise error rate is not further controlled). The $\alpha/27$ divisor is applied uniformly (conservative for positive findings); null findings are additionally checked against their family-specific divisor (App. R). Three-seed repeated-split stability is reported in Appendix E.

All main results use a fixed default $\beta=0.5, \lambda=30$ (half-split calibration supports $\beta=0.5$; $\lambda=30$ sits on the plateau of Figure 2; Appendix K). The overfitting bound vs. per-cell grid optima is $\leq 0.52\%$ (§6), comparable to the smallest effect in Table 2 and below all others. Cross-dataset transfer and 3-seed stability are verified post-hoc (Appendices B, E).

5 Results

5.1 Main Results

Table 2 and Figure 1 present the main results using the default configuration ($\beta=0.5, \lambda=30$) for cross-encoder evidence. 8 out of 9 improvements are statistically significant (McNemar $p <$

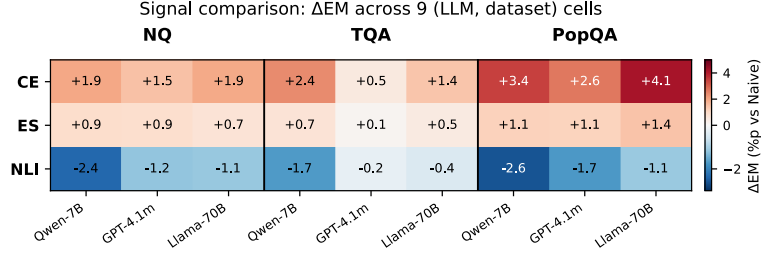


Figure 1: Signal comparison: ΔEM (%p vs Naive) across 9 (LLM, dataset) cells at the default configuration ($\beta=0.5, \lambda=30$). CE improves EM in all 9 cells; ES gives smaller positive gains; NLI consistently degrades EM.

0.001, Bonferroni-corrected for 27 comparisons). The only exception is TriviaQA with GPT-4.1-mini ($p=0.058$ raw, above the Bonferroni threshold $\alpha/27 \approx 0.0019$), where both Naive and Dir-CE are already near the per-LLM ceiling. Full per-cell McNemar statistics are in Appendix D.

Cross-encoder evidence improves EM in all 9 conditions (3 LLMs $\times$ 3 datasets; 8/9 Bonferroni-significant), from +0.5%p (GPT-4.1-mini on TQA) to +4.1%p (Llama-70B on PopQA). First, improvements are largest on PopQA (+2.6 to +4.1%p, the largest single gain being on Llama-70B), where long-tail entities make retrieval noisier and evidence-based filtering more valuable. Second, similarity weighting (SimW) underperforms Naive on NQ for all three LLMs: retriever similarity is at best neutral and at worst counter-productive as a voting signal (§6).

Across all 9 cells, Dir-CE recovers on average 11.6% of the Generation Oracle headroom (range 4.0–18.3%; C_4 in Appendix O).

These improvements are robust to the retriever: replacing the dense retrievers (DPR/Contriever) with sparse BM25 reproduces the CE gain on all three datasets (+1.2 to +2.0%p EM; Appendix Q).

5.2 Isolating the Weighting Stage

Aggregate EM conflates retrieval, weighting, and generation failures. To isolate the weighting stage, we partition each cell’s queries by whether the top-10 contains a gold document under the dataset-native annotation (has-gold: 72.7% / 75.3% / 62.5% of NQ / TQA / PopQA; the split is retriever-determined, hence identical across LLMs). We then re-run the McNemar analysis per subset (Table 3; full statistics in Appendix R).

The gain grows on the has-gold subset (+2.19 $\rightarrow$ +3.39%p average), positive in all nine cells (eight robustly significant; TQA/GPT-4.1-mini

marginal: the one cell not significant on the full test set (Appendix D), moved to the boundary once retrieval-failure queries are removed). On the no-gold subset the effect vanishes (−0.38%p; no cell significant in either direction at $\alpha/27$): reweighting cannot recover evidence that was never retrieved, and the small negative drifts are consistent with down-weighting documents that elicit correct parametric answers. The decomposition thus locates the CE gain at the weighting stage.

5.3 Evidence Type Hierarchy

Table 4 compares the three evidence types. CE achieves the highest discrimination AUC (0.728–0.776) and the largest EM gains (+2.2%p average). However, the relationship between AUC and EM improvement is not monotonic: TriviaQA has the highest CE AUC (0.776) but the smallest average gain (+1.42%p), whereas PopQA has the lowest AUC (0.728) yet the largest gain (+3.40%p). Raw discrimination captures only part of the picture; the benefit of reweighting also depends on how often the retriever alone is insufficient, which is more frequent for PopQA’s long-tail entities.

ES provides modest but consistent improvements (+0.82%p average), positive in all 9 cells (8/9 Bonferroni-significant; App. D). It is a viable alternative when cross-encoder inference is unavailable, though its higher computational cost (~ 0.16 s/query vs. CE’s ~ 0.02 s/query) makes it less attractive in practice.

In this setting, NLI reweighting is not only ineffective but tends to reduce EM: at the default configuration, the average change is −1.38%p across all 9 (LLM, dataset) cells, with 0/9 showing a Bonferroni-significant improvement and 7/9 showing a significant decrease (per-cell McNemar statistics: Table 14, Appendix C). We analyze this in §6.1.

Soft weighting vs. hard selection. Across 9

Dataset	LLM	Naive		SimW		Dir-CE		Δ EM
		EM	F1	EM	F1	EM	F1	
NQ	Qwen-7B	28.4 $\pm$ 0.37	38.9 $\pm$ 0.39	27.7 $\pm$ 0.49	37.8 $\pm$ 0.55	30.4 $\pm$ 0.38	41.8 $\pm$ 0.33	+1.9
	GPT-4.1-mini	32.4 $\pm$ 0.15	46.8 $\pm$ 0.01	32.1 $\pm$ 0.12	46.3 $\pm$ 0.05	33.9 $\pm$ 0.14	48.6 $\pm$ 0.16	+1.5
	Llama-70B	28.1 $\pm$ 0.27	41.6 $\pm$ 0.17	27.7 $\pm$ 0.18	40.8 $\pm$ 0.09	30.0 $\pm$ 0.24	44.5 $\pm$ 0.13	+1.9
TQA	Qwen-7B	52.4 $\pm$ 0.58	60.8 $\pm$ 0.49	52.2 $\pm$ 0.49	60.6 $\pm$ 0.40	54.7 $\pm$ 0.57	64.3 $\pm$ 0.59	+2.4
	GPT-4.1-mini	62.9 $\pm$ 0.21	73.2 $\pm$ 0.40	63.1 $\pm$ 0.13	73.3 $\pm$ 0.30	63.4 $\pm$ 0.21	74.1 $\pm$ 0.26	+0.5
	Llama-70B	60.1 $\pm$ 0.09	71.4 $\pm$ 0.21	60.1 $\pm$ 0.21	71.4 $\pm$ 0.23	61.5 $\pm$ 0.15	73.5 $\pm$ 0.18	+1.4
PopQA	Qwen-7B	28.4 $\pm$ 0.34	33.3 $\pm$ 0.32	28.4 $\pm$ 0.40	33.4 $\pm$ 0.38	31.8 $\pm$ 0.18	37.5 $\pm$ 0.09	+3.4
	GPT-4.1-mini	31.8 $\pm$ 0.28	37.4 $\pm$ 0.24	31.9 $\pm$ 0.24	37.4 $\pm$ 0.23	34.5 $\pm$ 0.26	40.8 $\pm$ 0.32	+2.6
	Llama-70B	26.1 $\pm$ 0.06	30.3 $\pm$ 0.16	26.2 $\pm$ 0.12	30.3 $\pm$ 0.20	30.2 $\pm$ 0.20	35.9 $\pm$ 0.18	+4.1

Table 2: Main results with the fixed default configuration ($\beta=0.5$, $\lambda=30$, CE evidence) applied uniformly across all cells. SimW = REPLUG-style similarity weighting ($\beta=0.5$). Bold = highest EM and F1 across Naive/SimW/Dir-CE in each row. EM and F1 are percentages rounded to one decimal place; Δ EM is the raw difference rounded to one decimal place and may differ from the difference of the displayed values by 0.1%p. All point estimates use the fixed default on the full test set; $\pm$ std is a separate 3-seed 50/50 cal/test repeated-split check (seeds 42, 123, 456; Appendix E). With $n=3$ it carries wide uncertainty, so all significance claims rely on McNemar’s test on the full test set (8/9 cells; see text), not on $\pm$ std.

Dataset	Full	Has-gold	No-gold
NQ	+1.75	+2.48	−0.17
TQA	+1.42	+2.11	−0.70
PopQA	+3.40	+5.59	−0.26
Average	+2.19	+3.39	−0.38

Table 3: Δ EM (Dir-CE − Naive, %p) by retrieval outcome, averaged over the 3 LLMs. Has-gold is positive in 9/9 cells: eight robustly Bonferroni-significant at $\alpha/27$, and TQA/GPT-4.1-mini passes numerically ($p=1.5\times 10^{-3}$ vs. 1.9×10^{-3}) but is reported as marginal; none of the no-gold cells is significant. Per-cell table and family-size sensitivity: App. R. Full-set values recomputed under strict $k=10$ (within 0.1%p of Table 2).

conditions, soft evidence weighting wins in 7 over hard top-5 CE reranking (margin up to +2.08%p), ties in 1 (NQ / GPT-4.1-mini, −0.06%p within noise), and loses in 1 (Llama-70B on PopQA, −0.64%p). Continuous weights preserve soft relevance information that binary selection discards.

Signal vs. functional form. We compare against three further training-free rules on the same cached answers (Appendix T). Dir-CE beats reciprocal rank fusion (RRF) (Cormack et al., 2009) of the similarity and CE ranks in all 9 cells (avg +1.71%p; 8/9 McNemar-significant), ties Borda voting over CE ranks (−0.16%p; no cell significant, nominal edge to Borda in four cells), and splits 5–4 with CE-Top1 selection (PopQA favors selection). Two functional forms over the same CE signal perform equivalently, while fusion re-admitting the similarity prior loses consistently:

Evid.	AUC			Avg AUC	Δ EM (avg)
	NQ	TQA	PopQA		
CE	.738	.776	.728	.747	+2.19
ES	.715	.734	.724	.724	+0.82
NLI	.504	.512	.502	.506	−1.38

Table 4: Evidence discrimination (AUC: gold vs. non-gold documents) and average EM improvement across 3 LLMs at default ($\beta=0.5$, $\lambda=30$). NLI AUC is reported on a 500-query per-dataset sample (App. C).

Method	EM	AURC↓	R@.8↓	R@.9↓
Gen-Oracle	46.6	—	—	—
Dir-CE	30.0	.539	.654	.679
EO-CE	30.0	.551	.656	.679
CE-Rerank	29.9	.541	.652	.680
Naive	28.1	.564	.678	.700
SimW	27.7	.560	.677	.698
Random	24.5	—	—	—

Table 5: Baseline comparison (Llama-70B, NQ, CE). R@.8 = Risk at 80% coverage (↓ = lower is better); ties and near-ties are read off in the text. Gen-Oracle = Generation Oracle (§4). Per-cell AUROC and AURC: Table 10.

the evidence signal matters more than the form incorporating it.

5.4 Selective Prediction

Beyond accuracy, evidence-based weighting also improves selective prediction, the ability to abstain on likely-incorrect answers while maintaining coverage.

Table 5 reports a full baseline comparison on one representative cell (Llama-70B, NQ);

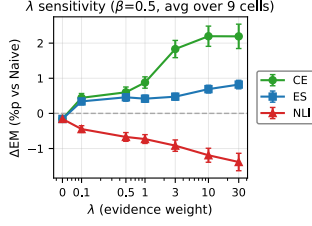


Figure 2: λ sensitivity at $\beta=0.5$ (average of 9 cells). CE plateaus at $\lambda \geq 10$; ES stays positive but modest; NLI stays below zero. Error bars: standard errors.

among the voting-based methods Dir-CE shows the lowest AURC, ties EO-CE on EM (30.0) and R@.9 (.679), and ties CE-Rerank on R@.8 (within noise). The pattern transfers across cells (e.g., Llama-70B/PopQA Risk@0.8: 0.704 $\rightarrow$ 0.654); per-LLM EM/F1 and the 9-cell selective-prediction table are in Appendix A (Table 10).

