TRIM: TRAINING-FREE RETRIEVAL FROM INTERMEDIATE MODEL STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Intermediate layers of LLMs often perform surprisingly poorly in dense retrieval when read out as a single vector, raising a fundamental question: do these layers lack retrieval-relevant information, or is the information present but not yet accessible to the native representation? We introduce training-free retrieval from intermediate model states (TRIM), a framework combining theoretical analysis and retrieval interventions to investigate this question without training or modifying the underlying model. First, we bypass single-vector pooling and perform token-level matching on shallow states. Surprisingly, even very shallow layers recover substantial retrieval performance despite near-zero performance with the native readout, revealing that retrieval-relevant evidence is already present but remains distributed across tokens. We then ask whether this evidence can be recovered into a single vector. Re-aggregating token states using the model's own attention substantially improves retrieval at middle depths, showing that insufficient aggregation is a major bottleneck. Finally, we find that the remaining gap can be largely explained by the geometry of the resulting embedding space: a training-free geometric calibration increasingly recovers retrieval quality at deeper layers. Together, these results reveal a depth-dependent transition in intermediate representations, i.e., from distributed retrieval evidence, to incomplete aggregation, to geometric mismatch. Building on this observation, we introduce a fully training-free early-exit retrieval pipeline that combines attention-based readout with geometric calibration. Without updating the backbone or using relevance labels, it retains 91.6% of full-depth nDCG@10 across seven datasets while reducing query encoding time by 26.8–30.5% in our timed setting. Our findings suggest that the poor retrieval performance of intermediate LLM states is not simply due to missing retrieval information, but to how that information is progressively aggregated and geometrically organized across depth.

1 INTRODUCTION

Large language models (LLMs) are increasingly used for dense retrieval, where queries and documents are encoded into vectors and ranked by similarity (Karpukhin et al., 2020; Wang et al., 2024; Zhang et al., 2025), but their retrieval quality comes with substantial encoding cost because every transformer block is executed before the final embedding is produced. Early exit offers a natural way to reduce this cost by using an intermediate layer as the retrieval representation (Xin et al., 2020). Yet a disappointing phenomenon emerges in practice: intermediate layers can retrieve extremely poorly even when the same model performs strongly at its final layer. For example, when both queries and documents exit at layer 22 of Qwen3-Embedding-4B, the native single-vector representation retains only 2.2% of the full-depth nDCG@10 on three BEIR datasets. This raises a fundamental question of how retrieval signal propagates across model depth. Poor intermediate-layer retrieval may have several explanations. The layer may lack retrieval-relevant information, or the information may already be present but distributed across token states and lost by the native single-vector readout. Even if successfully aggregated, the resulting representation may still lie in a geometry poorly suited for document ranking. Distinguishing these possibilities is important for both early exit and understanding how retrieval representations emerge across model depth.

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.

We approach this question through a sequence of training-free[1] diagnostic probes, each designed to remove one possible bottleneck. We first examine whether retrieval-relevant evidence is already present in intermediate token states. To bypass the single-vector readout entirely, we apply token-level MaxSim matching directly to frozen intermediate states, following the late-interaction principle of ColBERT (Khattab & Zaharia, 2020). The result is striking: at layer 8 of E5-Mistral-7B, token-level matching retains 74.0% of full-depth nDCG@10, compared with only 1.8% using the native pooled representation. Thus, weak dense retrieval at an intermediate layer does not imply that the layer lacks retrieval-relevant information. Instead, substantial matching evidence is already available, but remains distributed across token states rather than being effectively expressed by a single vector.

This leads to another important question of how to recover this evidence from native readout. We hypothesize that intermediate token states have not yet been aggregated in a way that preserves the information needed for retrieval. We therefore introduce Attention^2 Readout, a training-free readout that re-aggregates the token states using the model's own attention from its designated pooling position. For example, at layer 20 of E5-Mistral-7B, retention increases from 40.9% with native pooling to 67.0% with Attention^2 alone, and to 96.5% when combined with geometric calibration. Across models, the attention-based readout is most beneficial at intermediate depths, where calibration alone leaves substantial retrieval quality unrecovered. These results suggest that aggregation, rather than information availability, is a major bottleneck in intermediate-layer retrieval.

Yet aggregation does not unfold the whole picture. As the exit approaches the final layer, the attention-based readout becomes less important, while geometric calibration becomes increasingly effective. At later exits, calibration alone can recover more than 90% of full-depth retrieval performance and, in some cases, approach or exceed the full-depth reference. This depth-dependent behavior reveals a broader pattern that the bottleneck of retrieval changes with depth. At shallow depths, useful evidence is present but distributed across tokens; at intermediate depths, insufficient aggregation becomes a dominant obstacle; toward the final layer, the remaining discrepancy is increasingly a matter of scoring geometry. These effects may overlap across depths and architectures rather than follow universal boundaries, but together explain why native intermediate representations retrieve poorly.

This perspective also reveals an important distinction between representation fidelity and retrieval utility. Although a natural strategy is to make an intermediate representation resemble its full-depth counterpart, retrieval depends on the relative ordering of document scores rather than reconstruction error. Indeed, our controlled experiments show that analytic scaling can improve retrieval while increasing anchor cosine reconstruction error in all twelve tested settings. This explains why small representation changes can alter rankings and why closer reconstruction need not improve retrieval. The above observations lead to a practical consequence: intermediate layers can be made useful for retrieval without retraining the model. We combine Attention^2 Readout with geometric calibration fitted offline on unlabeled anchors, yielding a training-free early-exit pipeline over a frozen backbone. The resulting query-side early exit can also operate directly against an existing full-depth document index, leaving stored document vectors untouched and eliminating collection-wide re-encoding. At layer 25 of Qwen3-Embedding-4B, our approach retains 91.6% of full-depth nDCG@10 across seven datasets while reducing batch-one query encoding time by 26.8–30.5% on an A800, including the additional readout, calibration, and normalization costs.

Our contributions are threefold. First, we introduce TRIM, a framework that combines three training-free methods with ranking analysis to study how retrieval evidence is aggregated and geometrically organized across model depth. Second, we demonstrate retrieval recovery across four embedding models, covering both early exit on queries and documents and query-only early exit with an unchanged full-depth index. Third, we explain why matching full-depth embeddings more closely need not improve retrieval, and derive a tight bound on query error for preserving document rankings in a fixed index.

2 Related work

Depth-adaptive embeddings and early exit. Matryoshka Representation Learning (Kusupati et al., 2022) trains nested embedding dimensions, and ESE (Li et al., 2025) extends this adaptivity to

[1]Here, training-free means keeping the backbone frozen and using no gradient-based optimization or relevance labels. Geometric calibration fits a closed-form map on unlabeled anchors.

depth, training shallow layers to express salient information. Early exits have also been trained for classification, generation and answer ranking (Xin et al., 2020; Elhoushi et al., 2024; Soldaini & Moschitti, 2020). Even without such training, intermediate layers can yield stronger text embeddings than the final layer (Skean et al., 2025), and training-free Token Prepending (TP; Fu et al., 2025) exits at a validation-selected intermediate layer. We instead study readout and scoring at a frozen retrieval embedder's early exits against shallow or existing full-depth indexes.

Pooling token evidence. E5-Mistral (Wang et al., 2024), GTE-Qwen2 (Li et al., 2023) and Qwen3-Embedding (Zhang et al., 2025) use a designated-position, full-depth single-vector readout. ColBERT and ColBERTv2 (Khattab & Zaharia, 2020; Santhanam et al., 2022) use token-level MaxSim; SMART (Zhang et al., 2026a) applies late interaction to frozen multimodal embedding models and examines different layers. Methods that keep one vector alter the input or information flow (Jiang et al., 2024; Springer et al., 2025; BehnamGhader et al., 2024; Fu et al., 2025; Tang & Yang, 2026) or learn an attention-based pooler (Lee et al., 2025). Training-free Ditto weights tokens by diagonal attention (Chen et al., 2023), hierarchical TP (HTP) uses mean pooling (Ding et al., 2026), and AlignedWVA aggregates value vectors across layers using last-token attention and output projections (Zhang et al., 2026b). Attention^2 re-pools block-output states at the exit; TRIM compares this readout with token matching and geometric recovery, including fixed-index reuse. FastE (Shu et al., 2026) uses readout attention to select states for later layers; we use it for the exit representation.

Scoring geometry and index compatibility. Similarity also depends on geometry: anisotropic embedding spaces degrade sentence similarity, which contrastive learning can alleviate (Xu et al., 2023), but the embedders we study apply this objective only at full depth. Index compatibility can be learned through shallow-query distillation against a fixed document encoder (Wang & Hong, 2023) or backward-compatible retrieval training (Cai et al., 2023); Drift-Adapter (Vejendla, 2025) instead fits paired-embedding adapters, including closed-form Procrustes, to map a new model's queries into a legacy index without retraining. Within one model, the Tuned Lens (Belrose et al., 2023) trains affine probes on intermediate states to predict the final-layer distribution. Our calibration instead maps a frozen embedder's exit to its own full-depth space using unlabeled anchors; we test this closed-form map alone and after re-pooling, and analyze ranking effects.

3 TRIM: A Framework for Retrieval across Model Depth

An *exit* ℓ is the last executed block of an L-block model, and a *readout* turns its token states into retrieval vectors. Native readout uses pooling position p: the last valid decoder token or CLS in BERT-style encoders. TRIM combines a depth-dependent account of retrieval, three training-free probes and ranking analysis. With the encoder frozen (Figure 1), MaxSim tests token evidence, Attention^2 Readout tests re-aggregation, and calibration tests geometric recovery. Section 3.5 characterizes order preservation in a fixed index.

3.1 Evidence availability: token matching with MaxSim

For unit query and document token vectors and scored token sets I_q, I_d,

$$S_{\text{token}}(q,d) = \sum_{i \in I_q} \max_{j \in I_d} \left\langle q_i^{(\ell_q)}, t_j^{(\ell_d)} \right\rangle . \tag{1}$$

Each query token independently chooses its best document token, bypassing pooling. A pooled query cannot exceed the correspondingly weighted best-match score (Appendix A.1), and Proposition 1 gives the exact centered-score fit available to one query vector on a fixed dense index; neither result orders documents. We use $\ell_q = \ell_d$: both sides are shallow, and each document keeps its token vectors.

3.2 Aggregation: Attention^2 Readout

Let $a_{pj}^{(\ell,h)}$ be the attention from the pooling position p to position j in head h of block ℓ, $\mathcal{V}_\ell(p)$ the valid positions, and $w_j = H^{-1} \sum_{h=1}^{H} a_{pj}^{(\ell,h)}$. Attention^2 applies these weights to the block's output

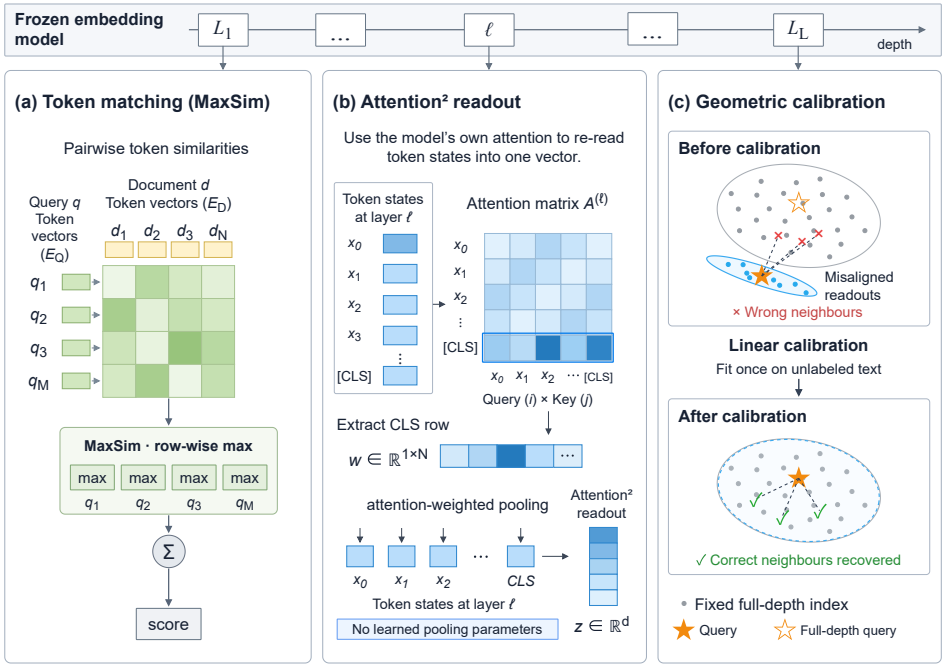


Figure 1: **The three training-free probes in TRIM.** (a) MaxSim scores each query token against its best document token. (b) Attention^2 Readout pools a layer's token states with the attention row of the pooling position, drawn as CLS; our decoders use the last valid token. (c) Geometric calibration maps a native readout toward a fixed full-depth index by a closed-form fit on unlabeled text. Depth placement is illustrative and useful ranges overlap; (b) and (c) can be composed.

states:

$$u_\ell^{\text{Attention}^2}(x) = \text{Normalize}\Big(\sum_{j\in\mathcal{V}_\ell(p)} w_j h_j^{(\ell)}\Big). \tag{2}$$

Attention is reused for readout; "2" is not a matrix square (Appendix C.7). No readout parameters are fitted; we retain backbone SDPA and separately recompute only attention row p from Q/K projections, including this cost in latency (Appendix B.1). Re-pooling alone does not ensure deep-index compatibility.

3.3 Geometry: Calibration of Native Pooling

The calibration-only probe keeps native pooling, $u_\ell = u_\ell^{\text{native}}$. Encoding n unlabeled texts at exit ℓ and at full depth gives row matrices X, Z with means μ_x, μ_z and centered versions X_c, Z_c. With $S_{xx} = X_c^\top X_c/(n-1)$, $S_{xz} = X_c^\top Z_c/(n-1)$ and source dimension d_x, ridge calibration fits

$$W = (S_{xx} + \lambda I)^{-1} S_{xz}, \qquad \lambda = 0.01\,\max\{\text{tr}(S_{xx})/d_x,\ 10^{-8}\}, \tag{3}$$

and deploys

$$f_\ell(x) = \text{Normalize}\big(\hat\alpha\, W^\top(u_\ell(x) - \mu_x) + \mu_z\big), \qquad \hat\alpha = \frac{\sqrt{\|Z_c\|_F^2/n}}{\sqrt{\text{tr}(W^\top S_{xx} W)}}. \tag{4}$$

The map centers and aligns the readout, matches centered target energy, restores the target mean, and normalizes. It is one linear solve without relevance labels or gradients; the fitting objective and numerical guards are in Appendix A.3. Composition uses the same recipe with a separate map for $u_\ell^{\text{Attention}^2}$ (Section 4.3).

