RECOVERING EVIDENCE SUFFICIENCY BEYOND RETRIEVAL SIMILARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

A retrieval system returns documents for every query, whether or not the answer is there. We would like to tell, from the embeddings alone, when the retrieved evidence is actually sufficient, before spending a generator on it. Prior work suggests this cannot be read off the geometry: fixed signals barely beat cosine similarity, so the geometry looks blind. We show the blindness was in the test, not the geometry. Those tests used an additive probe on the two vectors, which cannot represent any interaction between them and therefore cannot even represent cosine; it was asked to find a signal in a language that could not express it. Restore the interaction and the signal reappears, on all nine encoders we test and on a second, structurally different entailment corpus. Two facts explain the earlier null. The signal is not held in any few directions but spread redundantly across many, which is exactly why a fixed cosine readout misses it while a probe that pools directions recovers it. And the probe is nearly free: one dot product over embeddings the retriever already holds, orders of magnitude below a cross-encoder. That is cheap enough to gate every query before generation, with a finite-sample guarantee on how often sufficient evidence is wrongly rejected.

1 INTRODUCTION

Retrieval-augmented generation returns documents for every query, relevant or not. In high-stakes settings we want a check before the generator runs: is the retrieved evidence enough to answer, or should we escalate? The obvious place to look is the retrieval embeddings, but fixed geometric readouts are weak: cosine is dominated by a few high-variance directions (Ethayarajh, 2019; Timkey & van Schijndel, 2021), and work on retrieval sufficiency accordingly relies on learned classifiers rather than reading the geometry directly (Joren et al., 2025). Together these have made embedding geometry look blind to sufficiency. Across nine relevance-trained encoders we reproduce the null: additive probes on the geometry add essentially nothing over cosine.

The conclusion does not follow, and the reason is the test itself. Those probes were additive: a linear model on the pair $[q; d]$ adds a term in q to a term in d and stops there, so it cannot represent any interaction between the two, cosine included. It was asked to detect a signal in a language that could not state it.

Once the hypothesis class contains interactions, the picture changes. A logistic probe on $[q; d; |q - d|; q \odot d]$, whose $q \odot d$ block contains cosine as the uniform-weight special case, beats cosine on all nine encoders on a factual-entailment task, with a mean out-of-fold gain of $+0.083$ AUROC, and by a larger margin ($+0.261$) on semantic opposition where cosine is at chance. A feature ablation shows the signal is genuinely pair-level: a query-only probe carries none of it.

We then ask *why* the signal is invisible to similarity, and answer it. We find (§5) that the learned readout is a *stable* direction, that the two regimes on which the probe succeeds are *anti-aligned* rather than merely distinct, and that on the opposition regime the readout runs *against* the cosine direction. An erasure diagnostic (§6) then identifies the mechanism: the signal is *distributed and redundant*, surviving the removal of any low-dimensional linear subspace, including the probe's own. This is not a failed search for a localized cause, it *is* the cause. Diffuse, redundant encoding is precisely the geometry in which a fixed isotropic readout is blind while a probe pooling across directions succeeds, which is why the same representation that "fails" under cosine yields a large signal under interactions.

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

The result upgrades an empirical probe into a mechanistic explanation of when, and why, retrieval geometry looks blind. We are careful throughout to separate evidence from proof: we make no claim of a necessary low-dimensional cause, and note (§9) that non-linear localization remains open.

Contributions. (1) A *diagnosis*: the reported blindness of embedding geometry to sufficiency is a hypothesis-class artifact; interaction probes recover a large signal on 9/9 encoders on two structurally different tasks (§3). (2) A *protocol*: a two-condition (rank-displacement vs cosine-matched) swap test with a matching certificate, reusable for auditing any proposed sufficiency signal (§4). The interaction template itself is standard (InferSent; DistMult); our contribution is not a new architecture but the diagnosis above and the geometric analysis that follows. (3) A *mechanistic explanation*: the readout is stable, anti-identity on opposition, and locally low-rank, and an erasure diagnostic shows the signal is globally *distributed and redundant*, the encoding geometry that makes cosine blind (§5–6). (4) A *cost accounting*: split-conformal escalation at fixed false-negative rate against a trained cross-encoder, on a cost–quality frontier (§7).

2 SETUP

Task and label. An encoder f retrieves the top-K documents by cosine ($K = 5$). A query is EVIDENCE-IN if a gold supporting document appears in the top-K and EVIDENCE-OUT otherwise. EVIDENCE-IN is a retrieval-sufficiency label, not an entailment label; §4 is the bridge, testing directly whether scores react to entailment-relevant content at fixed similarity. We use four BEIR (Thakur et al., 2021) tasks with deliberately different relevance criteria: SciFact (Wadden et al., 2020) and Climate-FEVER (Diggelmann et al., 2020) (factual entailment, on a scientific and a 5.4M-document Wikipedia corpus respectively), FiQA-2018 (Maia et al., 2018) (opinion retrieval), and ArguAna (Wachsmuth et al., 2018) (semantic opposition, where the target is a counter-argument).