6 What the Framework Reveals

We use the framework’s λ parameterization to isolate why the patterns in §5 arise: when the retrieval prior helps, when it is redundant, and when evidence alone suffices.

6.1 Evidence-Only Sufficiency

Evidence-only weighting (without the retrieval-similarity prior) achieves near-identical EM to the full formula on NQ and PopQA (gap $\leq 0.2\%$ p in all 6 cells), though not as identical classifiers. On NQ with Qwen-7B they disagree on 543 predictions and flip correctness on 140, with wins and losses approximately balanced.

Practically, on NQ and PopQA the retriever score can be skipped using evidence-only weighting ($w_i \propto e_i$) with negligible loss. The exception is TriviaQA, where the combined formula wins by +0.70 to +0.95%p across LLMs, plausibly because TriviaQA has a higher retriever-gold-document correlation (Spearman ρ between the retriever score and the binary gold-document label over all query-document pairs: 0.30 vs. NQ 0.20, PopQA 0.29) and gold-document ratio (35.3% vs. 21.9–22.7%). When the retriever tracks answer utility its scores add complementary signal; otherwise they add noise. This is consistent with SimW underperforming Naive on NQ, where DPR’s topical relevance does not align with answer utility (McNemar: SimW is significantly worse than Naive in 2 of the 3 NQ cells at $\alpha/27$ and never significantly better in any of the 9

cells; App. D). This redundancy is specific to the evidence-dominant default: at $\lambda \leq 1$ the prior is active, widening the Dir-CE vs. EO-CE gap to up to 3.8%p (Appendix L).

Why does NLI underperform? When used as a document-weighting signal, the NLI score ($P(\text{entailment}) + 0.5 \cdot P(\text{neutral})$) tends to reduce EM rather than improve it. At the default configuration ($\beta=0.5$, $\lambda=30$), the average EM change is -1.38% p across the 9 (LLM, dataset) cells, with 0/9 Bonferroni-significant improvements and 7/9 Bonferroni-significant decreases. Across the full 24-point (β, λ) grid ($24 \times 9 = 216$ runs), no NLI configuration we tested yields a Bonferroni-significant improvement in any cell, while 131 of the 216 runs are Bonferroni-significant decreases (Appendix C). A plausible explanation is a train-test mismatch: NLI models judge strict entailment between short premise-hypothesis pairs, whereas RAG documents are long passages in which the answer-bearing span occupies only a small fraction of the text (Liu et al., 2024), diluting the entailment signal. This remains a plausibility argument rather than a mechanistic verification; our contribution is to quantify, under a unified weighting framework, how strongly the failure mode propagates to downstream-task EM across all 216 evaluations.

Supporting diagnostics are in Appendices C (direction, McNemar) and M (roberta-large-mnli, still low-AUC); off-the-shelf NLI does not appear to justify its extra cost (~ 0.11 s/query) as a document-weighting signal in this setting. Two qualitative case studies (a CE success and an NLI failure on NQ with Qwen-7B) are in Appendix H.

6.2 A Modern High-Precision Retriever

Do these findings survive a modern retriever? We re-encoded the full 21M-passage corpus with e5-base-v2 (Wang et al., 2022) and re-ran all nine cells (App. S). Dir-CE still improves EM in 9/9 cells, but the average gain shrinks ($+2.19 \rightarrow +0.71\%$ p; on corpus-comparable PopQA, $+3.40 \rightarrow +1.63\%$ p) and significance concentrates where retrieval remains hardest (the three PopQA cells; largest $+2.14\%$ p on Llama-70B). The similarity-only prior becomes fully redundant (SimW – Naive $< 0.1\%$ p in every cell), and a (β, λ) grid finds the fixed default within 0.41%p of every per-cell optimum ($\lambda \geq 10$ optimal throughout). As Appendix Q anticipated, E5 retrieves an answer-

Source $\rightarrow$	NQ	TQA	PopQA
NQ	30.0	61.5	30.2
TQA	29.5	61.9	29.7
PopQA	30.0	61.5	30.2
<i>Grid Oracle</i>	30.0	61.9	30.2
<i>Max loss</i>		0.53%p	

Table 6: Cross-dataset transfer (Llama-70B, CE). The source dataset’s best (β, λ) is applied to each target. EM in percent. *Grid Oracle* = the per-dataset best EM achievable within the tested (β, λ) grid (distinct from the *Generation Oracle* in Table 5).

bearing document for 70–81% of queries with a compressed similarity spread (≈ 0.03), so weighting choices matter little where retrieval is already strong (NQ/TQA) and still pay in the long-tail regime (PopQA). The diagnostic’s value is identifying when weighting matters.

6.3 Hyperparameter Robustness

We examine sensitivity and transferability in two ways.

Default configuration. A single setting ($\beta=0.5$, $\lambda=30$) achieves grid-best or near-grid-best performance: 5 of the 9 conditions match the per-dataset best (to two decimal places; 6 of 9 at the displayed one-decimal precision), and the maximum loss is 0.52%p. The same robustness replicates on E5 (gap $\leq 0.41\%$ p in all 9 cells; App. S). Practitioners can adopt this default on similar English open-domain factoid QA benchmarks, with the caveat that it was informed by tuning on these datasets, so it is not fully dataset-independent; the transfer results below bound, but do not eliminate, this dependence.

Cross-dataset transfer. We take the best (β, λ) from one dataset and apply it unchanged to others (Table 6; all three LLMs in Appendix B). Here “best” is selected on each source dataset’s full test set, so diagonal entries are upper bounds while off-diagonal entries represent the realistic transfer metric (held-out split stability: Appendix E). The maximum off-diagonal transfer loss is 0.53%p for Llama-70B (0.69%p Qwen-7B, 0.52%p GPT-4.1-mini). NQ and PopQA share near-identical optimal configurations ($\beta=0.5$, $\lambda=30$ for two of three LLMs).

6.4 Accuracy–Calibration Trade-off

Evidence reweighting can affect calibration. CE concentrates weights, reducing entropy

from $\log 10=2.303$ to 2.073/1.989/2.087 on NQ/TQA/PopQA ($-10.0/-13.6/-9.4\%$), sharpening the vote distribution (App. O). Under SmoothECE (Błasiok and Nakkiran, 2024), the Dir family wins in 10/18 (LLM, dataset, evidence) conditions: Dir-ES 8/9 cells, Dir-CE 2/9 (Dir-ES trades $\sim 1-2\%$ p EM for better SmoothECE). The signal type, CE (accuracy) vs. ES (calibration), dominates over scale. Binned ECE (App. G) and Brier decomposition (App. I) corroborate.

Efficiency note. On a cost-normalized basis (C_5 , Appendix O), CE is the only signal whose EM gain justifies its inference latency. CE yields $+0.110 \Delta\text{EM}/\text{ms}$ at 20 ms, ES $+0.005$ at 160 ms ($\sim 21\times$ less per ms), and NLI -0.013 at 110 ms, actively harmful.

6.5 Effect of Retrieval Depth (k)

The benefit of evidence weighting grows monotonically with k : averaged over the 9 cells, ΔEM rises from $+1.4\%$ p at $k=3$ to $+1.5$ at $k=5$ and $+2.2$ at $k=10$ (Table 16, Appendix F), as CE filters more non-gold documents at larger k . Figure 2 shows a broad λ plateau.

6.6 Voting vs. Concat-Prompt

We benchmark against the dominant concat-prompt paradigm, in which the top- k documents are concatenated into a single LLM call. Concat wins EM in 6/9 cells (avg $+2.09\%$ p over Dir-CE, range -3.00 to $+11.00\%$ p; Appendix P, with PopQA voting metrics on the strict $k=10$ subset for fair comparison). Three observations:

(i) **The advantage is concentrated on the largest model.** Concat – Dir-CE averages -1.27% p on Qwen-7B (voting wins 2/3), -0.09% p on GPT-4.1-mini (mixed), and $+7.64\%$ p on Llama-70B (concat wins 3/3). This is consistent with the documented positional degradation of long-context inputs (Liu et al., 2024); the size dependence itself is an empirical observation across our 9 cells, not a general scaling claim.

(ii) **Voting offers complementary value beyond EM.** The vote-mass distribution is a native confidence signal: even on Llama-70B/NQ where concat wins EM (37.59 vs. 30.01), Dir-CE attains the lowest AURC (.539) and tied-best Risk@0.9 (.679) among voting methods (Table 5). Concat produces a single output, lacking graded confidence without auxiliary scoring.

(iii) Complementary regimes, plus motivating scenarios. Our evidence supports a scoping rather than a winner. Concat-prompt is preferable for raw EM with large models. Evidence-weighted voting is preferable when calibrated abstention is needed: Dir-CE has the lowest mean AURC of the three voting rules in all 9 cells (Table 10). It is also preferable when the LLM is small: on Qwen-7B, the smallest open-weights model tested, voting beats concat on 2/3 datasets (avg +1.27%p). Quantized/on-device deployment and multi-agent RAG systems (SuRe, MAIN-RAG, etc.; §2), where per-document generation is the native mechanism, are motivating scenarios for this regime, not evaluated deployments. The framework does not claim to outperform concat-prompt generally; it gives the per-document voting regime a unified evidence-based weighting interface and a λ -controlled diagnostic.

7 Conclusion

We proposed a unified two-parameter framework generalizing REPLUG via an evidence term (training-free, black-box, low-overhead) and used it to diagnose which off-the-shelf signals help. Across 3 LLMs and 3 datasets, CE improves EM in all 9 cells (+0.5–4.1%p; 8/9 Bonferroni-significant, larger on has-gold), ES gives smaller gains, and NLI reduces EM. The retrieval prior is redundant on NQ/PopQA, on both the DPR/Contriever and the E5 retrieval sets, but helps on TriviaQA. The default ($\beta=0.5$, $\lambda=30$) stays within 0.52%p of grid optima across 7B–70B, and vote-mass yields a native selective-prediction signal (§6.6) that concat-prompt lacks. Practically: prefer CE, use ES for calibration, avoid NLI, omit the prior on NQ/PopQA, and test new evidence scores under the same formula. Future work: multi-hop QA and adaptive signals.

Limitations

Our study has several limitations. First, we evaluate only on open-domain factoid QA with extractive answers; our findings may not transfer to long-form generation, summarization, or multi-hop QA (e.g., HotpotQA). Second, our comparison with REPLUG uses similarity-weighted voting rather than its original token-level probability mixing, since our setting assumes text-only outputs (a stricter black-box than the no-fine-tuning regime REPLUG targets). Third, our evalua-

tion covers only English; multilingual settings and the BGE retriever remain untested. Beyond the two dense retrievers of the main experiments (DPR, Contriever), a BM25 check (Appendix Q) confirms the CE signal transfers across retriever families, and a full 9-cell E5 evaluation (Appendix S) confirms the predicted high-precision regime (gains positive in all cells but smaller, concentrated on long-tail PopQA). Fourth, the accuracy–calibration trade-off depends on the evidence signal rather than model scale (Dir-CE improves accuracy but often worsens SmoothECE; Dir-ES trades accuracy for better SmoothECE in 8/9 cells); we treat calibration as a diagnostic rather than jointly optimizing both. Fifth, we do not benchmark against multi-LLM-call training-free frameworks (SuRe, MAIN-RAG, Speculative RAG, VOTE-RAG, LCR), as a fair comparison would require controlling for prompt design, token budget, and per-method tuning; a controlled same-budget comparison is future work. We report instead a concat-prompt comparison (§6.6, Appendix P), in which concat wins EM in 6/9 cells (especially Llama-70B) while voting remains preferable on Qwen-7B and for selective prediction. We also report same-cost comparisons against training-free aggregation rules (RRF, Borda, CE-Top1; Appendix T). Sixth, several methodological choices remain to be refined. Evidence signals use different truncation lengths (CE/NLI 1,000, ES 512, LLM 800 chars), though an ES@1,000-char ablation (Appendix N) shows the confound is small ($|\Delta\text{AUC}| \leq 0.013$). NLI is passage-level rather than sentence-level or decomposed (Appendix C). Our evidence signals are off-the-shelf, not RAG-optimized. The default (β , λ) was selected on our three tuning datasets; cross-dataset transfer ($\leq 0.69\%$ loss) and the E5/BM25 grids bound this dependence, but the default is not guaranteed to be dataset-independent.