Why the scale changes rankings. For a fixed document index, write $y = W^\top(u_\ell(x) - \mu_x)$ and $m = \mu_z$. For $\alpha > 0$ and $\alpha y + m \neq 0$,

$$\big\langle \text{Normalize}(\alpha y + m), d_j \big\rangle = \frac{\alpha}{\|\alpha y + m\|}\big(y^\top d_j + \alpha^{-1} m^\top d_j\big). \tag{5}$$

The shared positive prefactor leaves ranks determined by $y^\top d_j + \alpha^{-1} m^\top d_j$. Scale balances the query-independent *document preference* $m^\top d_j$ against query-specific scores; each pair changes order at most once as positive α varies (Proposition 2). Scale can change rankings; Section 4.6 shows retrieval improving while anchor cosine reconstruction error increases.

3.4 Retrieval settings and composition

Let $v_\ell^r(x)$ be the vector from a dense method r and $z_L(d)$ the native full-depth document vector. We evaluate a rebuilt shallow index and reuse of the deep index:

$$S^r_{\text{rebuild}}(q,d) = (v_\ell^r(q))^\top v_\ell^r(d), \qquad S^r_{\text{reuse}}(q,d) = (v_\ell^r(q))^\top z_L(d). \tag{6}$$

Rebuild uses the same readout and fitted map, if any, on both sides; reuse changes only queries; MaxSim requires a token index. Before scaling, mean restoration and normalization, the shared-map inner product is $(u_q - \mu_x)^\top W W^\top (u_d - \mu_x)$. Ridge changes similarity, not merely orientation (Appendix C.6). Its benefit is measured, not assumed. Appendix A.7 describes an optional unlabeled selector between the calibrated native and Attention^2 readouts.

3.5 When does query error preserve the ranking?

TRIM's ranking analysis uses a fixed index, as in reuse, and writes the shallow query as $z + e$ with full-depth reference z. Query normalization and document-independent score shifts preserve rankings, motivating centered score perturbations:

$$C_D = \frac{1}{N}\sum_{j=1}^{N}(d_j - \bar{d})(d_j - \bar{d})^\top, \qquad E_D(e) = e^\top C_D e. \tag{7}$$

For a nonempty set $\mathcal{P}$ of required strict document orders with margins $\gamma_{ij} = z^\top(d_i - d_j) > 0$:

Theorem 1 (Exact threshold for preserving specified orders)**.** *With $C_D^\dagger$ the Moore–Penrose pseudoinverse of C_D, let*

$$r_* = \min_{(i,j)\in\mathcal{P}} \frac{\gamma_{ij}}{\sqrt{(d_i - d_j)^\top C_D^\dagger (d_i - d_j)}}. \tag{8}$$

If $r < r_*$, every error with $E_D(e) \le r^2$ preserves all orders in $\mathcal{P}$. If $r > r_*$, some error with $E_D(e) \le r^2$ reverses a required pair; at $r = r_*$, a tie is attainable.

The threshold compares margins against sensitivity to query error. Since $\Delta = d_i - d_j$ lies in the range of C_D, Cauchy–Schwarz gives $|e^\top \Delta| \le \sqrt{E_D(e)}\sqrt{\Delta^\top C_D^\dagger \Delta}$, and $e \propto -C_D^\dagger \Delta$ attains the bound (Appendix A.5). Top-K membership uses inside–outside pairs; within-list order uses adjacent pairs. Margins matter: even on a fixed index with fixed relevance labels, vanishing relative centered score error can accompany an nDCG@10 drop from 1 to 0 (Theorem 2, Appendix A.6); fidelity alone cannot certify retrieval quality. The threshold requires full-depth scores; it is a diagnostic, not an online exit rule.

4 Experiments

We examine TRIM across depth (Sections 4.2–4.3), in index reuse and transfer (Sections 4.4–4.5), and through reconstruction and ranking (Section 4.6).

4.1 Setup

Models and document settings. Depth studies use Qwen3-Embedding-4B (36 layers), e5-mistral-7b-instruct (32 layers) and Qwen3-Embedding-0.6B (28 layers); GTE-Qwen2-7B-instruct (28 layers) adds a fourth decoder, and BGE-large, BGE-base and e5-base-v2 are encoder extensions. *Both sides shallow* encodes queries and documents to the same exit and rebuilds the index on the full SciFact, NFCorpus and FiQA corpora from BEIR (Thakur et al., 2021) (240/258/518 test queries). *Deep-index reuse* keeps full-depth native document vectors and changes only the query; it adds SciDocs, TREC-COVID, Touché and Quora (seven datasets). Seven-dataset evaluation uses 9,895 test queries and configuration selection 2,475 development queries.

Preparation. MaxSim and independent Attention^2 involve no fitting. Analytic-scale ridge calibration uses 45,000 unlabeled anchors from the three core corpora (5,000 held out for diagnostics); each model–readout–exit map serves all seven datasets, with Attention^2 + calibration fitted separately by the same recipe. No evaluation query or relevance label enters any fit. Anchors are in-corpus for the three core datasets, out-of-corpus for the other four and MS MARCO.

Metric. We report mean retention across datasets, giving each dataset equal weight:

$$R(c) = \frac{1}{B} \sum_{b=1}^{B} \frac{U_b(c)}{U_b(F)}, \tag{9}$$

Here, B is the number of evaluated datasets, and $U_b(c)$ is the mean nDCG@10 over test queries in dataset b under configuration c. F denotes the same model's native full-depth reference within the same study (published-score qualification: Appendix A.8). We report retention as a percentage. Table 1 also reports absolute macro nDCG@10; retention differences are given in percentage points (pp).

Reading the tables. Tables 1–3 show representative exits selected retrospectively; complete depth results and paired intervals are in Figure 2 and Appendix E. Table 4 reports the separate timed study with its own full-depth reference.

4.2 TRIM ACROSS DEPTH: EVIDENCE, AGGREGATION AND GEOMETRY

Among the three strategies under the fixed fitting recipe, token matching leads early and the calibrated attention readout at intermediate exits; calibrated native pooling catches up or leads later, depending on the model (Table 1, Figure 2; queries and documents at the same exit).

Shallow exits: accessible token evidence. Under the fixed recipe, MaxSim leads both calibrated strategies early: 71.9% at Qwen L12 versus 28.8% for Attention^2 + calibration and 16.7% for calibrated native pooling; Qwen 0.6B L8 reaches 74.4%. Without fitting, Mistral's first-twelve-layer peak is L6/32: 76.1% (0.3966 macro nDCG@10); single-block retention is 51.4–65.1% across the three models. Its lead over both calibrated strategies is significant at every exit from L8 to L12 on both Qwen models and from L8 to L10 on Mistral (simultaneous intervals, Appendix E.8). Useful token evidence remains despite weak native pooling. In matched Qwen 4B controls, macro nDCG@10 rises from 0.275 with input-embedding MaxSim to 0.389/0.422 at L8/L12 (Appendix E.7). Removing the ridge floor raises Qwen 4B L8/L12 Attention^2 retention to 32.8%/54.1%; calibrated content mean reaches 64.2%/69.1% in the matched rerun (Appendix E.6).

Middle exits: re-aggregation adds to calibration. In Table 1, Attention^2 + calibration leads from L20 on both Qwen models and from L14 to L21 on Mistral (the lead alternates in point estimate over L11–L13). At L20, retention is 65.2%, 96.5% and 84.6% on Qwen 4B, Mistral and Qwen 0.6B; native calibration alone gives only 24.9% on Qwen 4B and 52.0% on Qwen 0.6B.

Late exits: recovery through calibration. At later exits, calibrated native pooling leads on Mistral from L22 (98.3% vs. 95.8%) and Qwen 0.6B from L26; Qwen 4B's dense strategies converge (0.3 pp apart at L33, no stable ordering afterwards). These late differences are point estimates. There is no universal switching layer; the readout for a given exit is an empirical, model-dependent choice.

4.3 COMPLEMENTARY EFFECTS OF READOUT AND CALIBRATION

Changing the readout recovers retrieval that calibrating the native vector leaves unrecovered. We compare native and Attention^2 readouts, each with and without calibration. At Qwen L22, both sides shallow, the four cells retain 2.2% (native), 11.2% (Attention^2 only), 32.8% (calibration only) and 80.3% (both): the joint gain exceeds the sum of the separate gains. At Mistral L20, retention rises from 40.9% with native pooling to 67.0% with Attention^2 alone, 90.6% with calibration alone, and 96.5% with both. Calibration recovers most of the gap, with re-pooling adding a further gain. At Qwen L22, the readout gain remains even when calibration uses no anchors from the evaluated corpus (Table 22).

Table 1: **TRIM across depth under three recovery strategies.** Both sides use the same exit on three datasets. Cells give macro nDCG@10 (retained %). Native is uncalibrated; the attention strategy includes geometric calibration.

Model	Exit	Native		MaxSim		Attention² + calibration		Native pooling + calibration	
Qwen 4B	8/36	0.006	(1.4)	**0.389**	**(65.4)**	0.041	(6.9)	0.070	(11.8)
	10/36	0.006	(1.4)	**0.417**	**(70.8)**	0.078	(12.9)	0.074	(12.6)
	12/36	0.006	(1.6)	**0.422**	**(71.9)**	0.166	(28.8)	0.097	(16.7)
	20/36	0.008	(1.9)	0.336	(56.2)	**0.384**	**(65.2)**	0.147	(24.9)
	22/36	0.010	(2.2)	0.344	(57.5)	**0.464**	**(80.3)**	0.196	(32.8)
	24/36	0.050	(8.3)	0.391	(65.9)	**0.496**	**(85.2)**	0.449	(76.0)
	30/36	0.486	(82.8)	0.469	(80.6)	**0.539**	**(93.3)**	0.535	(92.5)
	31/36	0.488	(83.1)	0.472	(81.1)	**0.538**	**(93.0)**	0.537	(92.8)
	33/36	0.505	(86.6)	0.483	(83.1)	0.538	(93.3)	**0.540**	**(93.6)**
Mistral	8/32	0.006	(1.8)	**0.384**	**(74.0)**	0.231	(43.9)	0.160	(32.0)
	9/32	0.007	(2.0)	**0.369**	**(70.9)**	0.284	(55.5)	0.158	(31.3)
	10/32	0.007	(1.9)	**0.349**	**(66.8)**	0.233	(44.8)	0.152	(30.3)
	14/32	0.006	(1.7)	0.262	(49.6)	**0.349**	**(69.6)**	0.152	(30.3)
	18/32	0.018	(4.1)	0.345	(65.8)	**0.466**	**(90.4)**	0.348	(68.0)
	20/32	0.196	(40.9)	0.412	(79.1)	**0.497**	**(96.5)**	0.465	(90.6)
	22/32	0.435	(85.0)	0.432	(83.1)	0.494	(95.8)	**0.505**	**(98.3)**
	26/32	0.486	(95.4)	0.434	(83.9)	0.495	(96.8)	**0.512**	**(99.8)**
	30/32	0.508	(99.1)	0.414	(79.8)	0.508	(99.2)	**0.517**	**(100.9)**
Qwen 0.6B	8/28	0.005	(1.5)	**0.382**	**(74.4)**	0.073	(14.7)	0.030	(6.0)
	20/28	0.065	(12.3)	0.401	(78.8)	**0.424**	**(84.6)**	0.269	(52.0)
	27/28	0.450	(89.7)	0.443	(88.3)	0.469	(94.2)	**0.473**	**(94.7)**

Exit is executed blocks / total blocks. Blue marks the highest observed retention in each row, not statistical significance. Each calibrated readout has its own map; the ridge floor can over-regularize low-variance attention readouts (Appendix E.6). Full-depth macro nDCG: Qwen 4B 0.5763, Mistral 0.5180, Qwen 0.6B 0.4933. Complete grids and intervals are provided in Appendix E.

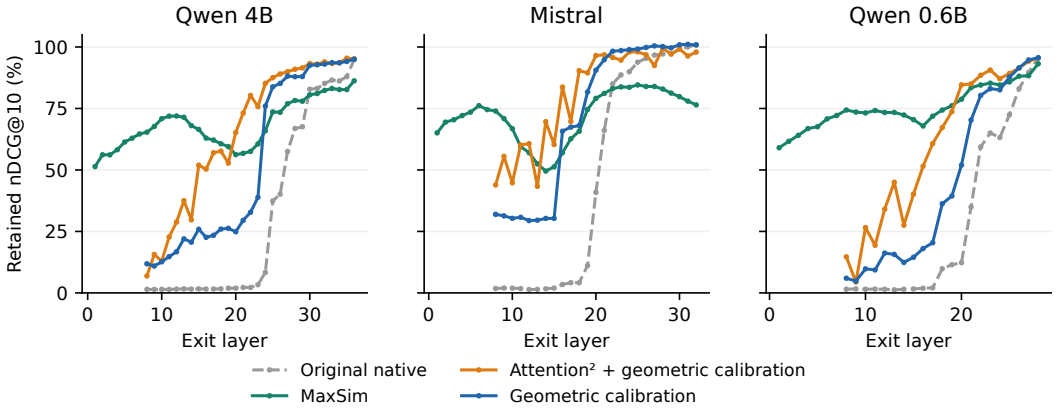


Figure 2: **From token matching to calibrated dense retrieval.** Both sides shallow, three datasets; every measured exit for the strategies of Table 1. The attention curve includes calibration; calibrated curves start at L8. Lines connect measured points without smoothing.

Table 2 reveals a clear depth pattern, where re-pooling helps most at intermediate layers, but its benefit fades toward the final layer. This suggests that attention-based re-pooling is especially useful at intermediate depths. With document inputs matched, replacing shallow documents with full-depth native documents changes retention by about 0.3 pp at Qwen L22 and Mistral L20 (Appendix C.4). Figure 5 shows the complete depth trend.

Table 2: **The attention readout adds quality under the same calibration recipe.** Three-dataset nDCG retention (%). Queries and documents exit together; both readouts are calibrated.

		Encode both sides		
Model	Exit	Native	Attention^2	Gain (pp)
Qwen 4B	16/36	22.6	50.4	+27.8
	22/36	32.8	80.3	+47.5
	30/36	92.5	93.3	+0.7
Mistral	14/32	30.3	69.6	+39.3
	20/32	90.6	96.5	+6.0
	26/32	99.8	96.8	−3.0
Qwen 0.6B	16/28	18.0	51.5	+33.5
	20/28	52.0	84.6	+32.6
	27/28	94.7	94.2	−0.5
GTE	10/28	20.3	68.8	+48.5
	16/28	30.2	79.3	+49.1
	22/28	77.6	86.3	+8.7

Gain: Attention^2 minus native, calculated before rounding; each readout has its own map. All exits: Appendix E. Index reuse: Table 36; input-matched controls: Appendix C.4.