Encoders. Nine relevance-trained encoders from three families: BGE (Xiao et al., 2023), E5 (Wang et al., 2022), and GTE (Li et al., 2023) (small/base/large). We treat these as correlated observations from three families and report all correlational statements as descriptive.

Probe. Logistic regression with an inner cross-validated regularization path (Alain & Bengio, 2016); the scaler and path are fit inside each outer fold, and out-of-fold predictions are pooled per encoder. Feature families: $\mathtt{q}$ only, $\mathtt{d}$ only, concat $[q; d]$, and interaction $[q; d; |q - d|; q \odot d]$ on L_2-normalized embeddings. The interaction template is that of InferSent (Conneau et al., 2017); the diagonal block $q \odot d$ weighted by w is exactly $q^\top \mathrm{diag}(w)\, d$, the DistMult relational scorer (Yang et al., 2015), and the full bilinear form $q^\top W d$ is RESCAL (Nickel et al., 2011).

Reproducibility scope. All nine encoders are shared across experiments. The Tier-1 rank sweep (§5), the erasure diagnostic (§6), and the conformal gate (§7) use a single fixed 5-fold partition stored with each encoder's embedding pack, so their splits are identical; the matched-swap audit (§4) is a separate protocol with its own held-out construction, described in place. We release the harness so every table regenerates end to end.

3 THE BLINDNESS IS IN THE READOUT, NOT THE REPRESENTATION

Theorem 1. *Let $s(q, d) = g(f(q), f(d))$ be any score computed from the embedding pair, and let P_1, P_0 be the joint distributions of $(f(q), f(d))$ conditional on the label. Then* $\mathrm{AUC}(s) \leq \frac{1}{2} + \mathrm{TV}(P_1, P_0)$.

The proof is a standard total-variation argument (Appendix A). Stating the bound on the *joint* matters: a document-marginal version admits scores that exploit query-conditional structure, and the ablation below shows that this is exactly where the sufficiency signal lives. The query-only probe carries none of the interaction gain, while document-only and concat track cosine and the interaction probe carries all of it.

We ran the blindness experiment twice. With the additive probe on $[q; d]$ (plus a global PCA, as in prior practice), the null holds. With interaction features the same data yield a mean out-of-fold gain

Table 1: Per-encoder cosine and interaction-probe AUROC on SciFact (out-of-fold). Every encoder improves over cosine.

Encoder	$\mathrm{AUC}_{\mathrm{cos}}$	$\mathrm{AUC}_{\mathrm{probe}}$	Δ
bge-small-en-v1.5	0.806	0.869	+0.063
bge-base-en-v1.5	0.798	0.902	+0.104
bge-large-en-v1.5	0.834	0.920	+0.086
e5-small-v2	0.759	0.853	+0.094
e5-base-v2	0.785	0.904	+0.119
e5-large-v2	0.820	0.907	+0.087
gte-small	0.837	0.875	+0.038
gte-base	0.830	0.910	+0.080
gte-large	0.841	0.916	+0.075

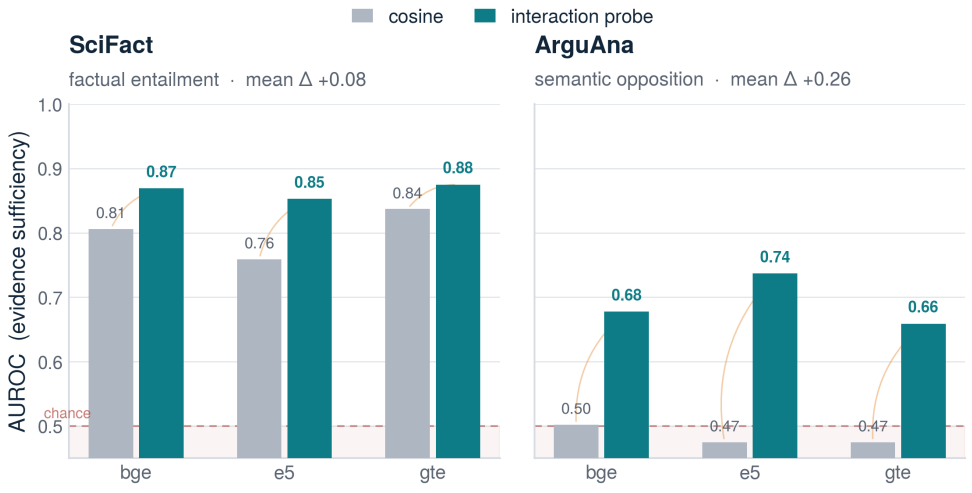

Figure 1: Cosine vs interaction-probe AUROC across three encoders on the two winning regimes (replication run; full nine-encoder results in Table 1 and Table 3). Where relevance means opposition (ArguAna), cosine is at chance and the probe recovers a large signal.

of $+0.083$ AUROC on SciFact, positive on 9 of 9 encoders (Table 1, Figure 1). A linear model on $[q; d]$ cannot represent $\sum_i q_i d_i$; the additive certificate of blindness could not have detected the signal even where it exists.