Ethics Statement

Data and models. This work uses publicly available QA datasets (NQ-Open, TriviaQA, PopQA) and pre-trained models (LLMs accessed via APIs; evidence models run locally). Artifact licenses are listed in Appendix J; all usage is for non-commercial academic research, consistent with each artifact’s stated intended use. No new data is collected; no human annotation or participant study is conducted.

Risk and dual use. Our method reweights per-document outputs of an already-deployed LLM; it does not increase the generative capability of any model and introduces no new training data. We therefore do not see an increased risk of harmful generation beyond the underlying LLM itself. One caveat is that evidence reweighting by construction amplifies documents scored high by the evidence signal: if the underlying retriever or cross-encoder systematically prefers a particular source distribution (e.g., Western-English Wikipedia), evidence weighting can reinforce that skew. Practitioners deploying evidence-weighted RAG in user-facing systems should audit the evidence model’s ranking behavior on their target domain.

Fairness and bias. The three datasets we use inherit known biases: NQ-Open and TriviaQA draw from English Wikipedia (Western/Anglophone skew), and PopQA over-represents entities with higher popularity scores even within its long-tail subset. We do not attempt demographic debiasing, and no subgroup analysis was conducted; the accuracy gains we report should be interpreted as conditional on these distributions rather than as evidence of fair performance across subpopulations.

Computational cost. We report computational costs explicitly: approximately \$300 in total across all experiments, comprising $\approx$ \$100 in LLM API costs for the main DPR/Contriever experiments plus $\approx$ \$200 for the E5 replication (App. S; $\approx$ \$10 of cloud-GPU time for corpus re-encoding, the remainder LLM API calls). Evidence computation adds at most 0.16 s per query; weight computation adds less than 1 ms per query. All local orchestration, evidence scoring, and analyses ran on a single consumer-grade laptop (36 GB RAM); LLM generation used provider-hosted APIs (Together AI and OpenAI); the E5 corpus re-encoding used one rented cloud GPU. We do not train or fine-tune any LLM. These figures reflect only the inference-time costs of our experiments and do not amortize the upstream training costs of Qwen2.5-7B, Llama-3.3-70B, or GPT-4.1-mini, which we do not control.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF), funded by the Korean government under Grant No. RS-2023-00273751.

References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. [Self-RAG: Learning to retrieve, generate, and critique through self-reflection](#). In *Proceedings of ICLR*.
- Jarosław Błasiok and Preetum Nakkiran. 2024. [Smooth ECE: Principled reliability diagrams via kernel smoothing](#). In *Proceedings of ICLR*.
- Glenn W. Brier. 1950. [Verification of forecasts expressed in terms of probability](#). *Monthly Weather Review*, 78(1):1–3.
- Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. 2025. [MAIN-RAG: Multi-agent filtering retrieval-augmented generation](#). In *Proceedings of ACL*.
- Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. [Reciprocal rank fusion outperforms Condorcet and individual rank learning methods](#). In *Proceedings of SIGIR*, pages 758–759.
- Boyi Deng, Wenjie Wang, Fengbin Zhu, Qifan Wang, and Fuli Feng. 2025. [CrAM: Credibility-aware attention modification in LLMs for combating misinformation in RAG](#). In *Proceedings of AAAI*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. [Retrieval-augmented generation for large language models: A survey](#). *arXiv preprint arXiv:2312.10997*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 540 others. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *Proceedings of ICLR*.
- Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. 2025. [Retrieval-augmented generation with estimation of source reliability](#). In *Proceedings of EMNLP*.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. [Unsupervised dense information retrieval with contrastive learning](#). *Transactions on Machine Learning Research*.
- Chaeyun Jang, Deukhwan Cho, Seanie Lee, Hyungi Lee, and Juho Lee. 2025. [Reliable decision-making via calibration-oriented retrieval-augmented generation](#). In *Proceedings of NeurIPS*.

- Jing Jin, Houfeng Wang, Hao Zhang, Xiaoguang Li, and Zhijiang Guo. 2024. [DVD: Dynamic contrastive decoding for knowledge amplification in multi-document question answering](#). In *Proceedings of EMNLP*.
- Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of ACL*.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of EMNLP*.
- Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024. [SuRe: Summarizing retrievals using answer candidates for open-domain QA of LLMs](#). In *Proceedings of ICLR*.
- Kiseung Kim and Jay-Yoon Lee. 2024. [RE-RAG: Improving open-domain QA performance and interpretability with relevance estimator in retrieval-augmented generation](#). In *Proceedings of EMNLP*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Proceedings of NeurIPS*.
- Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. [Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2356–2362.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [When not to trust language models: Investigating effectiveness of parametric and non-parametric memories](#). In *Proceedings of ACL*.
- Yuren Mao, Xuemei Dong, Wenyi Xu, Yunjun Gao, Bin Wei, and Ying Zhang. 2025. [FIT-RAG: Black-box RAG with factual information and token reduction](#). *ACM Transactions on Information Systems*, 43(2):40:1–40:27.
- Quinn McNemar. 1947. [Note on the sampling error of the difference between correlated proportions or percentages](#). *Psychometrika*, 12(2):153–157.
- Allan H. Murphy. 1973. [A new vector partition of the probability score](#). *Journal of Applied Meteorology*, 12(4):595–600.
- Lebede Ngartera, Saralees Nadarajah, and Rodoumta Koina. 2026. [Bayesian RAG: uncertainty-aware retrieval for reliable financial question answering](#). *Frontiers in Artificial Intelligence*, 8.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. [Passage re-ranking with BERT](#). *arXiv preprint arXiv:1901.04085*.
- Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. [Document ranking with a pre-trained sequence-to-sequence model](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.
- Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2025. [Entropy-based decoding for retrieval-augmented large language models](#). In *Proceedings of NAACL*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of EMNLP-IJCNLP*.
- Alireza Salemi and Hamed Zamani. 2024. [Evaluating retrieval quality in retrieval-augmented generation](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. [REPLUG: Retrieval-augmented black-box language models](#). In *Proceedings of NAACL*.
- Zhipeng Song, Xiangyu Kong, Xinrui Bao, Yizhi Zhou, Jiulong Jiao, Sitong Liu, Yuhang Zhou, and Heng Qi. 2026. [LLM-confidence reranker: A training-free approach for enhancing retrieval-augmented generation systems](#). *Expert Systems with Applications*, 314:131627.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. [Text embeddings by weakly-supervised contrastive pre-training](#). *arXiv preprint arXiv:2212.03533*.
- Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024. [REAR: A relevance-aware retrieval-augmented framework for](#)

open-domain question answering. In *Proceedings of EMNLP*.

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. 2025a. [InfoGain-RAG: Boosting retrieval-augmented generation through document information gain-based reranking and filtering](#). In *Proceedings of EMNLP*.

Zilong Wang, Zifeng Wang, Long Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2025b. [Speculative RAG: Enhancing retrieval augmented generation through drafting](#). In *Proceedings of ICLR*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of NAACL-HLT*.

Zequan Xie and Zhengyang Sun. 2026. [Mitigating hallucination on hallucination in RAG via ensemble voting](#). In *2026 29th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 273–278. IEEE.

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2025. [A theory for token-level harmonization in retrieval-augmented generation](#). In *Proceedings of ICLR*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. [Corrective retrieval augmented generation](#). *arXiv preprint arXiv:2401.15884*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jixia Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. [RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs](#). In *Proceedings of NeurIPS*.

A Per-LLM Detailed Results

Tables 7, 8, and 9 show full results for each LLM with default configuration ($\beta=0.5$, $\lambda=30$). NLI results are not shown here: at this default configuration, NLI’s ΔEM is negative in all 9 cells, and even at the most favorable (β, λ) setting we found, NLI does not produce a Bonferroni-significant improvement.

B Per-LLM Cross-Dataset Transfer

Table 6 (§6) reports the Llama-70B transfer matrix; Tables 11 and 12 extend it to Qwen-7B and GPT-4.1-mini. The source dataset’s best (β, λ) is applied to each target; EM values are in percent. Max loss across all 9 source→target combinations per LLM: 0.69%p (Qwen-7B), 0.52%p (GPT-4.1-mini), 0.53%p (Llama-70B).

C NLI Direction Comparison

To verify that our NLI direction choice does not drive the negative NLI finding, we compute gold-vs-non-gold AUC for both directions on a 500-query random sample (fixed seed 42) from each dataset. The forward direction (query as premise, document as hypothesis) matches our main-text choice; the reverse direction (document as premise, query as hypothesis) matches the standard NLI task formulation. Table 13 summarizes the comparison.

The pattern aligns with our NLI failure analysis in §6: long retrieval passages produce entailment/neutral probabilities that are poorly correlated with whether the passage actually contains the gold answer, in both directions. We use the forward direction in the main text as the more conservative of the two (lower per-dataset AUC on 2/3 datasets), to avoid overstating the NLI signal. All AUC values in the main text and appendices are computed as a single pooled ROC over the concatenation of (query, document) pairs across the dataset, rather than as a per-query average, matching the gold-vs-non-gold framing of the evidence-score classifier.

Per-cell McNemar statistics. Table 14 reports the paired test of Dir-NLI (default $\beta=0.5$, $\lambda=30$) against Naive in all nine cells, computed from the per-query correctness vectors of the main runs (full retrieval set, as in Table 15). Dir-NLI lowers EM in every cell; the decrease is Bonferroni-significant in 7/9 (the exceptions are the two larger LLMs on TQA). Sweeping all 24 (β, λ) configurations in every cell (216 runs) yields no Bonferroni-significant improvement and 131 significant decreases; the best NLI configuration per cell ranges from -0.72 to $+0.20\%$ over Naive (results/dir_nli_vs_naive_mcnemar.json).

DS	Ev	Naive		SimW		Dir		Δ EM
		EM	F1	EM	F1	EM	F1	
NQ	CE	28.4	38.9	27.7	37.8	30.4	41.8	+1.9
NQ	ES	28.4	38.9	27.7	37.7	29.3	40.1	+0.9
TQA	CE	52.4	60.8	52.2	60.6	54.7	64.3	+2.4
TQA	ES	52.4	60.7	52.2	60.6	53.0	61.7	+0.7
PopQA	CE	28.4	33.3	28.4	33.4	31.8	37.5	+3.4
PopQA	ES	28.4	33.3	28.4	33.4	29.5	34.6	+1.1

Table 7: Qwen2.5-7B full results (CE/ES, default config). CE improvements are Bonferroni-significant on all 3 datasets ($p < 0.001$).

DS	Ev	Naive		SimW		Dir		Δ EM
		EM	F1	EM	F1	EM	F1	
NQ	CE	32.4	46.8	32.1	46.3	33.9	48.6	+1.5
NQ	ES	32.4	46.8	32.1	46.3	33.3	47.7	+0.9
TQA	CE	62.9	73.2	63.1	73.3	63.4	74.1	+0.5
TQA	ES	63.0	73.2	63.1	73.3	63.1	73.4	+0.1
PopQA	CE	31.8	37.4	31.9	37.4	34.5	40.8	+2.6
PopQA	ES	31.8	37.4	31.9	37.4	32.9	38.6	+1.1

Table 8: GPT-4.1-mini full results (CE/ES, default config). CE improvement on TriviaQA is not Bonferroni-significant ($p=0.058$ raw, above $\alpha/27$); NQ and PopQA are significant at $p < 0.001$.

D McNemar Statistical Significance

Table 15 reports McNemar’s test results for Dir-CE (default config) vs. Naive, with Bonferroni correction for 27 comparisons ($\alpha/27 \approx 0.00185$).

SimW vs. Naive. Applying the same protocol to the similarity-only prior: SimW never significantly improves over Naive in any of the nine cells, and is significantly worse in two, both on NQ (Qwen-7B: $n_{01}=86$, $n_{10}=150$, $p=4.1 \times 10^{-5}$; Llama-70B: $40/78$, $p=6.6 \times 10^{-4}$; both at $\alpha/27$), with NQ/GPT-4.1-mini negative at nominal significance ($51/81$, $p=0.012$). All TQA/PopQA cells are statistically indistinguishable ($|\Delta\text{EM}| \leq 0.22\%$). This quantifies the regime in which similarity weighting hurts: the harm concentrates on NQ, where the DPR prior misaligns with answer utility (§6.1).