Table 3: **Calibrating native queries while reusing the document index.** Seven-dataset nDCG retention (%). Native pooling in both columns; only the second is calibrated.

		Native query		
Model	Exit	Original	Calibrated	Gain (pp)
Qwen 4B	18/36	0.7	26.3	+25.6
	24/36	5.9	77.2	+71.3
	30/36	77.1	96.0	+18.8
Mistral	16/32	5.1	42.7	+37.6
	21/32	36.0	85.6	+49.6
	26/32	69.6	95.3	+25.7
Qwen 0.6B	16/28	1.2	20.3	+19.1
	20/28	8.2	64.9	+56.7
	26/28	92.0	97.7	+5.7

Documents stay at full depth. Gains are calculated before rounding. All exits and absolute scores: Appendix E.

4.4 Early-exit queries against an unchanged full-depth index

Using TRIM's single-vector methods, calibrated early-exit queries recover most of the full-depth quality without re-encoding documents, and the saving survives the cost of the readout and map.

Calibrating the native query. Table 3 shows that calibration becomes increasingly effective with depth, recovering much of the lost quality at middle and later exits but remaining insufficient earlier, where re-pooling is still needed (Section 4.3). This distinguishes the roles of aggregation and geometry: re-pooling improves how token evidence is compressed into a query vector, whereas calibration improves how that vector ranks documents.

Measured quality–efficiency trade-off. At Qwen L25, Attention^2 + calibration retains 91.6% of full-depth nDCG@10 and over 94% of both Recall@100 and Recall@1000, while reducing query-encoding time by 26.8–30.5% relative to same-length full-depth encoding. Across all nine timed configurations, the measured reduction ranges from 12.3% to 40.1% (Table 4). Timing includes early-exit encoding, readout, calibration and normalization; tokenization, content-mask preparation, transfer and index search are excluded. Full timing details and offline costs are in Appendices B.6 and A.9.

Comparison with training-free alternatives. Attention^2 is both effective and lightweight: it adds no learned pooling parameters and recomputes only the pooling-position attention row, while calibration is fitted offline. At Qwen L25, it improves calibrated content-mean pooling by 6.05 pp across seven datasets; on Quora at Qwen L27, it retains 96.5% of full-depth nDCG@10 and exceeds HTP, the strongest tested external alternative, by 0.95 pp, with both gains significant after correction. Its overhead is included in the measured early-exit savings above. Content mean and HTP remain competitive on Mistral (Table 13; Appendix B.5).

4.5 Scale and architecture

The benefit of Attention^2 extends beyond the settings used in our main experiments. On the 8.8M-passage MS MARCO index, Attention^2 with calibration retains 83.1% of full-depth MRR@10 at Qwen L24, compared with 52.0% for calibrated native pooling. Both maps are fitted using only BEIR anchors. The gain also transfers to CLS-pooled encoders: on BGE-large, replacing calibrated CLS pooling with calibrated Attention^2 raises nDCG@10 retention from 13.4% to 62.3% at L20,

Table 4: **Quality and measured encoding savings for Attention^2 + calibration.** All nine timed server configurations, seven datasets, A800 at batch one; each input length uses its own full-depth timing reference.

		Retrieval quality retained (%)			Encoding time saved (%)	
Model	Exit	nDCG@10	Recall@100	Recall@1000	32 tokens	128 tokens
Qwen 4B	22/36	81.0	87.8	93.5	37.2	33.8
	24/36	89.4	92.8	96.2	31.5	30.0
	25/36	91.6	94.8	96.9	30.5	26.8
Mistral	18/32	83.7	89.1	95.3	40.1	40.0
	20/32	89.1	93.1	96.5	34.2	29.9
	22/32	91.1	94.2	95.8	26.2	24.4
GTE	16/28	85.3	91.3	96.8	37.2	38.2
	21/28	85.3	92.0	96.7	17.3	22.0
	23/28	92.8	95.8	98.3	12.3	13.1

Timed path: forward pass to the exit, readout, map and normalization. Run separately from Table 3, with its own full-depth reference. Intervals and absolute metrics: Appendix B; timing distributions and the cross-length inversion of absolute times: Appendix B.6.

with Attention^2 reaching 97.0% at L23. Frozen terminal attention, a map-free encoder readout, also recovers intermediate-layer retrieval on BGE-base and e5-base-v2 (Appendix E.4), extending the evidence-availability result to encoders. These results show that the readout gain is not specific to either a small index or a particular pooling architecture; the best exit, however, remains model- and dataset-dependent (Appendix E.3).

4.6 Better Reconstruction Does Not Imply Better Retrieval

We now test the distinction between representation fidelity and retrieval utility. We change only the calibration scale α, keeping the fitted map and means fixed. The analytic scale improves the retrieval point estimate in all 12 tested settings, with nine gains remaining significant after correction, even though cosine reconstruction error increases on both fitting and held-out anchors in all 12. The same mismatch appears within individual queries: on Quora at Qwen L22, 928 of the 1,098 queries with higher nDCG also have larger cosine and centered-score errors. Better reconstruction is therefore not a reliable proxy for better retrieval. TRIM's ranking analysis explains why: retrieval depends on whether score changes alter document rankings, not on reconstruction error alone (see Proposition 2; Theorems 1 and 2; Figure 4; Appendix D.1).

5 Conclusion

TRIM shows that weak intermediate-layer retrieval need not indicate missing information: shallow token states already support retrieval, re-aggregation improves calibrated single vectors at middle exits, and calibration alone often recovers later-layer quality. With the encoder frozen, MaxSim bypasses pooling, Attention^2 re-aggregates evidence without fitted parameters, and calibration adjusts scoring geometry. Their value varies with depth, document representation and preparation cost. These complementary roles support depth-aware choices, not universally disjoint winning ranges.

6 Limitations

The depth grid is a transparent summary of existing experiments, not a prospective exit policy. MaxSim lacks matched serving-cost measurements, and document-ingestion time is not timed. Matched small-encoder timing and throughput, and independent new-domain tests remain open (quality comparisons: Appendix B.4).

AI USE STATEMENT

Generative AI was primarily utilized to support manuscript preparation, assisting with language refinement and the verification of presented experimental results and tables. Generative AI was also used to search for and identify relevant literature. The authors oversaw all research directions, conducted the analyses, and assume full responsibility for the content and conclusions.

REPRODUCIBILITY STATEMENT

Proofs and statistical procedures are in Appendix A. Model inputs and evaluation settings are in Appendix B.1; bilateral controls, scale analyses and complete depth results are in Appendices C–E. Code, configurations, per-query results and reproduction instructions will be released with the camera-ready version.

REFERENCES

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=IW1PR7vEBf.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. doi: 10.48550/arXiv.2303.08112. URL https://arxiv.org/abs/2303.08112.

Yinqiong Cai, Keping Bi, Yixing Fan, Jiafeng Guo, Wei Chen, and Xueqi Cheng. L^2R: Lifelong learning for first-stage retrieval with backward-compatible representations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 183–192. ACM, 2023. doi: 10.1145/3583780.3614947. URL https://doi.org/10.1145/3583780.3614947.

Qian Chen, Wen Wang, Qinglin Zhang, Siqi Zheng, Chong Deng, Hai Yu, Jiaqing Liu, Yukun Ma, and Chong Zhang. Ditto: A simple and efficient approach to improve sentence embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5868–5875. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.359. URL https://aclanthology.org/2023.emnlp-main.359/.

Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A^2-Nets: Double attention networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 352–361, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/e165421110ba03099a1c0393373c5b43-Abstract.html.

Xueying Ding, Xingyue Huang, Mingxuan Ju, Liam Collins, Yozen Liu, Leman Akoglu, Neil Shah, and Tong Zhao. Hierarchical token prepending: Enhancing information flow in decoder-based LLM embeddings. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19047–19066. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.870. URL https://aclanthology.org/2026.acl-long.870/.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. LayerSkip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12622–12642. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.681. URL https://aclanthology.org/2024.acl-long.681/.

Yuchen Fu, Zifeng Cheng, Zhiwei Jiang, Zhonghui Wang, Yafeng Yin, Zhengliang Li, and Qing Gu. Token prepending: A training-free approach for eliciting better sentence embeddings from

LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3168–3181. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.159. URL https://aclanthology.org/2025.acl-long.159/.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3182–3196. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.181. URL https://aclanthology.org/2024.findings-emnlp.181/.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249, 2022. doi: 10.52202/068431-2192. URL https://papers.nips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af9993100e42-Abstract-Conference.html.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c4bf73386022473a652a18941e9ea6f8-Abstract-Conference.html.

Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. ESE: Espresso sentence embeddings. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/60dc26558762425a465cb0409fc3dc52-Abstract-Conference.html.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. doi: 10.48550/arXiv.2308.03281. URL https://arxiv.org/abs/2308.03281.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.272. URL https://aclanthology.org/2022.naacl-main.272/.

Jinsong Shu, Jinyong Wen, Baokun Wang, Zhongle Xie, Lidan Shou, Weiqiang Wang, and Gang Chen. FastE: Readout-triggered token compression for LLM embedding inference. *arXiv preprint arXiv:2609.08407*, 2026. doi: 10.48550/arXiv.2609.08407. URL https://arxiv.org/abs/2609.08407.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55854–55875. PMLR, 2025. URL https://proceedings.mlr.press/v267/skean25a.html.

Luca Soldaini and Alessandro Moschitti. The cascade transformer: an application for efficient answer sentence selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5697–5708. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.504. URL https://aclanthology.org/2020.acl-main.504/.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. Repetition improves language model embeddings. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e9b0c9480b03f9a6bca24ed8c08ee058-Abstract-Conference.html.

Yixuan Tang and Yi Yang. KV-embedding: Training-free text embedding via internal KV re-routing in decoder-only LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11773–11794. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.540. URL https://aclanthology.org/2026.acl-long.540/.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

Harshil Vejendla. Drift-adapter: A practical approach to near zero-downtime embedding model upgrades in vector databases. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15938–15949. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.805. URL https://aclanthology.org/2025.emnlp-main.805/.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.642. URL https://aclanthology.org/2024.acl-long.642/.

Yuxuan Wang and Lyu Hong. Query encoder distillation via embedding alignment is a strong baseline method to boost dense retriever online efficiency. In *Proceedings of the Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pp. 290–298. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.sustainlp-1.23. URL https://aclanthology.org/2023.sustainlp-1.23/.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2246–2251. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.204. URL https://aclanthology.org/2020.acl-main.204/.

Lingling Xu, Haoran Xie, Zongxi Li, Fu Lee Wang, Weiming Wang, and Qing Li. Contrastive learning models for sentence representations. *ACM Transactions on Intelligent Systems and Technology*, 14(4):67:1–67:34, 2023. doi: 10.1145/3593590. URL https://doi.org/10.1145/3593590.

Jianrui Zhang, Hyun Jung Lee, Sukanta Ganguly, Tae-Eui Kam, Donghyun Kim, and Yong Jae Lee. Your embedding model is SMARTer than you think. *arXiv preprint arXiv:2605.24938*, 2026a. doi: 10.48550/arXiv.2605.24938. URL https://arxiv.org/abs/2605.24938.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. doi: 10.48550/arXiv.2506.05176. URL https://arxiv.org/abs/2506.05176.

Yeqin Zhang, Yunfei Wang, Jiaxuan Chen, Ke Qin, Yizheng Zhao, and Cam-Tu Nguyen. LLM-based embeddings: Attention values encode sentence semantics better than hidden states. *arXiv preprint arXiv:2602.01572*, 2026b. doi: 10.48550/arXiv.2602.01572. URL https://arxiv.org/abs/2602.01572v2.

APPENDIX CONTENTS

A Proofs and evaluation details **15**
A.1 Token matching before pooling . . . 15
A.2 The best score fit available to a single query vector . . . 15
A.3 Attention^2 and ridge calibration . . . 16
A.4 Calibration scale and document preference . . . 16
A.5 Exact threshold for preserving document order . . . 17
A.6 A small-error example with maximal nDCG loss . . . 17
A.7 An empirical heuristic for two calibrated readouts . . . 18
A.8 Evaluation and uncertainty . . . 18
A.9 Encoding cost and amortization . . . 19

B Experimental settings and complete comparisons **19**
B.1 Models, coordinates, data and fitting . . . 19
B.2 Absolute metrics and aggregation sensitivity . . . 20
B.3 Attention versus mean-pooling baselines . . . 22
B.4 Small encoder comparisons . . . 23
B.5 Raw and calibrated neighboring-method controls . . . 23
B.6 Measured encoding-time savings . . . 24

C Both-sided shallow retrieval and readout interpretation **25**
C.1 Bilateral retrieval and document controls . . . 25
C.2 Calibration scale under bilateral retrieval . . . 25
C.3 Readout gains and document-side effects . . . 26
C.4 Input-matched deep documents and exact difference accounting . . . 27
C.5 Recall alongside nDCG . . . 28
C.6 Why a shared calibration can change bilateral similarity . . . 28
C.7 Attention-squared means a second aggregation . . . 29
C.8 Empirical connection to score-level theory . . . 29

D Additional depth, scale and transfer analyses **30**
D.1 Scale, restored means and retrieval geometry . . . 30

E Complete results across layers and retrieval settings **31**
E.1 Both queries and documents shallow: all measured depths . . . 31
E.2 Deep-index reuse: readouts with and without calibration . . . 33
E.3 Detailed scale, neighboring-method and encoder results . . . 35
E.4 Recovering encoder representations with a frozen readout . . . 36
E.5 Readout gains across document settings . . . 37
E.6 Effect of the regularization floor at shallow exits . . . 38
E.7 Retrieval gains beyond input embeddings . . . 39
E.8 Paired shallow-layer comparisons . . . 40

A PROOFS AND EVALUATION DETAILS

Complete depth results are in Appendix E; readout controls and timing in Appendix B; document controls in Appendix C; encoder extensions in Appendix E.4. The proofs below connect token aggregation and calibrated geometry to retrieval scores and ranking stability.

A.1 TOKEN MATCHING BEFORE POOLING

Pooling forces all query tokens to use the same document-token match. For query weights $\beta_i \geq 0$ with $\sum_i \beta_i = 1$,

$$\max_j \left\langle \sum_i \beta_i q_i, t_j \right\rangle \leq \sum_i \beta_i \max_j \langle q_i, t_j \rangle. \tag{10}$$

For each document token t_j, every term $\langle q_i, t_j \rangle$ is bounded by its individual maximum. Summing with nonnegative weights and maximizing over j proves the inequality. Equality holds exactly when one document token maximizes every positively weighted term: the nonnegative gaps must all vanish. MaxSim retains the ability of different query tokens to match different document tokens. The bound concerns matching scores; document-dependent gaps and vector normalization prevent it from implying a retrieval-ranking guarantee.