Two regimes, one negative case. We repeat the comparison across regimes (SciFact and ArguAna in Table 3; FiQA as a negative control). The interaction probe beats cosine on SciFact ($+0.083$, 9/9) and, by a wide margin, on ArguAna ($+0.261$, 9/9, where cosine is at chance), and adds nothing on the opinion-retrieval control (FiQA, -0.03), a regime where relevance is not opposition-based. We treat ArguAna as an extreme stress test rather than a sufficiency task in the strict sense: relevance there is semantic opposition, not entailment, so it asks whether the readout adapts to a regime cosine cannot handle, not whether sufficiency reduces to opposition. ArguAna is notable because cosine there is near chance: when relevance means opposition, similarity is actively misleading, and the learned readout recovers a large signal the cosine cannot. On FiQA interaction structure injects noise where cosine already suffices. We report FiQA as a single control point rather than per encoder, since our claim is that it should show no gain, and it does not.

A second entailment dataset. To check that the effect is a property of factual entailment and not of SciFact, we repeat the two core diagnostics on Climate-FEVER (Diggelmann et al., 2020), whose corpus (5.4M Wikipedia sentences) and domain differ entirely from SciFact. The interaction probe beats cosine by $+0.158$ and $+0.195$ on the two encoders we run (Table 2), *larger* gains than on SciFact, and the cosine-matched swap test gives AUROC 0.83 and 0.89 with the certificate satisfied, a much larger separation than the 0.54–0.59 on SciFact, on $\sim$1000 matched pairs. The claim thus rests

Table 2: Second entailment corpus (Climate-FEVER, 5.4M-doc Wikipedia). Interaction probe vs cosine, and the cosine-matched swap test (probe AUROC gold vs matched distractor, with the cosine certificate). Both diagnostics agree with SciFact and are stronger here.

Encoder	$\mathrm{AUC}_{\cos}$	$\mathrm{AUC}_{\text{probe}}$	Δ	swap AUROC (CI)	cert.
bge-small-en-v1.5	0.678	0.836	+0.158	0.826 [0.809, 0.841]	0.504
e5-small-v2	0.699	0.894	+0.195	0.890 [0.878, 0.901]	0.508

Table 3: Per-encoder ΔAUROC (interaction probe $-$ cosine) on the two regimes evaluated across all nine encoders. Every encoder improves on both; the probe wins by a wide margin on ArguAna, where cosine is at chance.

Encoder	SciFact	ArguAna
bge-small-en-v1.5	+0.063	+0.176
bge-base-en-v1.5	+0.104	+0.233
bge-large-en-v1.5	+0.086	+0.227
e5-small-v2	+0.094	+0.261
e5-base-v2	+0.119	+0.346
e5-large-v2	+0.087	+0.353
gte-small	+0.038	+0.183
gte-base	+0.080	+0.280
gte-large	+0.075	+0.286
mean	+0.083	+0.261

on two independent entailment corpora, and the content signal the swap test isolates is, if anything, stronger on the larger corpus.

4 Content, not position: a two-condition swap protocol

A probe trained on EVIDENCE-IN labels might learn retrieval-position patterns rather than content. The following protocol separates the two and is reusable as an audit for any candidate sufficiency signal.

Control (rank displacement). For queries whose top-1 is gold, replace it with the highest-ranked non-gold document. Cosine detects this swap essentially perfectly, and must: the swapped pair's cosine is lower than the original's in 100% of pairs by construction. Reporting only this condition manufactures a spurious win for similarity baselines; we report it as the control it is.

Test (cosine-matched). Replace the gold document with the top-100 non-gold document whose cosine is closest to the gold's, keeping pairs with $|\Delta \cos| \leq 0.01$. A certificate accompanies every run: the cosine AUC on the resulting original-vs-swap pairs must lie in $[0.45, 0.55]$, verifying that similarity carries no usable information on this set. Matched pairs are abundant on SciFact, which is itself telling: non-entailing neighbors at the evidence's own similarity level are the typical case, the regime where similarity-based gating must fail.

Result. Under a pre-registered rule (bootstrap CI of the original-vs-swap AUC above 0.5, at least 100 matched pairs, certificate satisfied), the interaction probe passes on 9 of 9 encoders (Table 4). Gold annotations are not exhaustive, so some matched distractors may in fact be entailing; this biases the test against the probe, making the result conservative. The separation is modest in absolute terms (AUROC 0.54–0.59; Table 4), as it must be once the similarity signal is certified away; we read it as a lower bound on the geometric content, which non-exhaustive gold pushes down further. We run this test on SciFact rather than ArguAna by necessity: the protocol needs pairs matched in cosine, and ArguAna, where cosine is already at chance, yields essentially none (no encoder clears the minimum pair count). This is consistent with the picture above, not a gap: a similarity certificate is only meaningful where similarity carries information to certify away.