Dir-ES vs. Naive. Dir-ES improves over Naive with Bonferroni significance in 8/9 cells; the exception is TQA/GPT-4.1-mini ($p=0.41$), the same cell that is the CE exception. Both analyses use strict $k=10$ and are released with the code (results/simw-vs-naive_mcnemar.json, results/dir-es-vs-naive_mcnemar.json).

E Repeated Split Stability

We verify robustness to data splits using 3 random seeds (42, 123, 456) with 50/50 cal/test partitions. Standard deviations are computed as population std (`np.std` with `ddof=0`); the sample-std convention (`ddof=1`) would multiply the reported values by $\sqrt{n/(n-1)} \approx 1.225$. Across all 18 (LLM, dataset, evidence) cells, EM and F1 standard deviations stay below 0.6%p under `ddof=0` (max EM std 0.57%p, max F1 std 0.59%p; $\leq 0.72\%$ under `ddof=1`; the Naive column of Table 2, outside these 18 evidence-weighted configurations, reaches ± 0.58); AUROC and AURC fluctuate more (up to 0.0099 and 0.0093 in absolute terms), as expected for ranking-based metrics on smaller split sizes. For each seed, the best (β, λ) is selected on the calibration half and evaluated on the test half; in 11 of 18 cells this selection picks different (β, λ) across seeds, so the reported standard deviations combine split randomness and hyperparameter-selection variance. Primary accuracy conclusions remain stable across splits; effects near the standard-deviation magnitude should be interpreted with caution.

F Top-k Ablation Details

Table 16 reports the top- k ablation summarized in §6. Full per-LLM per-dataset top- k results are available in the released code (Appendix J).

PopQA $k=10$ subset. At $k=10$, PopQA

DS	Ev	Naive		SimW		Dir		Δ EM
		EM	F1	EM	F1	EM	F1	
NQ	CE	28.1	41.6	27.7	40.8	30.0	44.5	+1.9
NQ	ES	28.1	41.6	27.7	40.8	28.8	42.7	+0.7
TQA	CE	60.1	71.4	60.1	71.4	61.5	73.5	+1.4
TQA	ES	60.1	71.4	60.1	71.4	60.6	72.0	+0.5
PopQA	CE	26.1	30.3	26.2	30.3	30.2	35.9	+4.1
PopQA	ES	26.1	30.3	26.2	30.3	27.5	32.0	+1.4

Table 9: Llama-3.3-70B full results (CE/ES, default config). PopQA CE shows the largest improvement (+4.1%p).

LLM	DS	AUROC $\uparrow$			AURC $\downarrow$		
		Naive	SimW	Dir-CE	Naive	SimW	Dir-CE
Qwen-7B	NQ	.698	.741	.688	.580	.572	.567
	TQA	.815	.829	.830	.253	.248	.227
	PopQA	.750	.800	.721	.516	.494	.493
GPT-4.1-mini	NQ	.726	.742	.727	.520	.516	.505
	TQA	.819	.820	.819	.172	.173	.169
	PopQA	.784	.808	.772	.456	.443	.432
Llama-70B	NQ	.710	.731	.711	.569	.564	.545
	TQA	.798	.802	.796	.205	.205	.198
	PopQA	.716	.740	.737	.571	.558	.511

Table 10: Selective-prediction metrics (AUROC $\uparrow$, AURC $\downarrow$) across all 9 (LLM, dataset) cells with CE evidence; 3-seed repeated-split means (same protocol as Table 2’s $\pm$ std; App. E). Dir-CE attains the lowest mean AURC in all 9 cells over both Naive and SimW (and in 26 of the 27 individual seed-level comparisons; the exception is Qwen-7B/PopQA seed 456, where SimW is lower by 0.001), confirming that the selective-prediction benefit shown for one cell in Table 5 generalizes across cells. AUROC is more mixed, with SimW (the retrieval-similarity prior) often ranking highest, indicating the two metrics emphasize different aspects (ranking vs. coverage-aware risk). Per-split std ≤ 0.0099 (AUROC) and ≤ 0.0093 (AURC) for Dir-CE; the Naive/SimW columns reach 0.0116 and 0.0101. Table 5 uses a single full-test split for detailed baselines, hence minor differences (e.g., Llama/NQ Dir-CE AURC .539 there vs. .545 here).

Source $\rightarrow$	NQ	TQA	PopQA
NQ	30.4	54.7	31.8
TQA	30.0	55.1	31.1
PopQA	30.4	54.7	31.8
Grid Oracle	30.4	55.1	31.8
Max loss		0.69%p	

Table 11: Cross-dataset transfer (Qwen-7B, CE). EM in percent.

Source $\rightarrow$	NQ	TQA	PopQA
NQ	33.9	63.6	34.4
TQA	33.8	63.9	34.3
PopQA	33.8	63.4	34.5
Grid Oracle	33.9	63.9	34.5
Max loss		0.52%p	

Table 12: Cross-dataset transfer (GPT-4.1-mini, CE). EM in percent.

uses 13,659 queries rather than the full 14,267: 608 queries (4.3%) had fewer than 10 retrieved documents in the Contriever cache (16 queries with 8 documents, 592 with 9 documents) and were excluded from the $k=10$ evaluation. The $k=3$ and $k=5$ evaluations use all 14,267 queries. The excluded queries correspond to retrieval-failure cases (Contriever returned fewer candidates), so the reported $k=10$ numbers are, if anything, slightly optimistic relative to the full set; the monotone Δ EM trend (+1.4 $\rightarrow$ +1.5 $\rightarrow$ +2.2%p) holds when $k=3$ and $k=5$ are restricted to the same 13,659 subset.

G Selective Prediction and Calibration Figures

Figures 3–4 visualize the selective-prediction and calibration claims of the main text across multiple cells.

H Qualitative Case Studies

To complement the mechanistic accounts in §6, we inspect two representative NQ queries under Qwen2.5-7B. The cases were selected from the earliest NQ queries in dataset order where (A) Dir-CE rescues a wrong Naive answer (the third such

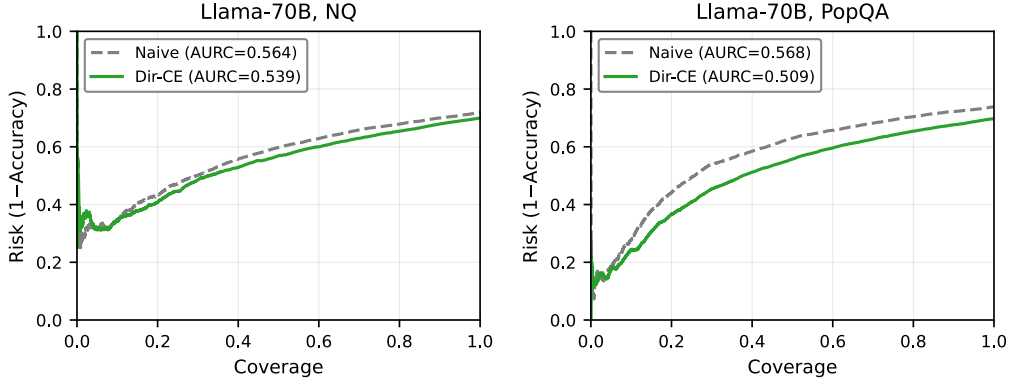


Figure 3: Risk-coverage curves (Llama-3.3-70B, CE evidence) on NQ (left) and PopQA (right). Lower risk at every coverage level is better. Dir-CE lies below Naive at every coverage level above roughly 10% (NQ) and 4% (PopQA); at the lowest coverages, where each point reflects only a handful of queries, the two curves cross. This is consistent with the AURC improvements reported in Table 5.

Dataset	Forward (q as premise)	Reverse (d as premise)	Gold frac.
NQ	.504	.570	.224
TriviaQA	.512	.590	.343
PopQA	.502	.450	.235

Table 13: Gold-vs-non-gold AUC for both NLI directions (500-query sample per dataset; 5,000 query-document pairs for NQ and TriviaQA, 4,976 for PopQA where some queries had <10 retrieved documents). Both directions yield AUC near 0.5 across datasets, indicating weak discrimination regardless of direction. The reverse direction is marginally better on NQ and TriviaQA but worse on PopQA; neither direction achieves strong separation. This supports rather than contradicts the main-text finding that off-the-shelf NLI is an unreliable document-weighting signal in this setting.

query) and (B) NLI-reweighting breaks a correct Naive/CE answer (the second such query). They are illustrative, not statistically selected.

Case A: Cross-encoder discriminates where uniform voting fails.

Q: “when was the last time the 76ers went to the finals” *Gold*: 2001.

Under Qwen-7B, 3 of the 10 retrieved documents prompt the LLM to answer “2022” and 2 prompt “2001”. Naive uniform voting therefore picks the wrong majority, “2022” (confidence 0.30). Cross-encoder scores, however, assign 0.995 to a document whose LLM answer is “2001”; weighting by CE under Eq. 2 gives that document enough mass to flip the vote to “2001” (Dir-CE, confidence 0.235, correct). SimW at $\beta=0.5$ does not help because the DPR cosine sim-

LLM	DS	n	Naive	Dir-NLI	ΔEM	n_{01}	n_{10}	p
Qwen-7B	NQ	8,757	28.43	26.08	-2.35	133	339	$3.9 \times 10^{-21} \dagger$
	TQA	8,837	52.37	50.67	-1.70	92	242	$3.6 \times 10^{-16} \dagger$
	PopQA	14,267	28.39	25.81	-2.58	206	574	$1.9 \times 10^{-39} \dagger$
GPT-4.1-mini	NQ	8,757	32.41	31.22	-1.19	85	189	$4.9 \times 10^{-10} \dagger$
	TQA	8,837	62.95	62.71	-0.24	60	81	0.092
	PopQA	14,267	31.81	30.08	-1.73	182	429	$2.5 \times 10^{-23} \dagger$
Llama-70B	NQ	8,757	28.11	26.98	-1.13	74	173	$4.5 \times 10^{-10} \dagger$
	TQA	8,837	60.11	59.68	-0.43	63	101	3.9×10^{-3}
	PopQA	14,267	26.11	25.00	-1.11	175	333	$3.3 \times 10^{-12} \dagger$

Table 14: McNemar’s test: Dir-NLI vs. Naive (EM in %, full retrieval set). n_{01} : Naive wrong, Dir-NLI right; n_{10} : vice versa. $\dagger$: Bonferroni-significant at $\alpha/27 \approx 1.9 \times 10^{-3}$ (all seven are decreases).

ilarities for the ten documents lie in a narrow band (0.688–0.711) and so produce near-uniform weights. Dir-NLI also fails here: NLI scores are close to 0 or 0.5 for most documents and do not reliably mark the “2001” document.

Case B: NLI fails to discriminate; cross-encoder succeeds.

Q: “who played v in the movie v for vendetta”
Gold: Hugo Weaving.

Here three documents with $CE \approx 1.00$ all elicit the correct answer “Hugo Weaving”, while one document with $CE \approx 1.00$ but a different topic elicits an incorrect “Ian McKellen”, and several remaining documents elicit short refusal/mention answers. Naive, SimW and Dir-CE all pick “Hugo Weaving”. Dir-NLI, however, picks “Stephen Greif did not play V” (a refusal phrasing from a lower-similarity document) because NLI scores are close in magnitude (0.01–0.50) across the ten documents and do not clearly separate the genuinely relevant Wikipedia passages from the distractors. This illustrates the long-passage dilution

DS	LLM	n_{01}	n_{10}	p -value
NQ	Qwen	329	159	<0.001***
NQ	GPT	283	156	<0.001***
NQ	Llama	286	121	<0.001***
TQA	Qwen	446	237	<0.001***
TQA	GPT	267	224	0.058
TQA	Llama	319	195	<0.001***
PopQA	Qwen	622	132	<0.001***
PopQA	GPT	589	211	<0.001***
PopQA	Llama	783	195	<0.001***

Table 15: McNemar’s test: Dir-CE vs. Naive. n_{01} : Naive wrong, Dir right. n_{10} : vice versa. 8/9 significant after Bonferroni ($\alpha/27 \approx 0.00185$); TQA/GPT not significant. PopQA rows use the full retrieval set ($n=14,267$); under strict $k=10$ ($n=13,659$; App. F) the discordant pairs are 599/128, 568/205, 748/190 (Qwen/GPT/Llama) and all three cells remain significant at $\alpha/27$.

k	Naive	SimW	Dir-CE	Δ EM
3	38.2	37.9	39.6	+1.4
5	39.3	39.0	40.7	+1.5
10	39.0	38.8	41.2	+2.2

Table 16: Top- k ablation (averaged across 3 LLMs $\times$ 3 datasets, CE, default config). EM as percentages.

failure mode discussed in §6.1: when the answer-bearing sentence is a small fraction of a long document, the per-document NLI probability barely shifts, so NLI weighting behaves similarly to uniform weighting but with extra noise.