A.2 THE BEST SCORE FIT AVAILABLE TO A SINGLE QUERY VECTOR

A fixed document index can produce only score vectors in the span of its columns. The proof separates a desired score vector into the part this index can produce and the remaining error.

Fix dense document matrix D and desired score vector s, such as a MaxSim score vector on the same document identities. Let $J = I - \mathbf{1}\mathbf{1}^\top/N$, $D_c = JD$ and $s_c = Js$. The exact unconstrained score-fit statement is

Proposition 1 (Best score fit with one vector)**.** *Let P_{D_c} be the orthogonal projector onto the column space of D_c, i.e., the set of vectors of the form $D_c v$. Then*

$$\min_{v \in \mathbb{R}^d} \|D_c v - s_c\|_2^2 = \|(I - P_{D_c}) s_c\|_2^2. \tag{11}$$

Orthogonally decompose

$$s_c = P_{D_c} s_c + (I - P_{D_c}) s_c.$$

Since $D_c v - P_{D_c} s_c$ lies in $\mathrm{col}(D_c)$ and the second component lies in its orthogonal complement,

$$\|D_c v - s_c\|_2^2 = \|D_c v - P_{D_c} s_c\|_2^2 + \|(I - P_{D_c}) s_c\|_2^2.$$

The first term can be zero by choosing a least-squares solution. This proves Proposition 1. Restricting v to normalized vectors or to the output of a fitted readout cannot lower the unconstrained minimum. The result concerns **centered numerical scores**, not the ability to reproduce rankings: two score vectors can differ outside a linear subspace and still induce identical total rankings. It also does not imply that every shallow layer has a nonzero or large projection residual.

A.3 ATTENTION^2 AND RIDGE CALIBRATION

We give the attention weights and the closed-form calibration parameters used in the experiments.

Attention weights. Let p be the designated pooling position, $\mathcal{V}_\ell(p)$ its valid attention positions, and $\mathbf{Q}_p^{(\ell,h)}, \mathbf{K}_j^{(\ell,h)}$ the query and key vectors of head h with dimension d_h. The weights used in Equation 2 are

$$a_{pj}^{(\ell,h)} = \mathrm{softmax}_{j\in\mathcal{V}_\ell(p)}\left(\frac{\langle \mathbf{Q}_p^{(\ell,h)}, \mathbf{K}_j^{(\ell,h)}\rangle}{\sqrt{d_h}}\right), \qquad w_j = \frac{1}{H}\sum_{h=1}^{H} a_{pj}^{(\ell,h)}. \tag{12}$$

Calibration parameters. The ridge penalty and analytic scale are defined in Equations 3 and 4. In implementation, the scale denominator is bounded below by 10^{-12}. The covariance uses $n-1$, whereas target energy uses n; when this numerical guard is inactive, $\hat{\alpha} = \sqrt{(n-1)/n}\,\|Z_c\|_F/\|X_cW\|_F$.

For row-oriented anchor matrices, consider

$$\mathcal{L}_{\mathrm{ridge}}(W,\mathbf{b}) = \frac{1}{n-1}\|XW + \mathbf{1}\mathbf{b}^\top - Z\|_F^2 + \lambda\|W\|_F^2.$$

The unpenalized intercept satisfies $\mathbf{b} = \mu_z - W^\top\mu_x$. Substitution yields the centered objective. Differentiating gives

$$S_{xx}W + \lambda W = S_{xz}.$$

For $\lambda > 0$, $S_{xx} + \lambda I$ is positive definite and the solution is unique. This solution minimizes the pre-normalization ridge loss. The scale in Equation 4 matches centered energy; retrieval quality is evaluated separately.

A.4 CALIBRATION SCALE AND DOCUMENT PREFERENCE

With $y = W^\top(u_\ell(x) - \mu_x)$ and $m = \mu_z$, for $\alpha > 0$ and $\alpha y + m \neq 0$, the pairwise consequence is

Proposition 2 (Why calibration scale can change a ranking)**.** *For a document pair (i,j), let $\Delta_{ij} = d_i - d_j$, $a_{ij} = y^\top\Delta_{ij}$ and $c_{ij} = m^\top\Delta_{ij}$. Their order is determined by the sign of $\alpha a_{ij} + c_{ij}$. If $a_{ij} \neq 0$, the pair can change order at most once for $\alpha > 0$, at $\alpha = -c_{ij}/a_{ij}$ when this value is positive.*

Equation 5 factors the normalized score into a positive prefactor common to all documents and $y^\top d_j + \alpha^{-1}m^\top d_j$. Thus the sign of a pairwise score difference is the sign of $\alpha a_{ij} + c_{ij}$, proving Proposition 2, including ties. If $a_{ij} = 0$, the pair never crosses and ties at every scale exactly when $c_{ij} = 0$.

If $C_D m = 0$, then $m^\top C_D m = N^{-1}\sum_j [m^\top(d_j - \bar{d})]^2 = 0$, so the document preference is constant across the index. Removing it cannot change ranks, and positive scales are equivalent. Otherwise the preference can affect ranking.

A.5 EXACT THRESHOLD FOR PRESERVING DOCUMENT ORDER

The ranking threshold compares each pair's score gap with its sensitivity to query error. The sensitivity is sometimes called a leverage term. The proof shows both that errors below the threshold are safe and that an error at the threshold can create a tie.

For additive error e,

$$\frac{1}{N}\sum_j [e^\top (d_j - \bar{d})]^2 = e^\top C_D e.$$

The uncentered error additionally contains $(e^\top \bar{d})^2$. That term shifts every document score by the same amount for the query and has zero rank effect.

Let $v_j = d_j - \bar{d}$. The range of $C_D = N^{-1}\sum_j v_j v_j^\top$ is the span of the v_j, because its null space is their common orthogonal complement. Therefore every $\Delta = d_i - d_j = v_i - v_j$ belongs to its range. Applying Cauchy–Schwarz on that range,

$$|e^\top \Delta| = |\langle C_D^{1/2} e, C_D^{\dagger/2}\Delta\rangle| \le \sqrt{e^\top C_D e}\sqrt{\Delta^\top C_D^\dagger \Delta}.$$

For a strict reference pair with margin $\gamma > 0$, the perturbed margin is at least $\gamma - r\sqrt{\Delta^\top C_D^\dagger \Delta}$. All specified margins stay positive when $r < r_*$. For a minimizing pair, choose

$$e_* = -\frac{r C_D^\dagger \Delta}{\sqrt{\Delta^\top C_D^\dagger \Delta}}.$$

Then $e_*^\top C_D e_* = r^2$ and its pairwise change attains the negative bound. At $r = r_*$ the pair ties; above it, it reverses. This proves Theorem 1. Since each required margin is positive, its document difference is nonzero and its leverage denominator is positive.

The threshold is exact over the additive-error ellipsoid; a model may realize only a subset of these errors. Boundary ties follow the declared document-ID rule. The proof uses the exact pseudoinverse, since truncating positive eigenvalues can discard ranking-relevant directions.

A.6 A SMALL-ERROR EXAMPLE WITH MAXIMAL NDCG LOSS

For $\rho(e, z) = e^\top C_D e/(z^\top C_D z)$ with positive denominator, the following result shows why a margin condition is essential.

Theorem 2 (Small score error need not preserve nDCG)**.** *There exists a fixed index of unit document vectors with positive-definite C_D and fixed binary relevance labels for which a sequence of unit query pairs $(z, z + e)$ satisfies $\rho(e, z) \to 0$, while nDCG@10 falls from 1 to 0.*

This construction keeps the documents and their relevance fixed. Only the query changes. As scores approach ties, increasingly small changes can move the relevant document out of the top ten.

Take twelve fixed unit documents in $\mathbb{R}^3$:

$$d_j = (\cos(j\pi/30), \sin(j\pi/30), 0), \quad j = 0, \ldots, 10, \qquad d_{11} = (0, 0, -1).$$

Only d_0 is relevant. For $0 < \epsilon < 1/2$, let

$$z_\epsilon = (\epsilon, 0, \sqrt{1-\epsilon^2}), \qquad \hat{z}_\epsilon = (-\epsilon, 0, \sqrt{1-\epsilon^2}).$$

Both queries are unit vectors. The first eleven document scores are strictly decreasing under z_ϵ and strictly increasing under $\hat{z}_\epsilon$. The last document is below all eleven under both queries. Hence d_0 moves from rank one to rank eleven, and nDCG@10 drops from one to zero without ties.

Three distinct arc points affinely span the plane containing the arc, and d_{11} lies outside it. Thus the centered covariance is positive definite. The error is $e_\epsilon = (-2\epsilon, 0, 0)$, so

$$e_\epsilon^\top C_D e_\epsilon = 4\epsilon^2 (C_D)_{11} \to 0, \qquad z_\epsilon^\top C_D z_\epsilon \to (C_D)_{33} = 11/144 > 0.$$

Hence $\rho(e_\epsilon, z_\epsilon) \to 0$, while the nDCG loss remains maximal. This proves Theorem 2: vanishing error permits maximal loss as queries approach a ranking boundary, consistent with the margin condition in Theorem 1.

A.7 An empirical heuristic for two calibrated readouts

At a fixed exit and document index, we test whether unlabeled anchors can select between calibrated native pooling and calibrated Attention^2. Five-fold cross-fitting fits a separate ridge map for each candidate on four anchor folds and evaluates it on the fifth. For candidate a and held-out anchor i, define

$$\tilde{z}_{a,i} = \mathrm{Normalize}\big(\beta_a W_a^\top (x_{a,i} - \mu_a) + \mu_z\big), \qquad \beta_a^2 = \frac{\sum_{i \in V} \|z_i - \mu_z\|^2}{\sum_{i \in V} \|W_a^\top (x_{a,i} - \mu_a)\|^2},$$

Here $\beta_a > 0$, V is the evaluated anchor fold, and means/maps come from the other folds. Pool the centered score energies across folds and choose

$$\widehat{R}_a = \frac{\sum_{i,j} [(d_j - \bar{d})^\top (\tilde{z}_{a,i} - z_i)]^2}{\sum_{i,j} [(d_j - \bar{d})^\top z_i]^2}, \qquad a^* = \arg \min_{a \in \{\mathrm{native}, \mathit{Attention}^2\}} \widehat{R}_a.$$

The choice is stored per index and exit; fitting and selection are offline, with no relevance labels or query-time decision. The diagnostic scale β_a is used only in the selection criterion. Retrieval in Table 5 uses unit-scale calibration, $\alpha = 1$.

Table 5: **Two-readout heuristic across tested depths.** Seven-dataset native/Attention^2 comparison with deployed $\alpha = 1$. A decision is one dataset–exit pair.

Model	Dataset–exit decisions	Higher-nDCG readout selected	Agreement (%)	Lower-nDCG selections with interval excluding zero
Qwen 4B	175	147	84.0	7
Mistral	147	136	92.5	5
Qwen 0.6B	119	108	90.8	1

Agreement is with the higher measured nDCG point estimate. The last column uses the original pointwise paired intervals, not a new family-adjusted test.

The heuristic selects the higher-nDCG readout in 391/441 decisions (88.7%); 13 selections favor the lower-quality arm with pointwise intervals excluding zero. This agreement is empirical and limited to these two calibrated readouts. Theorem 2 shows why centered score error alone gives no general ranking guarantee. The heuristic does not choose MaxSim or the main-table exits.

A.8 Evaluation and uncertainty

The bilateral depth study, seven-dataset index-reuse study and Quora readout comparison use separate maps and full-depth references. Their scores are compared within each study. Inputs and readouts are specified in Appendix B.1; the complete timing protocol is in Appendix B.6.

All intervals pair queries within each dataset and hold fitted maps fixed. Bilateral comparisons use 5,000 stratified bootstrap replicates with fixed full-depth denominator means. The seven-dataset comparison uses 10,000 replicates and resamples the full-depth denominator; its pooling intervals adjust over eighteen comparisons. The twelve-cell scale-gain analysis uses a twelve-comparison adjustment. The eight-configuration scale-control analysis uses 20,000 replicates, with sixteen comparisons against fixed scale/mean removal and eight separate comparisons against unit scale. Quora neighboring-readout intervals adjust over three selected families; the Mistral and anchor-source

supplements state their own interval procedures beside their tables. These are separate statistical families, conditional on the evaluated datasets and fitted maps. The bootstrap does not include variation from refitting the maps.

Full-depth references. Retained quality is relative to our measured full-depth runs under the stated inputs, rather than a reproduced leaderboard result. Different studies retain their own denominators.

A.9 Encoding cost and amortization

Let Q be the number of queries and M the number of documents that are ingested or refreshed. Let t_L^q, t_ℓ^q denote full- and shallow-depth query encoding time, and define t_L^d, t_ℓ^d analogously for documents. The chosen readout and any calibration are included in these times, while anchor encoding and map fitting are collected in C_{prep}. Independent MaxSim and Attention^2 need no calibration anchors; building their document indexes still has a cost. Relative to full-depth encoding, rebuild mode has net encoding-time benefit

$$Q(t_L^q - t_\ell^q) + M(t_L^d - t_\ell^d) - C_{\text{prep}}, \tag{13}$$

before any one-time index migration or construction cost. In reuse mode, if c is the unchanged per-query work outside the encoder, then

$$s_{\text{enc}} = 1 - \frac{t_\ell^q}{t_L^q}, \qquad s_{\text{total}} = s_{\text{enc}} \frac{t_L^q}{t_L^q + c}. \tag{14}$$

We measure query encoding time; document-ingestion savings in rebuild mode are not measured. Both dense modes store $O(Nd)$ values. A dense calibration map costs $O(d^2)$ per vector and $O(nd^2 + d^3)$ to fit. MaxSim stores multiple token vectors per document, approximately $O(N\bar{T}_d d_t)$ values, and therefore trades higher storage and scoring cost for token-level matching flexibility. Layer fractions report skipped transformer blocks; they should not be interpreted as measured latency or memory savings.

B Experimental settings and complete comparisons

Unless marked raw, the readouts below include geometric calibration. Independent, uncalibrated readouts are reported in Appendix E.