Table 4: Cosine-matched swap test: probe AUROC (gold vs matched distractor) with bootstrap CI, and cosine AUROC on the same pairs (certificate, all in $[0.45, 0.55]$). All nine pass (CI above 0.5).

Encoder	n	$\mathrm{AUC}_{\mathrm{probe}}$	CI_{95}	$\mathrm{AUC}_{\cos}$
bge-small-en-v1.5	232	0.563	$[0.532, 0.594]$	0.507
bge-base-en-v1.5	232	0.583	$[0.550, 0.618]$	0.509
bge-large-en-v1.5	219	0.542	$[0.518, 0.567]$	0.507
e5-small-v2	376	0.586	$[0.564, 0.607]$	0.533
e5-base-v2	329	0.584	$[0.559, 0.612]$	0.528
e5-large-v2	327	0.561	$[0.540, 0.583]$	0.529
gte-small	283	0.562	$[0.533, 0.589]$	0.517
gte-base	305	0.569	$[0.547, 0.594]$	0.520
gte-large	296	0.572	$[0.550, 0.597]$	0.520

5 Characterizing the readout

We now ask how the geometry organizes the recovered signal. Unless noted, this section analyzes the three small encoders (bge-small, e5-small, gte-small) with the diagonal bilinear probe $q^\top \mathrm{diag}(w)\, d$; all statements are descriptive.[1]

The direction is a stable object. Median within-task cosine between refit weight vectors is 0.82–0.92 across encoders, against a random-direction null $\mathbb{E}|\cos| = \sqrt{2/(\pi d)} = 0.04$ at $d = 384$: the readout is reproducible, not fit noise, by roughly $20\times$ over chance. This is the precondition for every subsequent claim.

The two regimes are anti-aligned, not merely distinct. The cosine between the SciFact and ArguAna weight directions is *negative* on every encoder (-0.03, -0.15, -0.08), with the confidence interval entirely below zero on two of three (e5 $[-0.19, -0.11]$, gte $[-0.12, -0.03]$). Orthogonality is the high-dimensional default; anti-alignment is not. The regimes do not merely use different directions, they use opposed ones along the coordinates they share. This is stronger than a "distinct directions" observation and is what distinguishes a genuinely different readout from a re-weighted one.

Only one regime is recalibrated cosine. On SciFact the weight aligns positively with the uniform direction (cosine of w with $\mathbf{1}/\sqrt{d}$ between $+0.54$ and $+0.68$) and puts little mass on negative coordinates (negative-mass share 0.05–0.12), consistent with a readout close to cosine, which already scores well there. On ArguAna the picture inverts: the uniform alignment is *negative* (-0.13, -0.22, -0.17) and the negative-mass share is large (0.59–0.66; Figure 2). Where relevance means opposition, the readout that works runs *against* the identity form; it is not a reweighted similarity but a different linear functional, which is why cosine is near or below chance there while the probe recovers a large signal. Crucially, a positive per-coordinate rescaling (whitening) cannot produce negative mass, so this separates a genuinely different readout from anisotropy correction (Ethayarajh, 2019; Timkey & van Schijndel, 2021).

The opposition signal is locally low-rank. We fit $\alpha I + UV^\top$ and sweep the rank (Figure 3). Recovery of the ArguAna gain saturates at *rank 2* and is flat thereafter; on cosine-matched pairs a single axis reaches AUROC 0.71–0.79, and the leading subspace is stable across folds (principal-angle cosine 0.55–0.66 vs 0.07 for a random subspace). The identity component (α) alone is at chance on ArguAna; the off-identity component carries the signal.

6 Why cosine is blind: an erasure diagnostic

We now turn the erasure toolkit into a *diagnostic* for how the signal is encoded, which in turn explains the blindness of §3. That a rank-2 form *recovers* the signal does not tell us whether it is *stored* in a

[1] The stability, anti-alignment, and dissociation statistics use the three small encoders; the erasure diagnostic (Table 5) covers all nine.

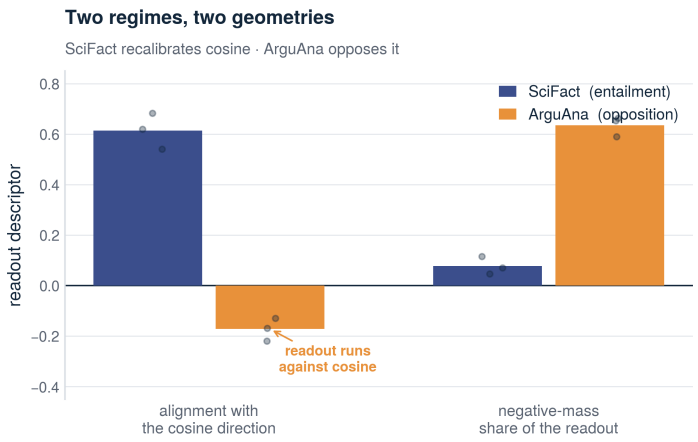

Figure 2: Alignment with the cosine direction and negative-mass share of the learned readout, per regime (means; per-encoder points overlaid). SciFact is a recalibrated cosine; ArguAna runs *against* the identity direction.