Both cases are consistent with the summary statistics: CE reliably discriminates the gold-bearing document among retrieved candidates, while off-the-shelf NLI on long passages does not, and plain similarity weighting is close to uniform when the retrieved documents are all topically close.

I Brier Score Decomposition

We decompose Brier scores (Brier, 1950) using the binned decomposition of Murphy (1973): Brier = Reliability – Resolution + Uncertainty. The decomposition reveals no single, uniform pattern. On TriviaQA, Dir-CE ties or improves reliability for Qwen-7B and Llama-70B (reliability drops from 0.0117/0.0100 to 0.0109/0.0095) but slightly worsens it for GPT-4.1-mini (0.0068→0.0072). On NQ and PopQA, Dir-CE typically increases reliability (worse calibration) in 5 of 6 cells across all three LLMs (Qwen-7B NQ 0.0029→0.0079, Qwen-7B PopQA 0.0007→0.0023, GPT-

4.1-mini NQ 0.0381→0.0410, GPT-4.1-mini PopQA 0.0058→0.0108, Llama-70B NQ 0.0515→0.0563), with resolution also decreasing for the Qwen-7B cases (NQ 0.0238→0.0216, PopQA 0.0428→0.0362), compounding the calibration loss and consistent with the overall accuracy–calibration trade-off reported in the main text. Llama-70B PopQA is the sole NQ/PopQA cell in which both components improve (reliability 0.0294→0.0212, resolution 0.0305→0.0373); note that even here the net Brier slightly increases (0.1918→0.1948, +0.003) because the uncertainty term shifts upward (0.1929→0.2109), so the component-level improvements at 70B do not translate into a net Brier gain.

J Implementation Details

Prompt template: “Based on the following context, answer the question in a few words. Context: {doc.text[:800]} Question: {question} Answer:”

API details: Qwen2.5-7B and Llama-3.3-70B are served via Together AI; GPT-4.1-mini is served via the OpenAI API. We use greedy decoding (temperature 0) with a 50-token generation limit for all models.

Artifact licenses. NQ-Open is released under CC-BY-SA 3.0 (Wikipedia-derived); TriviaQA under Apache 2.0; PopQA under MIT; DPR and Contriever-MSMARCO encoders under CC-BY-NC 4.0; intfloat/e5-base-v2 (Appendix S) under MIT; Pyserini (BM25 index, Appendix Q) under Apache 2.0; cross-encoder/ms-marco-MiniLM-L-6-v2, sentence-transformers/all-MiniLM-L6-v2, and cross-encoder/nli-deberta-v3-base under Apache 2.0; roberta-large-mnli (Appendix M) under MIT; Qwen2.5-7B-Instruct-Turbo under Apache 2.0; Llama-3.3-70B-Instruct-Turbo under the Llama 3.3 Community License (dated 2024-12-06); GPT-4.1-mini under the OpenAI API Terms of Service. All usage is for non-commercial academic research, consistent with each artifact’s stated intended use.

Computational cost: Total cost $\approx$ \$300: $\approx$ \$100 LLM API for the main DPR/Contriever experiments, plus $\approx$ \$200 for the E5 replication ($\approx$ \$10 cloud-GPU corpus re-encoding, $\approx$ \$190 LLM API; App. S). Evidence computation: CE $\sim$ 0.02 s/query, ES $\sim$ 0.16 s/query, NLI

~ 0.11 s/query. Weight computation: < 1 ms (NumPy). All local orchestration, evidence scoring, and analyses ran on a consumer-grade laptop (36 GB RAM); LLM generation used provider-hosted APIs (Together AI and OpenAI); the E5 corpus re-encoding used one cloud GPU.

API failure rates: GPT-4.1-mini: 98 failures out of the 307,409 unique (question, document) calls (0.03%); the 317,986 query–document pairs reduce to 307,409 unique pairs because repeated question strings are served from the cache. Llama-3.3-70B: 3 failures ($< 0.01\%$). Qwen2.5-7B: 0 failures in the main runs (the BM25 PopQA run is the exception; App. Q). Failed calls are cached as abstentions and counted as incorrect in evaluation.

Document truncation. Cross-encoder and NLI operate on the first 1,000 characters of each document, embedding stability operates on the first 512 characters, and the LLM prompt uses the first 800 characters.

Calibration metrics. SmoothECE uses a Gaussian kernel with bandwidth 0.1 evaluated on a 100-point grid $\{g_m\}_{m=1}^{100}$ over $[0, 1]$. At each grid point g_m we compute a kernel-weighted average accuracy $\bar{a}_m = \sum_i K(g_m, c_i) a_i / \sum_i K(g_m, c_i)$, and report the grid-integrated absolute calibration error $\frac{\Delta g}{n} \sum_m |g_m - \bar{a}_m| \cdot \sum_i K(g_m, c_i)$ with grid spacing $\Delta g = 1/99$. This is a discretized fixed-bandwidth variant of the outcome-smoothed reliability-diagram quantity $\widehat{\text{smECE}}_\sigma$ of Błasiok and Nakkiran (2024), evaluated at $\sigma=0.1$ with an unnormalized, unreflected Gaussian kernel; it is not their residual-smoothed smECE_σ or their self-consistent smECE , so absolute values are not comparable to their `relplot` package. All SmoothECE values in this paper use this same fixed scale. The 18 SmoothECE cell values summarized in §6 are stored in `results/smooth_ece_v2.json` and can be reproduced by running `verify_unchecked_numbers.py` on the cached predictions. Binned ECE uses 10 equal-width bins with integer-floor bin assignment $b(c) = \min(\lfloor 10c \rfloor, 9)$; Brier score is decomposed using the same bins (Murphy, 1973).

Rounding convention. Percentages are rounded half-up to one decimal place unless a table states otherwise (Tables 3 and 30 use two). Differences (Δ columns) are computed from raw values before rounding, so they may differ from

the difference of the displayed values by one unit in the last place.

Statistical tests. We report McNemar’s test using the chi-squared approximation with Yates’s continuity correction ($\chi^2 = \max(0, |n_{01} - n_{10}| - 1)^2 / (n_{01} + n_{10})$; the clamp matters only when $n_{01} = n_{10}$). For the discordant-pair counts we observe (hundreds per cell), this agrees with the exact binomial test to three decimals. The Bonferroni correction uses $\alpha/27$ where the 27 comparisons denote 9 (LLM, dataset) cells $\times$ 3 evidence signals (CE, ES, NLI); reported p -values are raw and are compared against this threshold.

API nondeterminism. Although we use greedy decoding (temperature=0), the Together AI and OpenAI APIs exhibit a small amount of run-to-run nondeterminism (observed $\leq 0.02\%$ EM drift per cell between pipeline runs); this accounts for the occasional 1-query difference between cache snapshots used by different sub-pipelines (e.g., the REPLUG baseline value as read from the CE result file versus the ES/NLI result files differs by up to 0.011%p on a subset of cells). For the same reason, original-run point estimates and later cache-based recomputations can differ by up to 0.013%p per cell, occasionally moving a one-decimal rounding boundary (e.g., GPT-4.1-mini/NQ Dir-CE: 33.86 in the original run vs. 33.85 in the released recomputation).

Vote tallying. Per-document answers are compared after the standard SQuAD normalization (lower-casing; removing punctuation, articles, and extra whitespace); empty outputs and the abstention token “unknown” receive no vote; ties between answers are broken in favor of the answer that appears first in retrieval order.

Per-query confidence. A method’s per-query confidence is defined as the largest vote mass assigned to any single answer, divided by the total vote mass over all retrieved documents (including those whose per-document LLM output is an abstention/“unknown” token). This choice has no effect on $\arg \max$ accuracy (EM, F1) and a negligible effect on selective-prediction metrics (AUROC, AURC, Risk@Coverage).

Reproducibility: Code and result JSON files are available at <https://github.com/DONGJUSEO/rag-document-weighting>; the cached LLM outputs and evidence caches (≈ 98 MB compressed) are archived on Zenodo (<https://doi.org/10.5281/zenodo.22054839>). The $\lambda = 0$ limit of Eq. 2 an-

alytically equals REPLUG’s softmax weights (max-shift cancels in the ratio). Embedding stability noise is drawn from PyTorch’s default generator, which is seeded once at module import with $\text{SEED} = 42$; running the pipeline in the same order as in our experiments reproduces the reported numbers exactly on the same device type (we used Apple MPS; the noise stream depends on the device’s random generator). The Contriever question encoder (used only to build the PopQA retrieval set) uses a plain mean over all output positions; because PopQA queries are short, the effect of including or excluding the small number of padding tokens in the mean is numerically negligible for top- k retrieval.

K Default Configuration Selection

To support the choice of the fixed default ($\beta=0.5, \lambda=30$) used in the main text (§4), we ran a preliminary half-split calibration experiment: for each (dataset, evidence) combination, queries are partitioned 50/50 into a calibration set and a test set (seed 42); (β^*, λ^*) is selected on the calibration half by maximizing EM and then reported.

Table 17 summarizes the cal-tuned best parameters. $\beta=0.5$ emerged as cal-best in all six (dataset, evidence) cells, directly supporting the main-text choice. The cal-tuned best λ^* varied across cells from 0.1 to 30, reflecting the shallow λ -sensitivity curve in the high- λ regime (Figure 2). We did not select a per-cell best λ ; instead, we fixed $\lambda=30$ as a single plateau-end value for all cells to keep the default simple and dataset-agnostic. Consequently, the main-text default is not formally held out for every cell, and admits up to 0.52%p EM inflation relative to per-cell grid optima (§6); this bound is comparable to the smallest headline effect in Table 2 and below the others.

Cell	n_{cal}	n_{test}	cal- β^*	cal- λ^*
NQ, CE	4,378	4,379	0.5	3.0
NQ, ES	4,378	4,379	0.5	0.5
TQA, CE	4,418	4,419	0.5	0.1
TQA, ES	4,418	4,419	0.5	0.1
PopQA, CE	7,133	7,134	0.5	30.0
PopQA, ES	7,133	7,134	0.5	3.0

Table 17: Half-split calibration cal-best (β^*, λ^*) per (dataset, evidence) cell, selected on a 50/50 cal/test partition (seed 42) via EM maximization on the calibration half. $\beta=0.5$ is cal-best across all 6 cells; λ^* varies from 0.1 to 30, and we adopt a single plateau-end default $\lambda=30$ for the main text.

L Lambda Sensitivity for the Evidence-Only Claim

§6.1 reports that evidence-only (EO-CE) weighting achieves near-identical EM to Dir-CE on NQ and PopQA at the default configuration ($\beta=0.5, \lambda=30$). This appendix documents how the Dir-CE versus EO-CE EM gap evolves across λ at $\beta=0.5$, showing that the redundancy claim is operating-point specific.

Table 18 reports Dir-CE EM at $\lambda \in \{0.1, 0.5, 1, 3, 10, 30\}$ alongside EO-CE for each (LLM, dataset) cell. At $\lambda \leq 1$, the retrieval prior $\exp(\beta \cdot s_i) \in [1.11, 1.54]$ (observed range over the three retrieval sets) contributes numerically on par with $\lambda \cdot e_i \in [0, 1]$, and Dir-CE differs from EO-CE by up to 3.81%p (Llama-70B, PopQA). As λ grows to 30, the evidence term dominates numerically ($\lambda \cdot e_i \in [0, 30]$ vs $\exp(\beta \cdot s_i) \in [1.11, 1.54]$), and Dir-CE converges to EO-CE in 6 of 9 cells (to within 0.2%p). The redundancy reported in §6.1 is thus specific to the evidence-dominant default; at smaller λ the prior is clearly active.