B.1 Models, coordinates, data and fitting

We use Qwen3-Embedding-4B and 0.6B, e5-mistral-7b-instruct, and gte-Qwen2-7B-instruct. Their full depths and embedding dimensions are 36/2560, 28/1024, 32/4096 and 28/3584, respectively. The seven-dataset study uses maximum length 192, BF16 encoding and FP32 vector normalization, with encoding batches of 4 (target models) and 24 (0.6B). GTE source vectors are stored in FP16 before conversion to FP32 for calibration.

Queries and query-like document anchors use the instruction "Given a web search query, retrieve relevant passages that answer the query," formatted as `Instruct: [instruction]` followed by a newline and `Query: [text]`. Qwen places the text directly after the colon, uses tokenizer special tokens and right padding, with no extra EOS. Mistral and GTE insert a space before the text, append EOS and left-pad; GTE disables automatic special tokens and uses bidirectional attention. Native pooling takes the last valid token. Intermediate states are measured before final-model normalization.

Content mean averages text-span tokens, excluding prompts, special tokens and padding; valid mean includes every non-padding token. Empty content uses Attention^2. For Attention^2, we retain the backbone attention implementation and compute only the pooling-position row from its Q/K projections, respecting normalization, positional encoding and grouped-query attention. Its head average weights the block-output states. Each readout is L2-normalized and receives its own ridge map (Appendix A.3).

Table 6: Anchor composition. The held-out rows are diagnostics only.

Anchor corpus	Total anchors	Fitting anchors	Held-out anchors
SciFact	3883	3510	373
NFCorpus	2725	2449	276
FiQA	43392	39041	4351

Of 50,000 unlabeled document anchors, 45,000 fit each map and 5,000 are held out. One map per model, exit and readout serves all seven datasets. Four datasets supply no anchors but retain equal weight in the macro average; they were seen during the project and are not new blind test domains.

Table 7: Query splits shared across target-model evaluations.

Dataset	Development queries	Test queries
SciFact	60	240
NFCorpus	65	258
FiQA	130	518
SciDocs	200	800
TREC-COVID	10	40
Touché-2020	10	39
Quora	2000	8000

FP32 inner-product scores are ordered by descending score and then descending document-ID string to resolve ties, including at the top-1000 boundary. We report nDCG@10 and Recall@100/1000. Bootstrap procedures are given in Appendix A.8.

Calibration cost. Each map stores $d_x d_z$ entries, two means and a scale. The Qwen 4B matrix occupies 25 MiB in FP32. Fitting a map to 45,000 precomputed anchor pairs takes 1.97–2.40 s on an RTX 5080; anchor encoding and retrieval evaluation are excluded. The complete query-encoding measurements use an A800-SXM4-40GB (Appendix B.6).

Supplementary Mistral L1–L7 single-vector measurements use an RTX 5080 (Table 26). At L12, differences from the archived evaluation are below 0.006 nDCG per dataset, with some query rankings differing.

B.2 Absolute metrics and aggregation sensitivity

Tables 8–10 use analytic-scale ridge calibration except for the full-depth native reference. The small encoder is Qwen3-Embedding-0.6B at full depth, calibrated separately to each target index. Scores average datasets equally; retention averages per-dataset ratios to the target model's full-depth reference (1.0000 = 100%).

Table 8: Qwen 4B: seven-dataset retrieval against the full-depth document index. All query readouts except the full-depth reference include calibration. Retention is a ratio; 1.0000 denotes 100%.

	nDCG@10		Recall@100		Recall@1000	
Readout and exit	Score	Retention (ratio)	Score	Retention (ratio)	Score	Retention (ratio)
Full-depth native	0.5456	1.0000	0.6275	1.0000	0.8442	1.0000
Qwen 0.6B + calibration	0.4812	0.8630	0.5853	0.9086	0.8135	0.9583
Native pooling L22	0.1964	0.3353	0.4060	0.5508	0.6581	0.7459
Attention^2 L22	0.4541	0.8097	0.5752	0.8780	0.7991	0.9355
Valid-token mean L22	0.1957	0.3190	0.3487	0.4892	0.5820	0.6688
Content-token mean L22	0.4402	0.7922	0.5697	0.8638	0.7913	0.9238
Native pooling L24	0.4355	0.7760	0.5662	0.8559	0.7889	0.9225
Attention^2 L24	0.4946	0.8944	0.5975	0.9279	0.8178	0.9622
Valid-token mean L24	0.2760	0.4595	0.4341	0.6269	0.6650	0.7680
Content-token mean L24	0.4684	0.8416	0.5811	0.8934	0.8034	0.9412
Native pooling L25	0.4744	0.8529	0.5946	0.9142	0.8140	0.9555

Table 8 (continued)

Readout and exit	nDCG@10 Score	nDCG@10 Retention (ratio)	Recall@100 Score	Recall@100 Retention (ratio)	Recall@1000 Score	Recall@1000 Retention (ratio)
Attention^2 L25	0.5047	0.9157	0.6081	0.9475	0.8234	0.9694
Valid-token mean L25	0.3264	0.5597	0.4807	0.7026	0.7030	0.8146
Content-token mean L25	0.4761	0.8552	0.5848	0.9005	0.8060	0.9455

Table 9: Mistral 7B: seven-dataset retrieval against the full-depth document index. All query readouts except the full-depth reference include calibration. Retention is a ratio; 1.0000 denotes 100%.

Readout and exit	nDCG@10 Score	nDCG@10 Retention (ratio)	Recall@100 Score	Recall@100 Retention (ratio)	Recall@1000 Score	Recall@1000 Retention (ratio)
Full-depth native	0.5075	1.0000	0.5709	1.0000	0.7776	1.0000
Qwen 0.6B + calibration	0.3831	0.7571	0.5050	0.8491	0.7317	0.9236
Native pooling L18	0.2626	0.4995	0.4375	0.6851	0.6598	0.8135
Attention^2 L18	0.4355	0.8370	0.5307	0.8906	0.7484	0.9529
Valid-token mean L18	0.0940	0.1753	0.1857	0.2794	0.3513	0.4301
Content-token mean L18	0.4134	0.8086	0.5296	0.8920	0.7473	0.9466
Native pooling L20	0.3945	0.7558	0.5163	0.8500	0.7246	0.9108
Attention^2 L20	0.4607	0.8909	0.5486	0.9310	0.7577	0.9654
Valid-token mean L20	0.1583	0.3028	0.2457	0.3870	0.4206	0.5221
Content-token mean L20	0.4508	0.8879	0.5505	0.9446	0.7591	0.9681
Native pooling L22	0.4715	0.9138	0.5501	0.9383	0.7608	0.9690
Attention^2 L22	0.4654	0.9114	0.5504	0.9416	0.7546	0.9583
Valid-token mean L22	0.1992	0.3846	0.2892	0.4769	0.4805	0.6007
Content-token mean L22	0.4566	0.8965	0.5553	0.9609	0.7635	0.9757

Table 10: GTE 7B: seven-dataset retrieval against the full-depth document index. All query readouts except the full-depth reference include calibration. Retention is a ratio; 1.0000 denotes 100%.

Readout and exit	nDCG@10 Score	nDCG@10 Retention (ratio)	Recall@100 Score	Recall@100 Retention (ratio)	Recall@1000 Score	Recall@1000 Retention (ratio)
Full-depth native	0.5651	1.0000	0.6395	1.0000	0.8457	1.0000
Qwen 0.6B + calibration	0.4839	0.8650	0.5870	0.8963	0.8097	0.9507
Native pooling L16	0.1667	0.2507	0.3660	0.4888	0.6339	0.7225
Attention^2 L16	0.4864	0.8531	0.5997	0.9129	0.8231	0.9683
Valid-token mean L16	0.3651	0.6120	0.5249	0.7854	0.7685	0.8995
Content-token mean L16	0.4971	0.8714	0.6050	0.9244	0.8250	0.9704
Native pooling L21	0.2852	0.4732	0.4888	0.6794	0.7331	0.8451
Attention^2 L21	0.4913	0.8534	0.6038	0.9200	0.8223	0.9673
Valid-token mean L21	0.4284	0.7193	0.5721	0.8513	0.7969	0.9332
Content-token mean L21	0.5032	0.8834	0.6093	0.9273	0.8248	0.9699
Native pooling L23	0.5032	0.8818	0.6134	0.9361	0.8265	0.9727
Attention^2 L23	0.5305	0.9283	0.6213	0.9576	0.8334	0.9833
Valid-token mean L23	0.4946	0.8488	0.6079	0.9276	0.8255	0.9725
Content-token mean L23	0.5267	0.9299	0.6204	0.9557	0.8332	0.9825

Aggregation matters: at Qwen L25, the mean of per-dataset retention ratios is 91.57%, while the ratio of macro nDCG means is 92.51%. The corresponding gains over content mean are 6.05 and 5.24 pp. We use the former throughout; the absolute macro nDCG gain is 0.0286.

Table 11: Qwen 4B, layer 25: seven-dataset quality with calibrated attention and content-token mean pooling. Gain is attention minus content mean in nDCG@10 retention points, with pointwise 95% intervals from 5,000 paired resamples per dataset and resampled full-depth denominators.

Dataset	Test queries	Attention nDCG@10	Content mean nDCG@10	Attention retention (ratio)	Gain (pp) [95% CI]
SciFact	240	0.7433	0.7161	0.9658	3.5329 [+0.596, +6.468]
NFCorpus	258	0.3841	0.3450	0.9574	9.7343 [+6.049, +13.930]
FiQA	518	0.5180	0.4677	0.9273	9.0141 [+5.916, +12.124]
SciDocs	800	0.2111	0.1863	0.8457	9.9357 [+6.688, +13.354]
TREC-COVID	40	0.5815	0.5622	0.8428	2.8083 [-5.540, +11.626]
Touché-2020	39	0.2434	0.2322	0.9046	4.1481 [-5.246, +12.892]
Quora	8000	0.8518	0.8235	0.9666	3.2115 [+2.786, +3.639]

Table 12: Pointwise 95% intervals for retained nDCG@10 (%) at the nine server configurations. Recall and timing are in Table 4; the paired bootstrap resamples the full-depth denominator.

Model	Exit	Retained nDCG@10 [95% interval]
Qwen 4B	L22	81.0 [78.8, 83.1]
	L24	89.4 [87.3, 91.5]
	L25	91.6 [89.7, 93.4]
Mistral	L18	83.7 [81.5, 86.0]
	L20	89.1 [87.5, 90.7]
	L22	91.1 [89.5, 92.8]
GTE	L16	85.3 [82.8, 88.0]
	L21	85.3 [82.9, 87.9]
	L23	92.8 [90.2, 96.0]

B.3 Attention versus mean-pooling baselines

Valid-token mean averages all non-padding tokens; content-token mean excludes prompts, special tokens and padding. Each readout has its own calibration map, so differences compare calibrated systems.

Table 13: Attention readout versus mean-pooling baselines on seven datasets. Differences are Attention^2 minus the named baseline in nDCG@10 retention percentage points (pp); positive values favor attention. Intervals use the eighteen-comparison Bonferroni adjustment.

Model	Exit layer	Baseline pooling	Retention difference (pp)	Bonferroni-adjusted interval (pp)
Qwen 4B	L22	Valid-token mean	+49.077	[+44.965, +52.782]
Qwen 4B	L22	Content-token mean	+1.751	[-1.388, +4.670]
Qwen 4B	L24	Valid-token mean	+43.491	[+39.238, +47.597]
Qwen 4B	L24	Content-token mean	+5.279	[+1.607, +8.527]
Qwen 4B	L25	Valid-token mean	+35.604	[+30.627, +40.175]
Qwen 4B	L25	Content-token mean	+6.055	[+2.945, +9.055]
Mistral 7B	L18	Valid-token mean	+66.170	[+62.318, +70.379]
Mistral 7B	L18	Content-token mean	+2.841	[-0.447, +6.670]
Mistral 7B	L20	Valid-token mean	+58.803	[+54.702, +62.953]
Mistral 7B	L20	Content-token mean	+0.295	[-2.641, +3.490]
Mistral 7B	L22	Valid-token mean	+52.675	[+48.354, +56.838]
Mistral 7B	L22	Content-token mean	+1.487	[-1.243, +4.166]
GTE 7B	L16	Valid-token mean	+24.108	[+19.796, +28.998]
GTE 7B	L16	Content-token mean	-1.833	[-4.473, +0.846]
GTE 7B	L21	Valid-token mean	+13.416	[+9.608, +18.120]
GTE 7B	L21	Content-token mean	-2.994	[-6.013, -0.305]
GTE 7B	L23	Valid-token mean	+7.946	[+4.815, +12.821]
GTE 7B	L23	Content-token mean	-0.166	[-3.737, +3.093]

B.4 SMALL ENCODER COMPARISONS

Qwen3-Embedding-0.6B runs at full depth with a separate map for each target index. These nine analytic-scale quality comparisons have unadjusted pointwise intervals; they use no test-selected recipe and have no matched cross-model timing.

Table 14: Calibrated early-exit Attention^2 versus full-depth Qwen 0.6B calibrated to the same index. Differences are attention minus the small encoder in nDCG@10 retention percentage points (pp). Positive values favor attention; intervals are unadjusted.

Target index model	Exit layer	Retention difference (pp)	Pointwise interval (pp; unadjusted)
Qwen 4B	L22	-5.330	[-7.801, -2.913]
Qwen 4B	L24	+3.133	[+0.949, +5.226]
Qwen 4B	L25	+5.270	[+3.076, +7.411]
Mistral 7B	L18	+7.997	[+5.468, +10.647]
Mistral 7B	L20	+13.378	[+11.007, +15.745]
Mistral 7B	L22	+15.431	[+13.119, +17.590]
GTE 7B	L16	-1.197	[-4.805, +1.942]
GTE 7B	L21	-1.158	[-4.763, +1.971]
GTE 7B	L23	+6.325	[+2.650, +9.700]

B.5 RAW AND CALIBRATED NEIGHBORING-METHOD CONTROLS

The Quora study fixes Qwen L27 and its full-depth index. Ditto varies three blocks, four heads and last/first–last pooling; Token Prepending (TP) varies three update endpoints and last/content pooling; Hierarchical Token Prepending (HTP) varies three endpoints and one/two sentence groups. Raw, unit-scale and analytic-scale variants yield 108 candidates. Each family is selected on 2,000 development queries and evaluated on 8,000 test queries; Attention^2 with analytic scaling is fixed. These are common-exit adaptations, not reproductions of the original benchmarks. The table gives development nDCG; Table 33 reports test comparisons with a three-family adjustment.

Method	Candidates	No calibration	Calibration unit scale	Calibration analytic scale
DITTO	72	0.026	0.433	0.588
TP	18	0.510	0.839	0.840
HTP	18	0.561	0.850	0.849

The selected configurations are Ditto (block 14, head 0, first–last, $\hat{\alpha}$), TP (endpoint 8, content mean, $\hat{\alpha}$), and HTP (one group, endpoint 27, $\alpha = 1$). Test results are in Table 33.