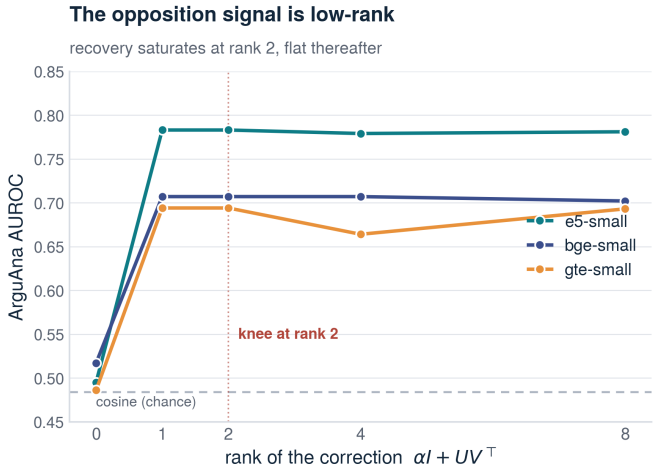

Figure 3: ArguAna AUROC of $\alpha I + UV^\top$ vs rank. Recovery saturates at rank 2; the dashed line is cosine (at chance here).

few directions or spread across many, and this distinction is exactly what determines whether a fixed isotropic readout can see it. Erasure answers it directly. Adopting the amnesic-probing methodology (Elazar et al., 2021), we define a concept axis *externally* (independent of the probe), erase it with LEACE (Belrose et al., 2023) (closed-form linear concept erasure, verified to remove all linear predictability of the concept), *retrain* the probe on the erased representation, and read off how much signal survives. The logic is a dichotomy: if the signal were localized, erasing a few directions would destroy it; if distributed, no low-dimensional erasure would. We use two external axes, negation minimal pairs, and a contradiction direction fit on MNLI (Williams et al., 2018) and transferred, plus an erasure-per-rank sweep on the probe's own subspace and the single-direction projection of INLP (Ravfogel et al., 2020) as controls.

The signal is distributed and redundant. The diagnosis is unambiguous, and three independent measurements converge on it.

(i) The erasure-per-rank curve is flat. Erasing k dimensions of the probe's own ArguAna subspace and retraining barely moves the gain, out to $k = 32$ (Figure 4): the fraction of the original gain surviving a 32-dim erasure is at least 78% on all nine encoders (small, base, and large across the three families; mean 0.94), and on several it is $\geq 100\%$ (the retrained probe recovers as much or more along

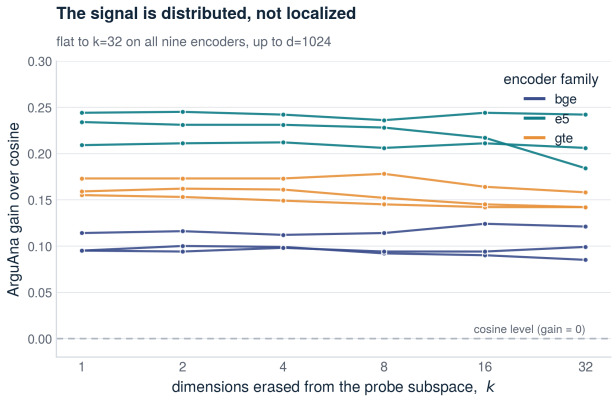

Figure 4: Erasing k dimensions of the probe's own ArguAna subspace and retraining. The gain over cosine is essentially unchanged out to $k = 32$: the sufficiency signal is not localized in any low-dimensional linear subspace.

Table 5: Erasure-per-rank on ArguAna, all nine encoders: gain over cosine after erasing k dimensions of the probe's own subspace and retraining. "surv." = gain(k=32)/gain(k=1). Every encoder stays flat (surv. near or above 1), including the large (d=1024) variants: the signal is distributed, not localized, at every scale. Base/large variants use capped ranks (r*=1–2 throughout).

Encoder	k=1	k=2	k=4	k=8	k=16	k=32	surv.
bge-small-en-v1.5	0.095	0.100	0.099	0.092	0.090	0.085	0.89
bge-base-en-v1.5	0.114	0.116	0.112	0.114	0.124	0.121	1.06
bge-large-en-v1.5	0.095	0.094	0.098	0.094	0.094	0.099	1.04
e5-small-v2	0.244	0.245	0.242	0.236	0.244	0.242	0.99
e5-base-v2	0.234	0.231	0.231	0.228	0.217	0.184	0.78
e5-large-v2	0.209	0.211	0.212	0.206	0.211	0.206	0.98
gte-small	0.155	0.153	0.149	0.145	0.142	0.142	0.92
gte-base	0.173	0.173	0.173	0.178	0.164	0.158	0.91
gte-large	0.159	0.162	0.161	0.152	0.145	0.142	0.90

other directions). A localized signal would show a knee at small k; this is a horizontal line, and it stays flat as the embedding dimension grows to 1024 (large variants), ruling out a low-dimensionality artifact. Table 5 reports all nine.