M Alternative NLI Model

To verify that the negative NLI finding in §6.1 is not an artifact of our choice of NLI model (nli-deberta-v3-base), we evaluate an alternative model, roberta-large-mnli, on a 500-query random sample per dataset (5,000 query-document pairs for NQ and TriviaQA, 4,976 for PopQA where some queries had <10 retrieved documents; seed 42; forward direction, query as premise, document as hypothesis, documents truncated to 1,000 characters to match the main setup). Table 19 compares gold-vs-non-gold AUC between the two models.

N Embedding Stability at Longer Truncation

The main-text ES implementation truncates documents at 512 characters (§3). Because CE and NLI use 1,000 characters, the shorter ES truncation is a potential cross-signal confound (Limitation 6). To check this, we recompute ES at 1,000-character truncation (with MiniLM’s max_seq_length raised to 512 tokens, the model maximum) on a 500-query random sample per dataset (seed 42) and compare gold-vs-non-gold AUC (Table 20).

LLM	Dataset	$\lambda=0.1$	$\lambda=0.5$	$\lambda=1$	$\lambda=3$	$\lambda=10$	$\lambda=30$	EO-CE	max $ \Delta $
Qwen-7B	NQ	28.57	28.79	29.02	30.01	30.24	30.38	30.38	1.80
Qwen-7B	TQA	53.37	53.57	53.75	55.05	54.93	54.75	54.05	1.01
Qwen-7B	PopQA	29.42	29.82	30.09	31.13	31.66	31.82	31.93	2.50
GPT-4.1-mini	NQ	32.50	32.64	32.87	33.44	33.87	33.86	33.73	1.23
GPT-4.1-mini	TQA	63.21	63.23	63.37	63.87	63.92	63.43	62.48	1.45
GPT-4.1-mini	PopQA	32.52	32.64	33.05	34.04	34.37	34.46	34.41	1.89
Llama-70B	NQ	28.08	28.09	28.38	29.22	29.79	30.01	30.02	1.94
Llama-70B	TQA	60.45	60.53	60.73	61.53	61.79	61.51	60.60	1.19
Llama-70B	PopQA	26.59	26.80	27.37	28.91	29.92	30.23	30.40	3.81

Table 18: Dir-CE EM (%) at $\beta=0.5$ across $\lambda \in \{0.1, 0.5, 1, 3, 10, 30\}$, compared to EO-CE (rightmost data column) for each (LLM, dataset) cell. “max $|\Delta|$ ” (last column) is the largest absolute Dir-CE vs. EO-CE gap across the tested λ values. At $\lambda=30$, the gap falls to $\leq 0.2\%$ in 6 of 9 cells; the maximum gap across smaller λ reaches 3.81%p, showing the prior is active at low λ and numerically redundant at the default.

Dataset	deberta-v3-base	roberta-large-mnli	Δ
NQ	0.504	0.511	+0.007
TriviaQA	0.512	0.517	+0.005
PopQA	0.502	0.502	−0.001

Table 19: NLI alternative-model gold-vs-non-gold AUC (500-query random sample per dataset, the same sample and gold definition as Table 13). `roberta-large-mnli` is within 0.01 AUC of `nli-deberta-v3-base`; both remain near chance and well below the AUC achieved by CE (0.73–0.78) and ES (0.71–0.73) in Table 4. The qualitative finding, that off-the-shelf NLI is a weak document-weighting signal in this setting, is robust to the choice of NLI model.

Dataset	ES@512	ES@1000	Δ
NQ	0.720	0.729	+0.009
TriviaQA	0.736	0.748	+0.013
PopQA	0.717	0.714	−0.003

Table 20: ES gold-vs-non-gold AUC at 512-character vs. 1,000-character document truncation (500-query random sample per dataset, the same sample and gold definition as Table 13; AUC values differ slightly from Table 4 due to the smaller sample). Increasing the truncation length from 512 to 1,000 characters changes AUC by at most 0.013 absolute; the ES signal quality is essentially flat with respect to truncation length in this range. The CE-vs-ES AUC gap is therefore not an artifact of the 512-char truncation.

O Mechanism Diagnostics

To complement the EM-level results in §6, this appendix reports five cache-based diagnostics that probe how Dir-CE redistributes weight (no new LLM inference is required). All values are computed at the default ($\beta=0.5$, $\lambda=30$) on the full test set. For reference, the mean Shannon entropy of the Dir-CE weight vector (the vote-sharpening figure quoted in §6) is 2.073 / 1.989 / 2.087 nats on NQ / TQA / PopQA, against $\log 10 = 2.303$ for uniform weights (−10.0 / −13.6 / −9.4%; `results/weight_entropy.json`).

Answer-support weight mass (C_1). For each query, we sum the voting weight assigned to documents whose generated answer matches the gold (after standard SQuAD/NQ normalization). Higher values indicate that weight is concentrated

on documents whose per-document answer is correct (Table 21).

Dataset	Naive	SimW	Dir-CE	EO-CE
NQ	0.196	0.196	0.220	0.224
TQA	0.480	0.480	0.517	0.526
PopQA	0.187	0.187	0.217	0.221
Avg.	0.288	0.288	0.318	0.323

Table 21: C_1 (answer-support weight mass), averaged across 3 LLMs. Dir-CE concentrates weight on answer-bearing documents in all 9 individual (LLM, dataset) cells; SimW matches Naive to three decimals.

Gold-doc weight mass (C_2). Sum of weights assigned to documents labeled `type: 'gold'` in the retrieval set, the dataset-native annotation used for the has-gold split (§5.2; at most one gold document per query, mean 0.63–0.75), which is

stricter than the answer-containment ratio of §4 (21.9–35.3% of documents). Complements C_1 at the document level (Table 22).

Dataset	Gold ratio	Naive	SimW	Dir-CE	EO-CE
NQ	0.073	0.073	0.073	0.108	0.115
TQA	0.075	0.075	0.076	0.139	0.161
PopQA	0.063	0.063	0.063	0.092	0.095

Table 22: C_2 (gold-doc weight mass). Dir-CE assigns 1.46–1.84 $\times$ more weight to gold-bearing documents than uniform; the largest amplification is on TriviaQA.

Wrong high-confidence (HC) rate (C_3). Fraction of queries where the prediction is wrong but vote mass exceeds 0.5. Lower is better. Quantifies overconfidence (Table 23).

Method	Wrong-HC $\downarrow$	Correct-HC $\uparrow$	Mean conf.	EM (%)
Naive	0.109	0.233	0.467	39.0
SimW	0.132	0.255	0.467	38.8
Dir-CE	0.161	0.290	0.522	41.1
EO-CE	0.201	0.295	0.553	40.9

Table 23: C_3 averaged across 9 (LLM, dataset) cells. Threshold $\tau=0.5$. Dir-CE sharpens both correct and incorrect predictions, quantifying the calibration trade-off discussed in §6.

Oracle gap closure (C_4). $(EM_{\text{Dir-CE}} - EM_{\text{Naive}}) / (EM_{\text{Oracle}} - EM_{\text{Naive}})$, where the Generation Oracle credits a query if any per-document answer matches the gold. Reframes small absolute gains as a fraction of recoverable headroom (Table 24).

LLM	Dataset	ΔEM	Gap	Closure
Qwen-7B	NQ	+1.94	20.46	9.5%
Qwen-7B	TQA	+2.37	15.85	14.9%
Qwen-7B	PopQA	+3.43	18.78	18.3%
GPT-4.1-mini	NQ	+1.45	19.40	7.5%
GPT-4.1-mini	TQA	+0.49	12.22	4.0%
GPT-4.1-mini	PopQA	+2.65	20.32	13.0%
Llama-70B	NQ	+1.88	18.43	10.2%
Llama-70B	TQA	+1.40	13.55	10.4%
Llama-70B	PopQA	+4.12	25.18	16.4%
Average		+2.19	18.24	11.6%

Table 24: C_4 : oracle gap closure per cell. Dir-CE recovers 11.6% of the Generation Oracle headroom on average (range 4.0–18.3%).

Cost-normalized gain (C_5). Mean ΔEM (vs. Naive, across 9 cells) divided by per-query evidence-computation latency.

C_5 (cost-normalized gain) is reported in the main text (“Efficiency note,” §6): CE delivers +0.110 $\Delta EM/ms$ (20 ms latency), $\sim 21\times$ more EM gain per millisecond than ES (+0.005 at 160 ms), while NLI is actively harmful (−0.013 at 110 ms).

Together, C_1 and C_2 show that Dir-CE re-distributes weight onto answer- and gold-bearing documents, consistent with the EM gains being driven by this mechanism; C_3 quantifies the calibration trade-off, C_4 frames the absolute gains as a non-trivial share of oracle headroom (11.6% on average), and C_5 shows CE is the dominant evidence signal once latency is accounted for.

P Concat-Prompt Baseline (Full Results)

This appendix complements the body discussion (§6.6) with full per-cell results (Table 25) comparing our voting framework to the standard concat-prompt RAG baseline. We feed the same top-10 retrieved documents (truncated to 800 chars each, matching the per-document setup) into a single LLM call with the prompt:

```
Based on the following
documents, answer the question
in a few words. Document 1:
{...} ... Document 10: {...}
Question: {q} Answer:
```

Greedy decoding, max 50 output tokens, identical to per-document generation.

Per-LLM averages (Concat – Dir-CE).

- Qwen-7B: −1.27%p (voting wins 2/3, Dir-CE preferred)
- GPT-4.1-mini: −0.09%p (mixed; voting wins 1/3, comparable)
- Llama-70B: +7.64%p (concat dominates, voting wins 0/3)

Selective prediction (Llama-70B / NQ). Despite Concat winning EM (37.59 vs. Dir-CE 30.01), Dir-CE achieves lower AURC and Risk@0.8 than the voting baselines Naive and SimW (Table 5, §5). Concat itself has no entry in that comparison: voting’s vote-mass distribution provides a native confidence signal that concat-prompt’s single-output regime lacks, so Concat has no risk–coverage curve to compare.

LLM	DS	Naive	Concat	Dir-CE	EO-CE	Δ_{C-N}	Δ_{C-D}
Qwen-7B	NQ	28.43	29.12	30.38	30.38	+0.69	-1.26
	TQA	52.38	51.88	54.75	54.05	-0.50	-2.86
	PopQA	28.37	32.13	31.82	31.92	+3.76	+0.31
GPT-4.1-mini	NQ	32.41	35.07	33.86	33.73	+2.66	+1.21
	TQA	62.94	60.43	63.43	62.48	-2.51	-3.00
	PopQA	31.74	35.93	34.40	34.34	+4.18	+1.52
Llama-70B	NQ	28.13	37.59	30.01	30.02	+9.47	+7.58
	TQA	60.11	65.86	61.51	60.60	+5.75	+4.35
	PopQA	26.12	41.20	30.21	30.38	+15.08	+11.00
Average (9 cells)		38.96	43.25	41.15	40.88	+4.29	+2.09

Table 25: Concat-prompt baseline vs. our voting methods on 9 (LLM, dataset) cells. EM in percent. Bold = best EM in row across Naive, Concat, Dir-CE, EO-CE. Δ_{C-N} = Concat - Naive; Δ_{C-D} = Concat - Dir-CE. PopQA voting metrics in this table are recomputed on the $k=10$ strict subset (13,659 queries) for fair comparison with Concat (which requires $k=10$); they differ from Table 2 by $\leq 0.07\%$. Concat outperforms Dir-CE in 6/9 cells (avg +2.10%p), with the gap dominated by Llama-70B (avg +7.64%p across 3 datasets). On Qwen-7B, Dir-CE wins in 2/3 cells (avg -1.27%p in favor of voting). The pattern reflects an inverse correlation between LLM size and voting effectiveness (treating GPT-4.1-mini, whose parameter count is undisclosed, as mid-tier by cost/latency class).

LLM size effect. The inverse correlation (smaller LLMs benefit from voting; larger LLMs benefit from concat) is consistent with the documented positional degradation of long-context inputs (Liu et al., 2024), the lost-in-the-middle phenomenon; the size dependence itself is our empirical observation. We observe this pattern across 9 cells; we do not claim general scaling laws and report the LLM size effect as an empirical observation rather than a causal claim.