Table 15: Qwen 4B L27 on Quora: calibration restores compatibility with the fixed full-depth index. All entries are absolute scores.

	No calibration			Analytic-scale calibration		
Readout	nDCG@10	Recall at 100	Recall at 1000	nDCG@10	Recall at 100	Recall at 1000
Native	0.421986	0.825475	0.959126	0.825768	0.993458	0.999875
Attention^2	0.000000	0.001358	0.010134	0.850112	0.995878	0.999886
Content-token mean	0.491112	0.883239	0.979819	0.835081	0.990773	0.999519

Intermediate exits on Mistral. Table 16 extends the comparison to L18/L20. All readouts receive their own map from the same 45,000 anchors and use $\hat{\alpha}$; each external family is selected by mean development nDCG over SciFact/SciDocs (60/200 queries), then evaluated on 240/800 test queries. This separate retrospective A40 run uses frozen A800 full-depth indexes. After cross-device exact replay failed, the rerun used documented deviation monitoring; it is a descriptive comparison, not a qualified replication.

HTP has the highest point estimates in this Mistral panel. Its gains over Attention^2 on SciDocs have adjusted intervals $[0.0109, 0.0315]$ at L18 and $[0.0021, 0.0245]$ at L20; both SciFact intervals include

Table 16: Mistral 7B: calibrated query readouts against the same full-depth index. Cells are test nDCG@10. TP is Token Prepending; HTP is Hierarchical Token Prepending. This retrospective panel is descriptive (protocol below).

	L18		L20	
Readout	SciFact	SciDocs	SciFact	SciDocs
Native	0.2859	0.0538	0.5748	0.1020
Attention^2	0.5669	0.1162	0.6719	0.1467
Content-token mean	0.5945	0.1313	0.6667	0.1560
Valid-token mean	0.3370	0.0348	0.4808	0.0714
Ditto	0.3204	0.0381	0.5189	0.0913
TP	0.6079	0.1326	0.6626	0.1552
HTP	0.6097	0.1376	0.6736	0.1598

zero. These retrospective paired bootstrap intervals use 20,000 resamples and Bonferroni adjustment over three families, two exits and two datasets.

Readout cost. Attention^2 and content mean preserve the prefix input; attention additionally recomputes one pooling-position row. TP/HTP insert summary tokens and replace prefix states, with sentence grouping for HTP. This A40 panel has no matched TP/HTP timing; the A800 measurements below time the full attention-plus-calibration path.

B.6 Measured encoding-time savings

Table 17 reports all timed configurations; quality is in Tables 8–10.

Table 17: Complete query-encoding time on A800 at batch size one, including readout and calibration. Times are medians in milliseconds; savings and speedups use the same-length full-depth reference and unrounded measurements.

Model	Exit layer	Input length (tokens)	Full depth (ms)	Attention^2 + calibration (ms)	Time saved (%)	Speedup
Qwen	L22	32	74.88	47.02	37.2	1.59×
Qwen	L22	128	69.92	46.29	33.8	1.51×
Qwen	L24	32	74.88	51.26	31.5	1.46×
Qwen	L24	128	69.92	48.94	30.0	1.43×
Qwen	L25	32	74.88	52.07	30.5	1.44×
Qwen	L25	128	69.92	51.15	26.8	1.37×
Mistral	L18	32	56.30	33.71	40.1	1.67×
Mistral	L18	128	55.25	33.12	40.0	1.67×
Mistral	L20	32	56.30	37.03	34.2	1.52×
Mistral	L20	128	55.25	38.70	29.9	1.43×
Mistral	L22	32	56.30	41.53	26.2	1.36×
Mistral	L22	128	55.25	41.75	24.4	1.32×
Gte	L16	32	43.33	27.19	37.2	1.59×
Gte	L16	128	42.59	26.34	38.2	1.62×
Gte	L21	32	43.33	35.86	17.3	1.21×
Gte	L21	128	42.59	33.22	22.0	1.28×
Gte	L23	32	43.33	38.01	12.3	1.14×
Gte	L23	128	42.59	37.01	13.1	1.15×

Measurements use an A800-SXM4-40GB at batch size one, BF16 execution, FP32 calibration, 20 warmups and 100 synchronized rounds, rotating methods within each input length. Timing includes the forward pass, readout, source normalization, map and output normalization, but excludes tokenization, content-mask preparation, transfer and search. Lengths 32 and 128 were timed in separate blocks; their unresolved absolute-time inversion precludes a length-scaling conclusion. Models and maps remain resident. These results measure query-encoding savings, not reduced memory use, document ingestion or end-to-end search.

C BOTH-SIDED SHALLOW RETRIEVAL AND READOUT INTERPRETATION

C.1 BILATERAL RETRIEVAL AND DOCUMENT CONTROLS

Both queries and documents use the chosen exit, readout and calibration. A map is shared between the two sides of a configuration and fitted separately for each model, exit and readout. The shallow document index is built once and searched by inner product. Calibration uses 45,000 query-like document anchors and 5,000 held-out rows; their full-depth targets remain an offline cost.

The three corpora are SciFact, NFCorpus and FiQA, with 240, 258 and 518 test queries. Shallow documents use the query-like template; Qwen's maximum length is 192. To separate this input change from the document representation, the input-matched control keeps each query fixed and encodes full-depth native documents with the shallow branch's prompt and truncation. The original deep index retains its document inputs. Anchor texts are drawn from these corpora rather than a verified text-disjoint source.

Full-depth reference scores. Retention averages the per-dataset ratios using the denominators below. These differ from the separate seven-dataset study.

Table 18: Native full-depth nDCG@10 references for the three-dataset bilateral study. These denominators differ from the seven-dataset index-reuse study.

Model	SciFact	NFCorpus	FiQA
Qwen 4B	0.768567	0.400679	0.559520
Qwen 0.6B	0.682644	0.359179	0.438068
Mistral	0.707534	0.339126	0.507221
GTE 7B	0.787721	0.394943	0.602645

Bilateral intervals are pointwise 95%, using the fixed-denominator bootstrap in Appendix A.8. Complete depth grids are in Appendix E.

C.2 CALIBRATION SCALE UNDER BILATERAL RETRIEVAL

Holding the readout and map fixed separates the effect of analytic scaling from the recovery due to calibration. Table 19 compares unit and analytic scales; uncalibrated baselines and complete analytic-scale depth results are in Appendix E.

Table 19: Bilateral scale ablation on three datasets. Cells are mean per-dataset nDCG@10 retention ratios (1 = 100%). Each readout has its own map, held fixed across its two scale columns.

Model	Exit	Native unit scale	Native analytic scale	Attention unit scale	Attention analytic scale
Qwen 4B	12	0.1616	0.1672	0.2205	0.2879
Qwen 4B	16	0.2255	0.2260	0.4997	0.5036
Qwen 4B	20	0.2455	0.2487	0.6443	0.6524
Qwen 4B	22	0.3265	0.3278	0.8026	0.8030
Qwen 4B	24	0.7543	0.7597	0.8530	0.8521
Qwen 4B	25	0.8370	0.8381	0.8749	0.8753
Qwen 4B	27	0.8772	0.8815	0.8968	0.8996
Qwen 4B	28	0.8757	0.8792	0.9068	0.9095
Qwen 4B	32	0.9296	0.9303	0.9380	0.9395
Qwen 4B	36	0.9503	0.9503	0.9516	0.9517
Qwen 0.6B	12	0.1590	0.1619	0.3017	0.3405
Qwen 0.6B	16	0.1802	0.1803	0.5106	0.5152
Qwen 0.6B	20	0.5191	0.5197	0.8404	0.8461
Qwen 0.6B	22	0.8038	0.8030	0.8859	0.8849
Qwen 0.6B	24	0.8255	0.8259	0.8645	0.8713
Qwen 0.6B	28	0.9564	0.9564	0.9552	0.9554
Mistral 7B	12	0.2941	0.2941	0.5763	0.6069
Mistral 7B	16	0.6530	0.6576	0.8067	0.8371

Table 19 (continued)

Model	Exit	Native unit scale	Native analytic scale	Attention unit scale	Attention analytic scale
Mistral 7B	18	0.6750	0.6804	0.9025	0.9038
Mistral 7B	20	0.9038	0.9056	0.9591	0.9652
Mistral 7B	22	0.9846	0.9828	0.9530	0.9580
Mistral 7B	24	0.9893	0.9893	0.9682	0.9802
Mistral 7B	28	1.0025	1.0019	0.9936	0.9941
Mistral 7B	32	1.0084	1.0084	0.9791	0.9793
GTE 7B	10	0.2013	0.2029	0.6845	0.6876
GTE 7B	12	0.1843	0.1822	0.6952	0.6911
GTE 7B	16	0.3042	0.3022	0.7922	0.7929
GTE 7B	22	0.7761	0.7756	0.8576	0.8629
GTE 7B	28	0.9653	0.9645	0.9646	0.9649

C.3 Readout gains and document-side effects

The two contrasts in Table 20 distinguish changing the calibrated readout from changing the document representation. Both benefits vary with model and depth; input-matched document controls follow below.

Table 20: Three-dataset paired changes in nDCG@10 retention points, with pointwise 95% intervals. The readout contrast uses bilateral analytic-scale calibration. The document contrast fixes the calibrated attention query and map; positive values favor the original full-depth index. Absolute scores are in Appendix E.

Model	Exit	Attention − native gain (pp) [95% CI]	Deep − shallow documents difference (pp) [95% CI]
Qwen 4B	12	+12.07 [+9.69, +14.53]	+12.02 [+8.74, +15.34]
Qwen 4B	16	+27.76 [+24.55, +31.07]	+10.53 [+7.12, +13.82]
Qwen 4B	20	+40.37 [+36.79, +43.98]	+11.28 [+8.23, +14.36]
Qwen 4B	22	+47.52 [+43.80, +51.39]	+4.90 [+2.66, +7.10]
Qwen 4B	24	+9.24 [+7.08, +11.46]	+7.08 [+5.10, +8.93]
Qwen 4B	25	+3.72 [+1.90, +5.53]	+7.52 [+5.65, +9.31]
Qwen 4B	27	+1.81 [+0.22, +3.39]	+5.54 [+4.00, +7.09]
Qwen 4B	28	+3.03 [+1.38, +4.66]	+5.51 [+3.97, +6.99]
Qwen 4B	32	+0.92 [-0.07, +1.89]	+4.96 [+3.56, +6.29]
Qwen 4B	36	+0.14 [-0.30, +0.55]	+4.93 [+3.59, +6.19]
Qwen 0.6B	12	+17.86 [+15.08, +20.68]	+11.35 [+7.84, +14.89]
Qwen 0.6B	16	+33.49 [+30.01, +37.17]	+12.14 [+8.96, +15.24]
Qwen 0.6B	20	+32.64 [+29.15, +36.42]	+4.17 [+1.72, +6.57]
Qwen 0.6B	22	+8.19 [+5.82, +10.64]	+4.41 [+2.54, +6.28]
Qwen 0.6B	24	+4.54 [+2.86, +6.32]	+5.53 [+3.49, +7.48]
Qwen 0.6B	28	-0.09 [-0.90, +0.68]	+5.07 [+3.50, +6.68]
Mistral 7B	12	+31.28 [+27.70, +34.84]	-4.29 [-8.37, -0.21]
Mistral 7B	16	+17.95 [+15.10, +20.81]	-10.41 [-13.95, -6.70]
Mistral 7B	18	+22.34 [+19.27, +25.56]	-1.75 [-4.64, +1.25]
Mistral 7B	20	+5.95 [+3.60, +8.33]	-2.49 [-5.43, +0.46]
Mistral 7B	22	-2.48 [-4.52, -0.47]	-3.35 [-6.13, -0.48]
Mistral 7B	24	-0.91 [-2.93, +1.11]	-1.08 [-3.80, +1.67]
Mistral 7B	28	-0.77 [-2.45, +0.89]	-1.71 [-4.33, +0.86]
Mistral 7B	32	-2.91 [-4.29, -1.57]	-0.75 [-3.44, +1.84]
GTE 7B	10	+48.48 [+44.93, +52.02]	+12.82 [+10.53, +15.18]
GTE 7B	12	+50.89 [+47.31, +54.55]	+12.27 [+9.90, +14.71]
GTE 7B	16	+49.06 [+45.29, +52.64]	+9.76 [+7.65, +11.87]
GTE 7B	22	+8.74 [+6.76, +10.78]	+5.86 [+3.99, +7.78]
GTE 7B	28	+0.04 [-0.68, +0.75]	+3.20 [+1.89, +4.51]

C.4 INPUT-MATCHED DEEP DOCUMENTS AND EXACT DIFFERENCE ACCOUNTING

With the query readout, map and scale fixed, let Q_{shallow} denote retention with shallow documents, Q_{deep} retention with the original full-depth index, and Q_{matched} retention with input-matched full-depth documents. Then

$$Q_{\text{deep}} - Q_{\text{shallow}} = (Q_{\text{deep}} - Q_{\text{matched}}) + (Q_{\text{matched}} - Q_{\text{shallow}}). \tag{C.1}$$

The first difference changes document inputs; the second captures the remaining document-representation difference, including depth, readout and calibration. This is an accounting decomposition rather than a causal attribution.

Table 21: Document-input controls with the query, map and analytic scale fixed. The first three score columns are nDCG@10 retention ratios; the last two are differences in percentage points. Input change is original minus input-matched deep; representation change is input-matched deep minus shallow, with a pointwise 95% interval.

Model	Exit layer	Shallow documents (ratio)	Original deep index (ratio)	Input-matched deep index (ratio)	Input change (pp)	Representation change (pp) [95% CI]
Qwen 4B	12	0.2879	0.4081	0.3656	+4.25	+7.77 [+4.67, +10.86]
Qwen 4B	20	0.6524	0.7652	0.7205	+4.47	+6.81 [+4.10, +9.38]
Qwen 4B	22	0.8030	0.8520	0.8003	+5.17	-0.27 [-2.10, +1.66]
Qwen 4B	24	0.8521	0.9230	0.8648	+5.82	+1.27 [-0.36, +2.89]
Qwen 4B	28	0.9095	0.9647	0.9129	+5.17	+0.34 [-0.91, +1.53]
Qwen 0.6B	16	0.5152	0.6366	0.5950	+4.15	+7.98 [+4.90, +11.04]
Qwen 0.6B	20	0.8461	0.8878	0.8315	+5.63	-1.47 [-3.53, +0.53]
Qwen 0.6B	22	0.8849	0.9290	0.8711	+5.79	-1.38 [-3.06, +0.26]
Mistral	16	0.8371	0.7331	0.8132	-8.01	-2.40 [-5.14, +0.48]
Mistral	18	0.9038	0.8863	0.9143	-2.80	+1.05 [-1.24, +3.44]
Mistral	20	0.9652	0.9403	0.9680	-2.77	+0.28 [-1.88, +2.39]
Mistral	22	0.9580	0.9245	0.9647	-4.02	+0.67 [-1.03, +2.38]

Matching document inputs reverses the Mistral L18/L20 point ordering relative to bilateral retrieval; the original deep-index differences have intervals spanning zero.