(ii) Erasing the probe's own axis fails. The tautological control that *should* work if the signal were low-dimensional, erasing the probe's own leading subspace, removes only $\approx 10\%$ of the gain on e5 and is slightly negative on bge and gte, because the retrain recovers the same AUROC along other directions.

(iii) A non-linear reader recovers it. After LEACE erasure of the opposition axis, a small MLP on the erased representation recovers essentially all of the original gain (recovered gain 0.24–0.32; the information was never removed, it is simply not confined to the erased directions).

Consistently, the external opposition axis has *low* overlap with the probe's ArguAna subspace (principal-angle cosine 0.14/0.14/0.035) and removes nothing ($\approx$ 2–4% of the gain, no better than a random subspace). Because overlap is checked *before* testing erasure, the axis is chosen on an independent criterion rather than to make the story work; and this low overlap shows that no better-matched *low-dimensional* axis would succeed, since the curve in Figure 4 is flat regardless.

Reconciliation. A rank-2 subspace is *sufficient* to read the signal but not *necessary*: after it is erased, retraining finds another rank-2 reading in a different subspace because the information is replicated across many directions. The signal is locally low-rank but globally redundant. This is the mechanistic counterpart of the "coherent irrelevance" intuition: the representation holds the sufficiency signal in diffuse form, no single direction owns it, which is exactly why a fixed isotropic readout (cosine)

Table 6: False-negative rate of the conformal gate, in-domain versus under task shift (mean over nine encoders). In domain the rate sits at the nominal ε; transferring a threshold across tasks can violate it, worst when calibrating where cosine is competent (SciFact) and testing where it is not (ArguAna).

Calibrate → test	$\varepsilon = 0.05$	$\varepsilon = 0.10$
In-domain (either task)	0.05	0.10
SciFact → ArguAna	0.97	0.98
ArguAna → SciFact	0.06	0.10

cannot express it while a probe that pools across directions can. We therefore make no claim of a necessary low-dimensional cause; the defensible statement is that the signal is *distributed, not localized*. Whether it is localized *non-linearly* (removable by a non-linear eraser) is open (§9).

7 CONFORMAL GATE ECONOMICS

AUROC does not measure what a gate buys, which is tail behavior. We report sufficiency gates as *escalation at fixed risk*: a split-conformal threshold (Papadopoulos et al., 2002; Vovk et al., 2005) calibrated on held-out EVIDENCE-OUT scores guarantees a false-negative rate $\leq \varepsilon$ with finite-sample validity (Bates et al., 2021; Angelopoulos et al., 2024), and the reported quantity is the fraction of queries escalated to the verifier. On SciFact, under identical train/calibration/test splits (10 seeds), the conformal probe gate resolves 24.3% of within-domain queries before any verifier call at $\varepsilon = 0.05$ versus 15.2% for a conformal cosine gate, and 39.3% versus 31.0% at $\varepsilon = 0.10$; the probe gate escalates less than the cosine gate on all nine encoders at both risk levels.

Distribution shift breaks the guarantee, and asymmetrically. The gate's finite-sample validity assumes calibration and test scores are exchangeable, which holds only in domain. In domain it does: across all nine encoders the empirical false-negative rate matches the nominal ε (Table 6, first row). Under a task shift it can fail completely, and the direction of the shift decides how badly. A threshold calibrated on SciFact, where cosine is already competent and scores are compressed near the top, is far too permissive on ArguAna, where the winning readout produces much larger scores: the false-negative rate jumps from ε to 0.88–1.00 on every encoder, passing almost all insufficient evidence. The reverse, calibrating on ArguAna and applying to SciFact, stays close to ε (mean 0.06 at $\varepsilon = 0.05$, 0.10 at $\varepsilon = 0.10$), with only mild violations. Shift toward the regime where cosine fails is the dangerous one. The gate is therefore a monitored control that must be recalibrated in domain, not a static guarantee; we regard surfacing this, rather than reporting only the in-domain number, as part of the contribution.