Cost. Concat uses 1 LLM call per query with $\sim 2,100$ input tokens; voting uses 10 LLM calls each with ~ 200 input tokens. Total token consumption is comparable, but voting parallelizes across documents (up to 10 concurrent calls on multi-tenant APIs), and voting’s per-document outputs cache more efficiently when a query is reused with different combinations of retrieved documents.

Q BM25 Retriever Robustness

To assess robustness beyond DPR and Contriever, we conducted a post-hoc evaluation using BM25 (Pyserini (Lin et al., 2021)/Lucene; $k_1=0.9$, $b=0.4$) as the retriever with Qwen2.5-7B-Instruct-Turbo and the same evaluation protocol ($k=10$, $\beta=0.5$, $\lambda=30$, CE evidence). BM25 scores are min-max normalized to $[0, 1]$ before computing $\exp(\beta \cdot s_i)$ to maintain compatibility with the REPLUG-style softmax in Eq. 2; without normalization, raw BM25 scores collapse

weight onto the top-1 document. GPT-4.1-mini and Llama-3.3-70B results are not available for BM25.

Finding 1: CE evidence signal transfers to BM25. As shown in Table 26, Dir-CE improves EM over Naive on all three datasets (+1.23/+1.26/+2.00%p), and EO-CE yields consistent gains (+2.07/+1.11/+2.33%p). The CE evidence signal therefore retains its accuracy benefit when the underlying retriever is changed from dense (DPR/Contriever) to lexical (BM25), despite BM25’s lower absolute Naive baseline.

Finding 2: Dataset-level prior behavior replicates qualitatively. A grid search over $\beta \in \{0.5, 1.0, 2.0, 4.0\}$ and $\lambda \in \{0.1, 0.5, 1.0, 3.0, 10.0, 30.0\}$ (24 Dir-CE configurations $\times$ 3 datasets) shows that EO-CE dominates all 24 Dir-CE points on NQ and PopQA, while on TriviaQA 7 of 24 configurations outperform EO-CE (best: $\beta=0.5$, $\lambda=3$, +0.43%p over EO-CE). The main-paper pattern, that retrieval-similarity scores add marginal value on TriviaQA but are redundant or harmful on NQ and PopQA, thus replicates: BM25 similarity scores act as counterproductive weighting priors on NQ and PopQA (EO-CE - Dir-CE_{default} = +0.83%p and +0.34%p), while on TriviaQA the best Dir-CE configuration ($\lambda=3$) outperforms EO-CE by +0.43%p. This suggests the value of the retrieval-similarity prior on TriviaQA is a dataset-level property (consistent with its higher

Dataset	Naive	SimW	Dir-CE	EO-CE	Δ_{Dir}	Δ_{EO}
NQ ($n=8,757$)	22.31	22.32	23.55	24.38	+1.23	+2.07
TriviaQA ($n=8,837$)	54.25	54.35	55.51	55.36	+1.26	+1.11
PopQA ($n=14,267$)	20.76	20.96	22.76	23.10	+2.00	+2.33

Table 26: BM25 robustness evaluation (Qwen2.5-7B; default $\beta=0.5$, $\lambda=30$). EM in %. $\Delta_{\text{Dir}} = \text{Dir-CE} - \text{Naive}$; $\Delta_{\text{EO}} = \text{EO-CE} - \text{Naive}$. Bold = highest EM in row. BM25 Naive baselines are lower than Table 2 due to retriever quality differences; absolute EM values are not directly comparable across retrievers, but Δ patterns are.

retriever–gold-document correlation, $\rho=0.30$ in §6.1) rather than retriever-specific.

Practical implication. For BM25 deployments, EO-CE ($\lambda \rightarrow \infty$) is recommended for NQ-style and PopQA-style tasks; on TriviaQA-style tasks a smaller λ (≈ 3) yields modest additional gains over EO-CE.

API failures in the PopQA run. During the original PopQA run, 12,219 of the 142,670 per-document calls (8.6%; 1,260 queries, of which 1,205 had all ten calls fail) hit the provider’s rate limit on every retry and were cached as abstentions, which receive no vote (App. J); the affected queries are therefore scored as incorrect for all four methods. The provider’s serverless endpoint for this model returned service-unavailable errors at camera-ready time, so the failed calls could not be repeated. On the 13,007 PopQA queries without any failed call, EM is 22.74 / 22.96 / 24.93 / 25.30 for Naive / SimW / Dir-CE / EO-CE ($\Delta_{\text{Dir}} +2.19$, $\Delta_{\text{EO}} +2.56\%$ p). The PopQA row of Table 26 therefore understates absolute EM but not the direction or approximate size of the gains. The NQ and TriviaQA runs had no failed calls.

Limitations of this analysis. Results cover Qwen2.5-7B only; GPT-4.1-mini and Llama-3.3-70B are not evaluated on BM25. The TriviaQA default $\lambda=30$ is suboptimal by 0.28%p relative to the grid optimum ($\lambda=3$), so the default-config transfer claim from the main text holds directionally but not optimally on BM25-TriviaQA. A BGE evaluation remains future work. We expected that high-precision modern dense retrievers would reduce absolute CE gains as retrieval-similarity and CE evidence scores become more correlated (i.e., when the retriever already ranks answer-bearing documents at the top), while the evidence-only ($\lambda \rightarrow \infty$) regime should remain applicable regardless of retriever type, since EO-CE bypasses the retrieval-similarity prior entirely. Both expectations are confirmed by the E5 evaluation (Ap-

DS	LLM	Subset	n	Naive	Dir-CE	ΔEM	McNemar p
NQ	Qwen-7B	has-gold	6,364	38.37	41.01	+2.64 [†]	<0.001
		no-gold	2,393	2.01	2.09	+0.08	0.823
	GPT-4.1-mini	has-gold	6,364	41.01	43.23	+2.22 [†]	<0.001
		no-gold	2,393	9.53	8.90	−0.63	0.046
	Llama-70B	has-gold	6,364	35.64	38.22	+2.58 [†]	<0.001
		no-gold	2,393	8.11	8.15	+0.04	1.000
TQA	Qwen-7B	has-gold	6,653	67.13	70.31	+3.19 [†]	<0.001
		no-gold	2,184	7.42	7.33	−0.09	0.899
	GPT-4.1-mini	has-gold	6,653	75.74	76.70	+0.96	0.0015
		no-gold	2,184	23.99	22.99	−1.01	0.036
	Llama-70B	has-gold	6,653	71.76	73.95	+2.19 [†]	<0.001
		no-gold	2,184	24.63	23.63	−1.01	0.019
PopQA	Qwen-7B	has-gold	8,537	45.16	50.57	+5.41 [†]	<0.001
		no-gold	5,122	0.39	0.57	+0.18	0.0077
	GPT-4.1-mini	has-gold	8,537	48.37	52.90	+4.53 [†]	<0.001
		no-gold	5,122	4.04	3.57	−0.47	0.0060
	Llama-70B	has-gold	8,537	39.89	46.71	+6.83 [†]	<0.001
		no-gold	5,122	3.18	2.69	−0.49	0.0039

Table 27: Full has-gold / no-gold decomposition: subset size, absolute EM (%), ΔEM (Dir-CE – Naive, %p), and McNemar p . [†]: Bonferroni-significant at $\alpha/27 \approx 1.9 \times 10^{-3}$. TQA/GPT-4.1-mini ($p=0.0015$) passes numerically but is reported as marginal (§5.2).

pendix S).

R Has-Gold / No-Gold Decomposition Details

Table 27 reports the full statistics behind Table 3 (§5.2): subset sizes, absolute EM, and McNemar p -values. A query is has-gold if any of its top-10 retrieved documents is marked gold under the dataset-native annotation (strict $k=10$); the split depends only on the retrieval run and is therefore identical across the three LLMs.

Six no-gold cells reach nominal $p<0.05$ (five negative, one positive), but none survives Bonferroni correction ($\alpha/27$) and all are small in magnitude ($\leq 1.01\%$ p). The two largest (both TQA) sit on a $\approx 24\%$ EM base, while the single positive cell (PopQA/Qwen-7B, +0.18%p) is driven by only 9 discordant pairs ($n_{01}=9$, $n_{10}=0$) at a below-1% EM base, consistent with noise rather than a genuine no-gold effect.

Family-size sensitivity. We apply the uniform $\alpha/27$ divisor throughout (§4). Under the family-

DS	LLM	Naive	SimW	Dir-CE	EO-CE	Δ_{Dir}	p
NQ	Qwen-7B	35.37	35.33	35.47	35.23	+0.10	0.66
	GPT-4.1-mini	36.13	36.17	36.33	36.23	+0.19	0.29
	Llama-70B	31.78	31.78	32.07	31.97	+0.29	0.11
TQA	Qwen-7B	59.00	58.91	59.53	59.28	+0.53	0.028
	GPT-4.1-mini	64.77	64.72	64.86	64.56	+0.09	0.68
	Llama-70B	62.46	62.48	62.77	62.36	+0.31	0.15
PopQA	Qwen-7B	36.31	36.32	37.85	37.85	+1.54 [†]	<0.001
	GPT-4.1-mini	39.69	39.69	40.91	40.93	+1.23 [†]	<0.001
	Llama-70B	34.57	34.56	36.71	36.81	+2.14 [†]	<0.001
Average Δ_{Dir}						+0.71	

Table 28: E5 full-set EM (%) for all nine cells ($n=8,757 / 8,837 / 14,267$). $\Delta_{\text{Dir}} = \text{Dir-CE} - \text{Naive}$ (raw difference rounded, as in Table 2). Bold = best in row. [†]: Bonferroni-significant (McNemar, $\alpha/27$). Absolute EM is not comparable to Table 2 on NQ/TQA (legacy DPR passage surface forms).

specific divisor for the nine no-gold tests ($\alpha/9 \approx 5.6 \times 10^{-3}$), one cell crosses the threshold in the negative direction: Llama-70B/PopQA ($\Delta_{\text{EM}} - 0.49\%p$, $p=3.9 \times 10^{-3}$). This does not change the reading that weighting offers no reliable gain without a gold document. If anything, it strengthens the recommendation to fall back to uniform weighting when the diagnostic predicts no separation. No has-gold conclusion is affected: all nine has-gold p -values clear both divisors (the marginal TQA/GPT-4.1-mini cell passes both numerically).

S E5 Retriever Evaluation

To assess the framework on a modern high-precision retriever, we re-encoded the full 21M-passage Wikipedia corpus with `intfloat/e5-base-v2` (using the required “query:”/“passage:” prefixes) and re-ran the complete pipeline on all nine (LLM, dataset) cells ($\approx 920k$ new LLM calls; same protocol: $k=10$, $\beta=0.5$, $\lambda=30$, CE evidence). Raw cosine similarities enter Eq. 2 directly (no min-max normalization). Because our NQ/TQA passage sets use the legacy DPR surface form (lower-cased/tokenized), absolute EM is not comparable to Table 2 on those datasets; we therefore report within-E5 comparisons only. For the has-gold decomposition below, “gold” means answer-string match, applied within E5.

Within E5, Dir-CE improves EM over Naive in all nine cells (Table 28; avg $+0.71\%p$), but only the three PopQA cells are Bonferroni-significant; TQA/Qwen-7B is nominally significant ($p=0.028$). Three observations:

1. Absolute gains shrink vs. DPR/Contriever

($+2.19 \rightarrow +0.71\%p$ average), the behavior Appendix Q anticipated for high-precision retrievers. E5 retrieves an answer-bearing document in the top-10 for 70–81% of queries (70.4% / 80.4% / 75.9% on NQ / TQA / PopQA), Naive voting is already strong, and the per-query top1–top10 cosine spread is compressed (0.028–0.031 mean), leaving similarity and evidence signals little room to disagree. The attenuation is not an artifact of the NQ/TQA passage surface-form change: on PopQA, the one dataset whose absolute EM is directly comparable across retrievers (Table 28 caption), the average gain also shrinks, $+3.40 \rightarrow +1.63\%p$.

2. **The similarity-only prior is effectively redundant on E5.** SimW differs from Naive by less than 0.1% p in every cell (avg $-0.02\%p$): the similarity prior alone provides no measurable benefit over uniform voting on the most modern retriever tested.