Anchor source at a fixed budget. For Qwen L22, we compare 5,000 mixed-corpus anchors with 5,000 anchors excluding the evaluated corpus, using three fixed subsets. Both queries and documents exit early. Table 22 reports absolute three-domain macro nDCG under separately replayed local references, not a replacement for the historical 45,000-anchor depth tables.

Table 22: Qwen 4B, layer 22: anchor-source control with queries and documents exiting together. Cells are absolute macro nDCG@10 across three fixed subsets. The difference is attention minus native pooling. The two fitted rows use separate maps for each readout; exclusion also uses a separate map per evaluated corpus.

Anchor source	Fitting anchors	Native pooling nDCG@10	Attention readout nDCG@10	Attention − native [95% CI]
No calibration	0	0.0104	0.0798	+0.0695 [0.0561, 0.0825]
Mixed corpora	5,000	0.0958	0.4403	+0.3445 [0.3244, 0.3647]
Evaluated corpus excluded	5,000	0.0450	0.4030	+0.3580 [0.3373, 0.3789]

Excluding the evaluated corpus lowers Attention^2 macro nDCG by 0.0373 (paired interval $[0.0291, 0.0455]$), chiefly on FiQA (0.0986), while preserving its gain over calibrated native pooling. Unadjusted intervals use 5,000 paired within-dataset resamples with maps fixed; they exclude anchor-population uncertainty. The historical ridge floor affects readouts differently, so this retrospective comparison measures complete pipelines.

C.5 RECALL ALONGSIDE NDCG

Recall complements the ranking-focused nDCG comparisons by measuring relevant-document coverage at larger cutoffs. Table 23 retains all measured Qwen configurations; the corresponding Mistral and GTE measurements are unavailable.

Table 23: Absolute bilateral Recall@100 and Recall@1000 on three datasets. Unit and analytic scales share a map within each readout. The corresponding nDCG results are in the depth tables.

Model	Exit	Readout and calibration	Recall@100	Recall@1000
Qwen 4B	22	Native pooling, no calibration	0.0510	0.2576
Qwen 4B	22	Native pooling, unit scale	0.4792	0.7779
Qwen 4B	22	Native pooling, analytic scale	0.4730	0.7795
Qwen 4B	22	Attention^2, no calibration	0.1908	0.4555
Qwen 4B	22	Attention^2, unit scale	0.6586	0.8613
Qwen 4B	22	Attention^2, analytic scale	0.6649	0.8618
Qwen 4B	24	Native pooling, no calibration	0.2053	0.5480
Qwen 4B	24	Native pooling, unit scale	0.6547	0.8634
Qwen 4B	24	Native pooling, analytic scale	0.6565	0.8660
Qwen 4B	24	Attention^2, no calibration	0.3893	0.6396
Qwen 4B	24	Attention^2, unit scale	0.6882	0.8773
Qwen 4B	24	Attention^2, analytic scale	0.6893	0.8788
Qwen 4B	28	Native pooling, no calibration	0.6193	0.8369
Qwen 4B	28	Native pooling, unit scale	0.7011	0.8835
Qwen 4B	28	Native pooling, analytic scale	0.7019	0.8845
Qwen 4B	28	Attention^2, no calibration	0.5642	0.7914
Qwen 4B	28	Attention^2, unit scale	0.7082	0.8819
Qwen 4B	28	Attention^2, analytic scale	0.7091	0.8829
Qwen 0.6B	20	Native pooling, no calibration	0.2251	0.5500
Qwen 0.6B	20	Native pooling, unit scale	0.5087	0.7777
Qwen 0.6B	20	Native pooling, analytic scale	0.5097	0.7778
Qwen 0.6B	20	Attention^2, no calibration	0.3538	0.6169
Qwen 0.6B	20	Attention^2, unit scale	0.6188	0.8337
Qwen 0.6B	20	Attention^2, analytic scale	0.6207	0.8360
Qwen 0.6B	22	Native pooling, no calibration	0.5385	0.7832
Qwen 0.6B	22	Native pooling, unit scale	0.6231	0.8403
Qwen 0.6B	22	Native pooling, analytic scale	0.6221	0.8400
Qwen 0.6B	22	Attention^2, no calibration	0.5189	0.7661
Qwen 0.6B	22	Attention^2, unit scale	0.6326	0.8475
Qwen 0.6B	22	Attention^2, analytic scale	0.6335	0.8452

C.6 WHY A SHARED CALIBRATION CAN CHANGE BILATERAL SIMILARITY

For a fixed source readout and layer, set $y_x = W^\top(u(x) - \mu_x)$, $m = \mu_z$ and $v_x = \alpha y_x + m$, with all v_x nonzero. The exact bilateral score is

$$\frac{v_q^\top v_d}{\|v_q\|\|v_d\|} = \frac{\alpha^2 y_q^\top y_d + \alpha m^\top (y_q + y_d) + \|m\|^2}{\|\alpha y_q + m\|\|\alpha y_d + m\|}. \tag{C.2}$$

The centered term is $y_q^\top y_d = (u_q - \mu_x)^\top W W^\top (u_d - \mu_x)$, where $WW^\top$ is positive semidefinite. A shared ridge map can therefore change how source directions contribute to similarity. Mean restoration and normalization also affect scores. An orthogonal map without shifts preserves cosine similarity; ridge need not. Retrieval gains are measured empirically and do not isolate an individual term.

Proposition 2 and Theorem 1 assume a fixed document index. In bilateral retrieval, document vectors and their norms also change with calibration. For normalized shallow vectors $z + e, d_j + f_j$, direct expansion gives

$$\widehat{s}_j - s_j = e^\top d_j + z^\top f_j + e^\top f_j. \tag{15}$$

Thus for $\gamma_{ij} = s_i - s_j$ and $\widehat{\gamma}_{ij} = \widehat{s}_i - \widehat{s}_j$,

$$\widehat{\gamma}_{ij} - \gamma_{ij} = e^\top (d_i - d_j) + z^\top (f_i - f_j) + e^\top (f_i - f_j). \tag{16}$$

Equation 16 separates the margin change into query error, document error and their interaction. The latter two vanish when the index is fixed. Query-side reconstruction error alone therefore cannot characterize bilateral retrieval.

C.7 Attention-squared means a second aggregation

Attention^2 reuses attention to aggregate the block outputs a second time; it does not generally square the attention matrix.

Let A_h be an attention matrix for head h, $\bar{A} = H^{-1} \sum_h A_h$, and H_{out} the block-output matrix. Before final vector normalization, Attention^2 is exactly

$$u_{\text{pre}}^\top = e_p^\top \bar{A} H_{\text{out}}. \tag{C.3}$$

Only the row $e_p^\top \bar{A}$ is needed for the readout. For a pre-normalized residual block with T tokens, let $V_h \in \mathbb{R}^{T \times d_h}$ contain head h's token value vectors and $O_h \in \mathbb{R}^{d_h \times d_{\text{model}}}$ be its output projection. Then

$$H_{\text{out}} = H_{\text{in}} + \sum_h A_h V_h O_h + G,$$

where G is the actual post-attention feed-forward/residual contribution. Then

$$u_{\text{pre}}^\top = e_p^\top \bar{A} H_{\text{in}} + \sum_h e_p^\top \bar{A} A_h V_h O_h + e_p^\top \bar{A} G. \tag{C.4}$$

Only a single-head block without residual or feed-forward paths reduces to $e_p^\top A^2 V O$, where V contains token values and O is the output projection. This special case motivates the name; the full-block expression above defines the general readout. It is distinct from the trained gathering/distribution operator of A^2-Nets (Chen et al., 2018).

C.8 Empirical connection to score-level theory

The Quora analysis compares $\hat{\alpha}$ with $\alpha = 1$ using the same map and 8,000 queries. The plot shows changes in centered score error, normalized by each query's native score energy, against changes in nDCG. At L22, 928 of the 1,098 improved queries also have higher cosine and centered score errors (counting changes above 10^{-6}).

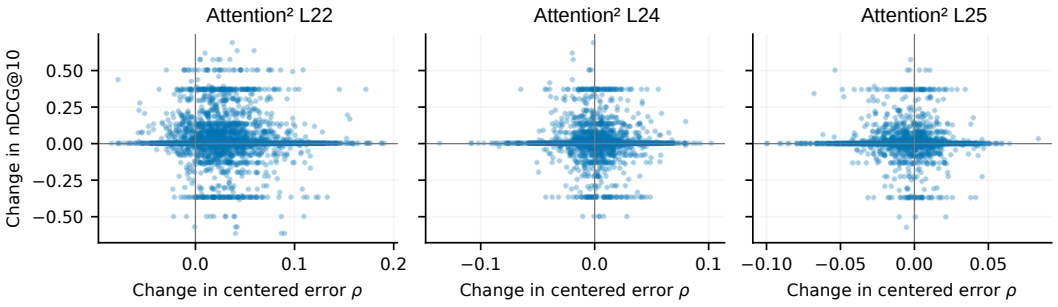


Figure 3: Same-query retrieval and distortion changes, $\hat{\alpha}$ minus one. Upper-right points improve nDCG despite larger centered distortion. All 8,000 queries per panel are drawn; many have unchanged retrieval. This is an empirical mismatch between the surrogate and retrieval, not a frequency estimate of the arbitrarily small-error counterexample.

This example shows why average reconstruction error is insufficient to predict retrieval. Applying Theorem 1 would additionally require margins and error projections for the same document pairs.

D ADDITIONAL DEPTH, SCALE AND TRANSFER ANALYSES

D.1 SCALE, RESTORED MEANS AND RETRIEVAL GEOMETRY

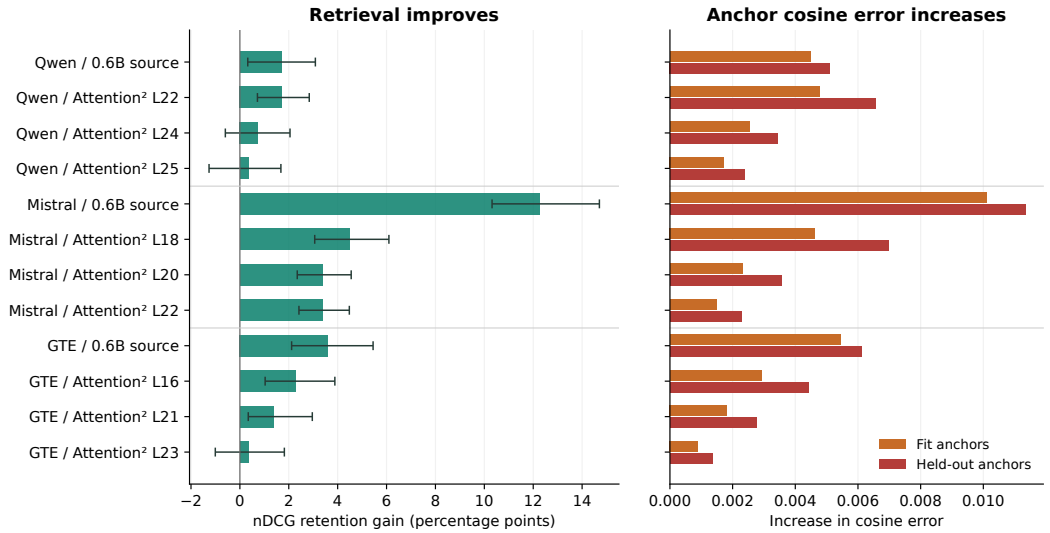


Figure 4: **Analytic scaling improves retrieval while increasing anchor cosine error.** Each row fixes the map and both means and changes only $\alpha = 1$ to $\hat{\alpha}$. Left: all twelve retrieval point gains are positive; nine intervals exclude zero after the original twelve-comparison adjustment. Right: fitting and held-out cosine errors increase in every case. The 0.6B rows align a small query encoder to the named target index. No fitted head or new optimization is introduced.

Table 24: **Fixed-map scale controls.** Qwen on the separate 8,000-query Quora study; nDCG retention (%). Matrix and source/target means are fixed. The last column omits target-mean restoration consistently from the centered transform.

		nDCG@10 retention (%)			
Readout and exit	Estimated scale $\hat{\alpha}$	Unit scale $\alpha = 1$	Analytic scale $\alpha = \hat{\alpha}$	Fixed scale $\alpha = 1.5$	Without target mean restoration
Attention^2 L22	1.380	89.3	90.4	90.5	87.1
Attention^2 L24	1.270	95.4	95.7	95.6	92.8
Attention^2 L25	1.219	96.4	96.7	96.5	94.1
Attention^2 L27	1.191	96.1	96.5	96.5	94.3
Native L24	1.386	81.8	84.5	84.8	80.5
Native L27	1.243	93.3	93.7	93.7	90.3
Content mean L27	1.234	94.7	94.8	94.6	91.8
Qwen 0.6B L28	1.290	92.8	94.3	94.8	92.7

All eight configurations favor mean restoration. Fixed scale 1.5 is better for native L24 and Qwen 0.6B alignment; the other six comparisons are unresolved after adjustment. Thus $\hat{\alpha}$ is an unlabeled rule, not an estimated retrieval optimum. The 0.6B row maps its full-depth vector into the Qwen-4B index.

For $q_\alpha = \mathrm{Normalize}(\alpha W^\top (x - \mu_x) + \mu_z)$, removing target-mean restoration means using $\mathrm{Normalize}(\alpha W^\top (x - \mu_x))$ consistently. A positive scalar then cancels.

Paired query-level distortion/quality diagnostics are in Appendix C.8; the additional document-error and interaction terms for bilateral retrieval are derived in Appendix C.6.

E COMPLETE RESULTS ACROSS LAYERS AND RETRIEVAL SETTINGS

The tables report every measured layer, including reversals absent from representative main-text rows. Cells give macro nDCG@10 (mean per-dataset retention, %). The bilateral tables compare uncalibrated readouts and calibrated combinations; fixed-index and combined-system results follow.