Cost. A trained cross-encoder reranker (Nogueira & Cho, 2019; Wang et al., 2020) on the same EVIDENCE-IN labels is more accurate, on ArguAna it reaches AUROC 0.814 versus the interaction probe's 0.723, but at the cost of a full transformer forward per pair ($\sim 10^{10}$ FLOPs), against a single weighted dot product for the probe ($\sim 4.6 \times 10^3$ FLOPs) over embeddings the retriever has already computed: a 10^6–$10^7\times$ gap (Figure 5). For a gate that must run on *every* query before generation, this asymmetry, not a small AUROC margin, is what decides whether pre-generation gating is deployable at all. The residual accuracy the cross-encoder adds is consistent with the non-linear share of the signal identified in §6. Between these two extremes sit late-interaction retrievers such as ColBERT (Khattab & Zaharia, 2020), which score with a MaxSim over token embeddings; we do not evaluate them here, but they occupy an intermediate cost regime and would be expected to fall between the interaction probe and the cross-encoder in Figure 5. A direct comparison is a natural extension.

8 RELATED WORK

Sufficient-context classifiers (Joren et al., 2025) and grounding evaluators such as RAGAS (Es et al., 2024) judge generated answers or context; we work pre-generation and contribute an audit protocol and cost accounting. Hallucination detectors from model internals (Manakul et al., 2023; Farquhar et al., 2024) operate after generation; our gate needs only embeddings and a logistic layer. Selective prediction (El-Yaniv & Wiener, 2010) and adaptive retrieval (Asai et al., 2024) abstain without

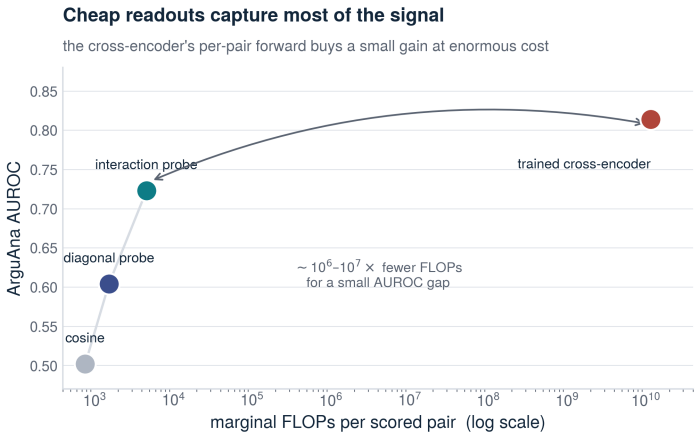

Figure 5: Quality vs. marginal cost on ArguAna (log-FLOPs), all points on the same encoder. Geometric readouts on the retriever's frozen embeddings, cosine, the diagonal probe, and the interaction probe, capture most of the separation at 10^6–$10^7\times$ lower marginal cost than a trained cross-encoder, whose per-pair forward pass buys a modest AUROC gain at enormous cost.

distribution-free guarantees; we use split conformal for the gate and note C-RAG (Kang et al., 2024) certifies generation risk rather than the pre-generation gate. Interaction features for pair classification date to InferSent (Conneau et al., 2017) and linear probing to Alain & Bengio (2016), with control-task caveats from Hewitt & Liang (2019); anisotropy of trained encoders (Ethayarajh, 2019; Timkey & van Schijndel, 2021) motivates why similarity-aligned readouts can miss information. Our erasure analysis uses LEACE (Belrose et al., 2023) and the INLP (Ravfogel et al., 2020) / amnesic-probing (Elazar et al., 2021) tradition; the diagonal/full bilinear connection is to DistMult (Yang et al., 2015) and RESCAL (Nickel et al., 2011).

9 Limitations and Outlook

We state the boundaries of the present study precisely, as each marks a concrete next question rather than a gap we paper over.

Depth vs. breadth of the characterization. The erasure diagnostic that establishes distributed encoding runs on all nine encoders and, crucially, holds as the embedding dimension grows to 1024 (Table 5), so the central claim is not an artifact of small models. The finer weight-space statistics (within-task stability, cross-task anti-alignment, §5) are reported on the three small encoders, whereas the erasure diagnostic that carries the necessity claim, and the probe and matched-swap results (§3–4), already span all nine.

Distribution shift. The conformal gate's guarantee is bound to exchangeability; §7 shows it holds in domain and can break under a task shift, so the gate needs in-domain recalibration. A broader study of shift, beyond the two-task transfer reported there, is future work.

Scope of the tasks. All experiments are English-language and single-hop: sufficiency is a property of a single query–document pair. Multi-hop retrieval, where evidence is distributed across documents and sufficiency need not be isolable from any single (q, d) pair, is not addressed by our pairwise formulation and is an important direction on its own, one where the failure of a pointwise similarity should be even more pronounced.

Linear vs. non-linear localization: a deliberate frontier. Our result is a precise statement about *linear* structure: the sufficiency signal is distributed across, and not removable from, any low-dimensional *linear* subspace, while a non-linear reader recovers it in full (§6). We read this not as an open loose end but as the first sharp coordinate on a map: it locates exactly where linear geometry ends and non-linear structure begins. The natural successor is a *non-linear* concept-erasure that asks

whether the signal, invisible to isotropic similarity and diffuse to linear readouts, admits a localized non-linear description. Answering it would turn "the readout is what matters" from a diagnosis into a design principle for a new class of geometric sufficiency evaluations, and, ultimately, for retrievers trained to be sufficiency-separable rather than merely similarity-aligned (§8). We regard drawing this boundary cleanly, rather than blurring it, as part of the contribution.