3. **The evidence-only limit remains competitive.** EO-CE tracks Dir-CE on PopQA ($|\text{Dir} - \text{EO}| \leq 0.11\%p$) and on NQ for GPT-4.1-mini and Llama-70B ($+0.09\%p$ each), trails it marginally on NQ for Qwen-7B ($+0.24\%p$), and most consistently on TQA ($\text{Dir} - \text{EO} = +0.25 / +0.31 / +0.41\%p$), consistent with the main text’s finding that the similarity prior adds the most value on TriviaQA, the dataset with the highest retriever–gold correlation ($\rho=0.30$) and gold ratio (35.3%; §6.1).

Has-gold / no-gold decomposition. The same analysis as §5.2 (with answer-string gold and full retrieval sets, without a strict- k filter, per the preamble above) yields avg $\Delta_{\text{Dir}} = +0.97\%p$ on has-gold vs. $-0.17\%p$ on no-gold. Per dataset, has-gold Δ_{Dir} : PopQA $+2.04 / +1.63 / +2.86$ (Qwen / GPT / Llama; all $p < 10^{-10}$), TQA $+0.75 / +0.30 / +0.58$, NQ $+0.15 / +0.19 / +0.24$, the same concentration pattern as on DPR/Contriever, at a smaller scale.

(β, λ) grid. Table 29 reports the same 24-configuration grid as Appendix Q ($\beta \in \{0.5, 1, 2, 4\}$, $\lambda \in \{0.1, 0.5, 1, 3, 10, 30\}$), here for all nine cells. The fixed default is robust on E5: the gap between the default and the per-cell grid optimum is $\leq 0.41\%p$ in every cell, tighter

DS	LLM	Default	Best	(β^*, λ^*)	Gap	EO-CE	>EO
NQ	Qwen-7B	35.47	35.61	(2, 30)	0.14	35.23	24/24
	GPT-4.1-mini	36.33	36.35	(2, 30)	0.02	36.23	9/24
	Llama-70B	32.07	32.11	(2, 30)	0.05	31.97	5/24
TQA	Qwen-7B	59.53	59.93	(1, 10)	0.40	59.28	10/24
	GPT-4.1-mini	64.86	65.27	(2, 30)	0.41	64.56	24/24
	Llama-70B	62.77	62.95	(2, 30)	0.18	62.36	24/24
PopQA	Qwen-7B	37.85	37.85	(0.5, 30)	0.00	37.85	0/24
	GPT-4.1-mini	40.91	40.91	(0.5, 30)	0.00	40.93	0/24
	Llama-70B	36.71	36.71	(0.5, 30)	0.00	36.81	0/24

Table 29: E5 (β, λ) grid summary per cell: EM of the fixed default ($\beta=0.5, \lambda=30$) vs. the per-cell grid optimum (β^*, λ^*); Gap = Best – Default (%p); >EO = number of the 24 Dir-CE grid points exceeding EO-CE.

than the $\leq 0.52\%$ p bound on DPR/Contriever, and on PopQA the default is the grid optimum in all three cells. Every per-cell optimum lies in the evidence-dominant regime ($\lambda \in \{10, 30\}$). The EO-CE comparison largely replicates the dataset-level pattern of Appendix Q: on PopQA no grid point beats EO-CE (0/24 in all three cells), on TQA most or all do (10–24 of 24), and on NQ, where BM25 showed full EO-CE dominance, the outcome is mixed with margins $\leq 0.4\%$ p.

T Aggregation-Rule Comparison

Table 30 compares Dir-CE against three additional training-free aggregation rules, computed from the same cached per-document answers with no additional LLM calls (all nine cells, strict $k=10$): **RRF** (Cormack et al., 2009) fuses the similarity rank and the CE rank via reciprocal-rank fusion ($1/(60+r_{\text{sim}}) + 1/(60+r_{\text{CE}})$) as voting weights; **Borda** converts CE ranks to weights $k-r+1$; **CE-Top1** returns the answer of the single highest-CE document (no vote).

Four observations. (1) Dir-CE outperforms RRF in all nine cells (avg +1.71%p; 8/9 McNemar-significant at $\alpha/27$, the exception being TQA/GPT-4.1-mini, $p=0.30$). RRF is the closest external counterpart to our formula, as it fuses the same two signals, but by rank, re-admitting the similarity prior. (2) Borda over CE ranks is statistically indistinguishable from Dir-CE (avg -0.16% p). No cell is significant in either direction; nominal $p \leq 0.05$ differences favor Borda in four cells (NQ/Qwen $p=0.018$; PopQA/Qwen $p=0.009$; PopQA/Llama $p=0.006$; NQ/GPT $p=0.0499$) and Dir-CE in none. Two different functional forms over the same CE signal thus perform equivalently, supporting the interpretation that the evidence signal itself, not the

DS	LLM	Naive	RRF	CE-Rerank	CE-Top1	Borda	Dir-CE
NQ	Qwen-7B	28.43	28.88	29.44	28.75	30.92	30.38
	GPT-4.1-mini	32.41	32.57	33.92	32.68	34.28	33.85
	Llama-70B	28.11	28.24	29.88	30.24	29.93	30.00
TQA	Qwen-7B	52.37	53.06	52.66	50.55	54.65	54.75
	GPT-4.1-mini	62.95	63.17	62.28	59.70	63.38	63.43
	Llama-70B	60.11	60.39	60.67	58.93	61.11	61.51
PopQA	Qwen-7B	28.37	29.51	30.61	33.16	32.23	31.82
	GPT-4.1-mini	31.74	32.45	33.40	34.54	34.52	34.40
	Llama-70B	26.12	26.72	30.87	34.21	30.75	30.21
Average		38.96	39.44	40.42	40.31	41.31	41.15

Table 30: Training-free aggregation rules on the same cached per-document answers (EM, %; no additional LLM calls). Bold = best in row. CE-Rerank = top-5 CE documents with uniform weights (the hard-selection baseline of §5), recomputed on the same strict- k answers. Borda’s average edge over Dir-CE (+0.16%p) is not significant in any cell (McNemar, $\alpha/27$); Dir-CE beats RRF in all 9 cells (8/9 significant), beats CE-Rerank in 7 (5 significant), and splits with CE-Top1 5–4 along dataset lines.

specific form, drives the gains. Borda, however, discards score magnitudes (rank-only) and has no prior-evidence interpolation parameter, so it cannot express the λ -diagnostic that is the point of the framework. We read it as a strong aggregation baseline within the per-document voting regime, not a replacement for it. (3) CE-Top1 is dataset-dependent: Dir-CE wins five cells (all three TQA, two NQ), CE-Top1 wins the three PopQA cells (by up to +4.00%p on Llama-70B) plus NQ/Llama-70B (+0.24%p, $p=0.53$). On PopQA, single entity-centric passages often contain the answer span directly, consistent with the one CE-Rerank loss in §5 (Llama-70B on PopQA). Selection-style rules also return a point answer only: they do not produce the vote-mass distribution used for selective prediction (§5.4). (4) CE-Rerank is the hard top-5 selection of §5 recomputed on the strict- k subset, and it reproduces the full-set pattern. Dir-CE wins 7 of 9 cells (5 McNemar-significant at $\alpha/27$; avg +0.73%p). It is within noise on NQ/GPT-4.1-mini (-0.07% p, $p=0.84$) and loses only on PopQA/Llama-70B (-0.67% p, $p=6.1 \times 10^{-3}$, not significant at $\alpha/27$), the same cell where CE-Top1 excels.

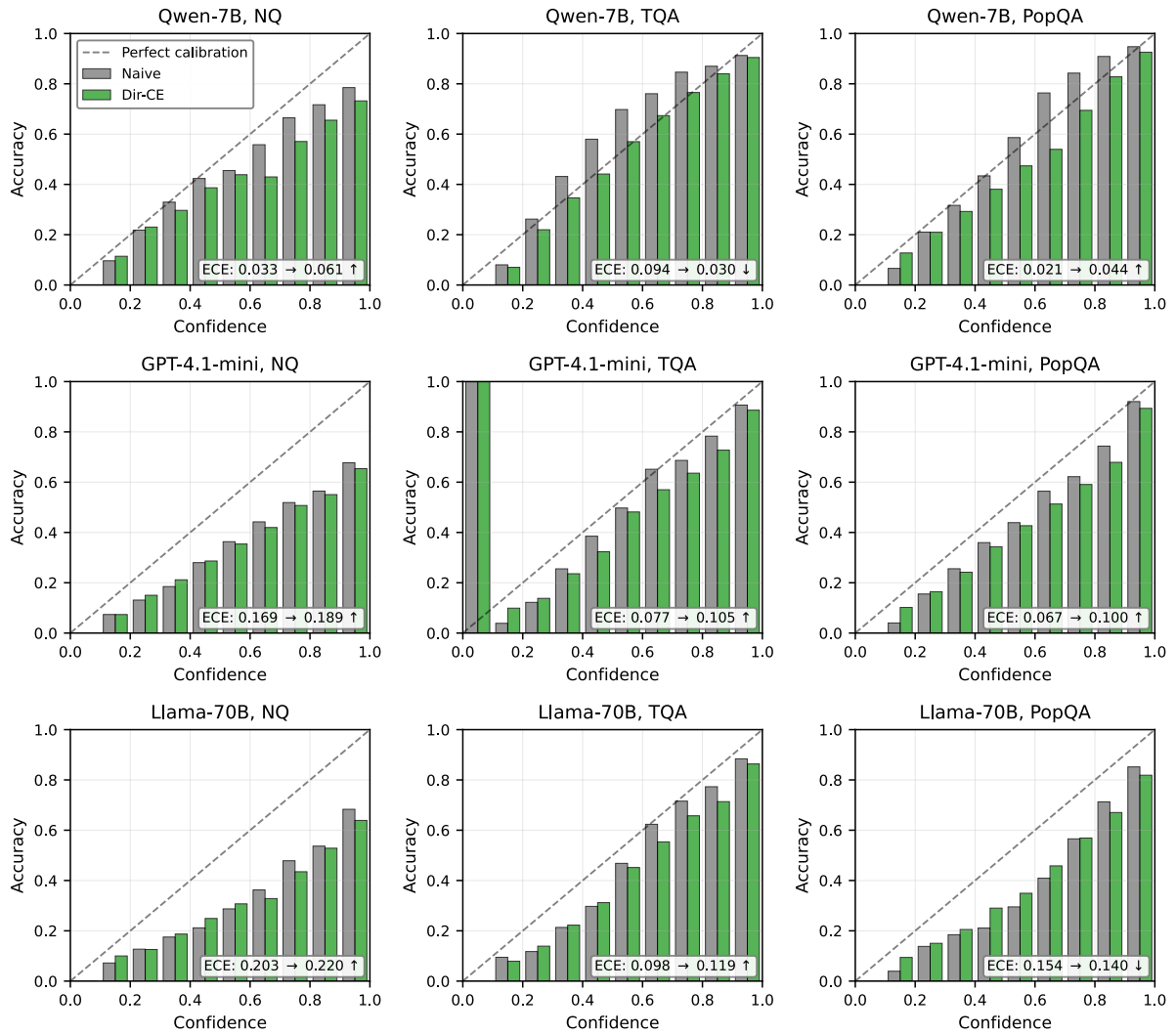


Figure 4: Full reliability grid: 3 LLMs (rows) $\times$ 3 datasets (columns), Naive vs. Dir-CE at $\beta=0.5, \lambda=30$. Each panel shows binned accuracy per confidence bin; dashed line = perfect calibration. ECE before/after is annotated per panel with direction arrow ($\uparrow$ = worsens, $\downarrow$ = improves). The pattern is heterogeneous rather than cleanly scale-monotonic: GPT-4.1-mini shows a consistent shift below the diagonal (overconfidence) across all three datasets; Qwen-7B behaves similarly on NQ and PopQA but shows the grid’s largest improvement on TriviaQA (ECE -0.064); Llama-70B improves on PopQA (-0.014) and moderately worsens on NQ/TriviaQA. Plain ECE magnitudes (binned) may differ in direction from the kernel-smoothed SmoothECE reported in §6; the figure is illustrative of the mechanism rather than a reproduction of the SmoothECE counts.