10 Conclusion

The blindness of retrieval geometry to evidence sufficiency, reported in prior work and reproduced here with fixed signals, is a property of the readouts, not the representations. A ten-line interaction probe recovers the signal on every encoder tested on two structurally different tasks; a matched-swap protocol certifies that it reads content at fixed similarity; a characterization shows the readout is stable and, on the opposition regime, anti-aligned with cosine and locally low-rank; and an erasure analysis shows the signal is globally distributed rather than localized. A conformal gate converts the probe into escalation savings with a finite-sample guarantee. The representation holds the signal; fixed isotropic readouts cannot express it.

Use of AI assistance

In line with the ICLR 2027 policy, we disclose the use of AI assistants in preparing this work. AI tools were used to help write and debug the experimental code and analysis pipeline, to generate the figures from computed results, and to edit and tighten the prose. All research decisions, the hypotheses, the experimental design, the choice of datasets and encoders, the interpretation of results, and every quantitative claim, were made and verified by the authors. All reported numbers were produced by the authors' own experimental runs and checked against the source data; no results, citations, or references were fabricated by an AI system. The authors take full responsibility for the content.

A Proof of Theorem 1

Let $X \sim s_{\#}P_1$, $Y \sim s_{\#}P_0$. Then $\mathrm{AUC} = \frac{1}{2} + \frac{1}{2}\,\mathbb{E}[\mathrm{sgn}(X - Y)]$. With $h(x) = \mathbb{E}_Y[\mathrm{sgn}(x - Y)] \in [-1, 1]$ and signed measure $\mu = s_{\#}P_1 - s_{\#}P_0$, $\mathbb{E}[\mathrm{sgn}(X - Y)] = \int h\, d\mu \le \|h\|_\infty \|\mu\|_{\mathrm{TV}} \le 2\,\mathrm{TV}(s_{\#}P_1, s_{\#}P_0) \le 2\,\mathrm{TV}(P_1, P_0)$, the last step by the data-processing inequality. □

References

G. Alain and Y. Bengio. Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644*, 2016.

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal risk control. *ICLR*, 2024.

A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. Self-RAG. *ICLR*, 2024.

S. Bates, A. Angelopoulos, L. Lei, J. Malik, and M. I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021.

N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman. LEACE: Perfect linear concept erasure in closed form. *NeurIPS*, 2023.

T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *arXiv:2012.00614*, 2020.

A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from NLI data. *EMNLP*, 2017.

Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *TACL*, 2021.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *JMLR*, 2010.

S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. RAGAS. *EACL*, 2024.

K. Ethayarajh. How contextual are contextualized word representations? *EMNLP*, 2019.

S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. Detecting hallucinations using semantic entropy. *Nature*, 2024.

J. Hewitt and P. Liang. Designing and interpreting probes with control tasks. *EMNLP*, 2019.

H. Joren, J. Zhang, C.-S. Ferng, D.-C. Juan, A. Taly, and C. Rashtchian. Sufficient context: A new lens on retrieval-augmented generation systems. *ICLR*, 2025.

O. Khattab and M. Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *SIGIR*, 2020.

M. Kang, N. M. Gürel, N. Yu, D. Song, and B. Li. C-RAG. *ICML*, 2024.

Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv:2308.03281*, 2023.

M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, and A. Balahur. WWW-2018 open challenge: Financial opinion mining and question answering (FiQA). *WWW Companion*, 2018.

P. Manakul, A. Liusie, and M. Gales. SelfCheckGPT. *EMNLP*, 2023.

M. Nickel, V. Tresp, and H.-P. Kriegel. A three-way model for collective learning on multi-relational data (RESCAL). *ICML*, 2011.

R. Nogueira and K. Cho. Passage re-ranking with BERT. *arXiv:1901.04085*, 2019.

H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. *ECML*, 2002.

S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *ACL*, 2020.

N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *NeurIPS Datasets and Benchmarks*, 2021.

W. Timkey and M. van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *EMNLP*, 2021.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

H. Wachsmuth, S. Syed, and B. Stein. Retrieval of the best counterargument without prior topic knowledge. *ACL*, 2018.

D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi. Fact or fiction: Verifying scientific claims. *EMNLP*, 2020.

L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei. Text embeddings by weakly-supervised contrastive pre-training (E5). *arXiv:2212.03533*, 2022.

W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression. *NeurIPS*, 2020.

A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference (MultiNLI). *NAACL*, 2018.

S. Xiao, Z. Liu, P. Zhang, and N. Muennighoff. C-Pack: Packed resources for general Chinese embeddings (BGE). *arXiv:2309.07597*, 2023.

B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng. Embedding entities and relations for learning and inference in knowledge bases (DistMult). *ICLR*, 2015.