Token and bilateral results share the three-dataset full-depth references, but were obtained in separate evaluations; full-curve comparisons are descriptive. The token study used different stored precision, encoding batches and scoring implementation. Paired shallow comparisons are in Appendix E.8. The decoder tables include supplementary single-vector measurements below L8. On NFCorpus, native calibration at Qwen 0.6B L4–L6 and attention calibration at Mistral L2/L3 underperform their uncalibrated readouts; these diagnostic values are excluded from row maxima. The seven-dataset ladders use their own maps and references and remain separate from the later pooling comparison in Appendix B.

E.1 BOTH QUERIES AND DOCUMENTS SHALLOW: ALL MEASURED DEPTHS

Table 25: Qwen 4B: all layers with queries and documents exiting together, on three datasets. Cells give macro nDCG@10 (retention, %). Bold marks the highest unrounded retention per row. Calibration uses a separate map for each readout.

Exit layer	Native pooling no calibration	MaxSim	Attention^2 no calibration	Attention^2 + calibration	Native pooling + calibration
1	0.009 (1.8)	0.310 (**51.4**)	0.013 (2.4)	0.203 (33.4)	0.071 (11.2)
2	0.008 (1.7)	0.335 (**56.2**)	0.005 (1.1)	0.139 (23.6)	0.068 (10.9)
3	0.005 (1.3)	0.335 (**56.1**)	0.006 (1.5)	0.187 (31.2)	0.062 (9.9)
4	0.005 (1.2)	0.347 (**58.2**)	0.006 (1.5)	0.227 (38.9)	0.086 (14.0)
5	0.005 (1.1)	0.364 (**61.4**)	0.009 (2.0)	0.197 (34.4)	0.082 (13.0)
6	0.006 (1.4)	0.372 (**62.9**)	0.005 (1.2)	0.217 (37.5)	0.066 (10.8)
7	0.006 (1.5)	0.384 (**64.6**)	0.007 (1.7)	0.205 (36.6)	0.094 (15.7)
8	0.006 (1.4)	0.389 (**65.4**)	0.004 (0.9)	0.041 (6.9)	0.070 (11.8)
9	0.006 (1.4)	0.401 (**67.7**)	0.005 (1.0)	0.096 (15.6)	0.066 (10.9)
10	0.006 (1.4)	0.417 (**70.8**)	0.004 (1.0)	0.078 (12.9)	0.074 (12.6)
11	0.006 (1.4)	0.422 (**71.9**)	0.003 (0.8)	0.137 (22.7)	0.089 (14.7)
12	0.006 (1.6)	0.422 (**71.9**)	0.006 (1.4)	0.166 (28.8)	0.097 (16.7)
13	0.007 (1.7)	0.419 (**71.4**)	0.007 (1.6)	0.219 (37.5)	0.132 (22.0)
14	0.006 (1.6)	0.401 (**68.0**)	0.005 (1.2)	0.174 (29.7)	0.124 (20.6)
15	0.007 (1.6)	0.395 (**66.5**)	0.006 (1.5)	0.306 (51.9)	0.156 (25.9)
16	0.007 (1.6)	0.374 (**62.9**)	0.006 (1.4)	0.298 (50.4)	0.138 (22.6)
17	0.007 (1.6)	0.371 (**62.1**)	0.008 (1.6)	0.331 (57.0)	0.140 (23.4)
18	0.007 (1.7)	0.362 (**60.6**)	0.014 (2.8)	0.336 (57.7)	0.157 (26.0)
19	0.008 (1.9)	0.356 (**59.5**)	0.014 (2.7)	0.310 (52.8)	0.158 (26.3)
20	0.008 (1.9)	0.336 (56.2)	0.004 (1.0)	0.384 (**65.2**)	0.147 (24.9)
21	0.011 (2.2)	0.340 (56.8)	0.047 (7.6)	0.426 (**73.1**)	0.179 (29.5)
22	0.010 (2.2)	0.344 (57.5)	0.066 (11.2)	0.464 (**80.3**)	0.196 (32.8)
23	0.017 (3.4)	0.360 (60.6)	0.035 (6.2)	0.441 (**75.8**)	0.235 (38.9)
24	0.050 (8.3)	0.391 (65.9)	0.168 (27.8)	0.496 (**85.2**)	0.449 (76.0)
25	0.228 (37.4)	0.432 (73.6)	0.169 (28.3)	0.506 (**87.5**)	0.489 (83.8)
26	0.247 (40.1)	0.433 (73.4)	0.236 (37.7)	0.514 (**89.0**)	0.497 (85.2)
27	0.350 (57.5)	0.451 (76.9)	0.214 (32.6)	0.520 (**90.0**)	0.512 (88.2)
28	0.403 (66.8)	0.458 (78.3)	0.348 (56.8)	0.525 (**91.0**)	0.511 (87.9)
29	0.407 (67.6)	0.457 (77.9)	0.387 (64.3)	0.528 (**91.5**)	0.513 (88.0)
30	0.486 (82.8)	0.469 (80.6)	0.364 (62.1)	0.539 (**93.3**)	0.535 (92.5)
31	0.488 (83.1)	0.472 (81.1)	0.386 (65.0)	0.538 (**93.0**)	0.537 (92.8)
32	0.498 (85.2)	0.479 (82.3)	0.405 (68.8)	0.542 (**94.0**)	0.536 (93.0)
33	0.505 (86.6)	0.483 (83.1)	0.320 (52.4)	0.538 (93.3)	0.540 (**93.6**)
34	0.505 (86.2)	0.481 (82.7)	0.316 (51.9)	0.541 (**93.6**)	0.541 (93.5)
35	0.513 (87.9)	0.481 (82.7)	0.524 (90.5)	0.551 (**95.4**)	0.544 (94.1)
36	0.549 (94.9)	0.499 (86.2)	0.548 (94.9)	0.550 (**95.2**)	0.549 (95.0)

The same fixed full-depth references are used across methods. Boldface does not imply statistical significance.

Table 26: Mistral 7B: all layers with queries and documents exiting together, on three datasets. Cells give macro nDCG@10 (retention, %). Bold marks the highest unrounded retention per row. Calibration uses a separate map for each readout.

Exit layer	Native pooling no calibration	MaxSim	Attention^2 no calibration	Attention^2 + calibration	Native pooling + calibration
1	0.011 (2.4)	0.338 (**65.1**)	0.006 (1.8)	0.165 (26.8)	0.181 (31.3)
2	0.009 (2.0)	0.363 (**69.4**)	0.005 (1.5)	0.005 (1.5)	0.136 (23.4)
3	0.005 (1.4)	0.367 (**70.4**)	0.005 (1.5)	0.006 (1.6)	0.145 (25.3)
4	0.007 (1.7)	0.377 (**72.1**)	0.005 (1.3)	0.028 (5.5)	0.110 (20.4)
5	0.006 (1.5)	0.384 (**73.5**)	0.005 (1.3)	0.059 (11.9)	0.109 (20.7)
6	0.008 (2.0)	0.397 (**76.1**)	0.003 (0.9)	0.041 (8.7)	0.133 (26.2)
7	0.007 (1.9)	0.387 (**74.5**)	0.004 (1.3)	0.113 (23.4)	0.140 (27.7)
8	0.006 (1.8)	0.384 (**74.0**)	0.003 (1.0)	0.231 (43.9)	0.160 (32.0)
9	0.007 (2.0)	0.369 (**70.9**)	0.007 (1.6)	0.284 (55.5)	0.158 (31.3)
10	0.007 (1.9)	0.349 (**66.8**)	0.002 (0.6)	0.233 (44.8)	0.152 (30.3)
11	0.006 (1.8)	0.312 (59.5)	0.003 (1.0)	0.307 (**60.2**)	0.154 (30.7)
12	0.005 (1.4)	0.301 (57.1)	0.005 (1.3)	0.309 (**60.7**)	0.147 (29.4)
13	0.005 (1.4)	0.277 (**52.5**)	0.004 (1.0)	0.211 (43.4)	0.146 (29.6)
14	0.006 (1.7)	0.262 (49.6)	0.007 (1.8)	0.349 (**69.6**)	0.152 (30.3)
15	0.007 (2.0)	0.270 (51.3)	0.008 (2.2)	0.310 (**60.3**)	0.152 (30.3)
16	0.013 (3.4)	0.300 (57.0)	0.052 (10.0)	0.434 (**83.7**)	0.336 (65.8)
17	0.017 (4.2)	0.329 (62.7)	0.040 (8.2)	0.365 (**69.7**)	0.345 (67.4)
18	0.018 (4.1)	0.345 (65.8)	0.104 (21.4)	0.466 (**90.4**)	0.348 (68.0)
19	0.050 (11.1)	0.390 (74.6)	0.078 (15.7)	0.462 (**89.5**)	0.419 (81.7)
20	0.196 (40.9)	0.412 (79.1)	0.353 (67.0)	0.497 (**96.5**)	0.465 (90.6)
21	0.336 (66.1)	0.422 (81.1)	0.281 (51.6)	0.500 (**96.8**)	0.488 (94.8)
22	0.435 (85.0)	0.432 (83.1)	0.294 (52.8)	0.494 (95.8)	0.505 (**98.3**)
23	0.454 (88.6)	0.435 (83.8)	0.085 (19.1)	0.487 (94.6)	0.506 (**98.5**)
24	0.460 (90.0)	0.433 (83.6)	0.183 (39.1)	0.503 (98.0)	0.507 (**98.9**)
25	0.478 (93.8)	0.437 (84.6)	0.148 (32.6)	0.502 (98.0)	0.507 (**99.2**)
26	0.486 (95.4)	0.434 (83.9)	0.128 (29.3)	0.495 (96.8)	0.512 (**99.8**)
27	0.494 (96.7)	0.434 (83.9)	0.177 (34.2)	0.475 (92.5)	0.515 (**100.5**)
28	0.496 (97.2)	0.429 (82.9)	0.219 (46.6)	0.510 (99.4)	0.514 (**100.2**)
29	0.497 (97.1)	0.421 (81.3)	0.349 (67.6)	0.499 (97.2)	0.511 (**99.7**)
30	0.508 (99.1)	0.414 (79.8)	0.363 (70.3)	0.508 (99.2)	0.517 (**100.9**)
31	0.512 (100.1)	0.406 (78.0)	0.329 (63.4)	0.496 (96.3)	0.518 (**101.1**)
32	0.517 (**100.9**)	0.399 (76.5)	0.394 (74.7)	0.503 (97.9)	0.517 (100.8)

The same fixed full-depth references are used across methods. Boldface does not imply statistical significance.

Table 27: Qwen 0.6B: all layers with queries and documents exiting together, on three datasets. Cells give macro nDCG@10 (retention, %). Bold marks the highest unrounded retention per row. Calibration uses a separate map for each readout.

Exit layer	Native pooling no calibration	MaxSim	Attention^2 no calibration	Attention^2 + calibration	Native pooling + calibration
1	0.006 (1.5)	0.311 (**59.0**)	0.008 (1.8)	0.160 (29.2)	0.056 (10.2)
2	0.005 (1.5)	0.323 (**61.6**)	0.007 (1.7)	0.132 (24.7)	0.040 (7.5)
3	0.006 (1.6)	0.335 (**64.1**)	0.005 (1.4)	0.101 (18.5)	0.024 (4.8)
4	0.006 (1.5)	0.347 (**66.8**)	0.003 (0.9)	0.012 (2.8)	0.017 (3.3)
5	0.006 (1.6)	0.350 (**67.6**)	0.003 (0.8)	0.007 (1.8)	0.016 (3.3)
6	0.006 (1.5)	0.365 (**70.8**)	0.004 (1.1)	0.015 (3.2)	0.018 (3.4)
7	0.006 (1.6)	0.371 (**72.1**)	0.003 (0.7)	0.013 (2.7)	0.011 (2.5)
8	0.005 (1.5)	0.382 (**74.4**)	0.003 (0.9)	0.073 (14.7)	0.030 (6.0)
9	0.006 (1.6)	0.377 (**73.5**)	0.003 (0.8)	0.023 (4.4)	0.024 (4.8)
10	0.005 (1.5)	0.375 (**73.2**)	0.007 (1.5)	0.136 (26.6)	0.049 (9.8)
11	0.006 (1.6)	0.380 (**74.2**)	0.005 (1.1)	0.094 (19.4)	0.047 (9.4)
12	0.006 (1.5)	0.377 (**73.4**)	0.005 (1.4)	0.175 (34.1)	0.085 (16.2)
13	0.005 (1.3)	0.378 (**73.4**)	0.005 (1.3)	0.225 (45.0)	0.079 (15.7)
14	0.005 (1.5)	0.374 (**72.3**)	0.004 (1.2)	0.143 (27.5)	0.064 (12.4)
15	0.006 (1.6)	0.364 (**70.5**)	0.014 (2.9)	0.201 (40.2)	0.073 (14.5)
16	0.007 (1.9)	0.351 (**67.9**)	0.020 (3.8)	0.259 (51.5)	0.091 (18.0)
17	0.009 (2.1)	0.370 (**71.9**)	0.050 (9.8)	0.303 (60.7)	0.100 (20.4)

Table 27 (continued)

Exit layer	Native pooling no calibration	MaxSim	Attention^2 no calibration	Attention^2 + calibration	Native pooling + calibration
18	0.051 (9.8)	0.382 (**74.4**)	0.132 (23.2)	0.339 (67.2)	0.187 (36.3)
19	0.059 (11.5)	0.390 (**76.1**)	0.079 (15.3)	0.373 (73.7)	0.204 (39.4)
20	0.065 (12.3)	0.401 (78.8)	0.174 (33.3)	0.424 (**84.6**)	0.269 (52.0)
21	0.187 (34.9)	0.424 (83.5)	0.291 (56.5)	0.426 (**85.0**)	0.355 (70.3)
22	0.311 (59.4)	0.429 (84.5)	0.300 (56.9)	0.445 (**88.5**)	0.405 (80.3)
23	0.339 (65.0)	0.431 (85.3)	0.391 (77.5)	0.452 (**90.6**)	0.419 (83.1)
24	0.332 (63.2)	0.428 (84.6)	0.284 (54.8)	0.436 (**87.1**)	0.416 (82.6)
25	0.372 (72.5)	0.433 (85.8)	0.175 (33.6)	0.446 (**89.2**)	0.440 (87.7)
26	0.420 (83.0)	0.444 (88.1)	0.211 (41.0)	0.456 (91.5)	0.459 (**91.5**)
27	0.450 (89.7)	0.443 (88.3)	0.281 (54.3)	0.469 (94.2)	0.473 (**94.7**)
28	0.477 (95.6)	0.467 (93.2)	0.462 (92.4)	0.477 (95.5)	0.477 (**95.6**)

The same fixed full-depth references are used across methods. Boldface does not imply statistical significance.

Table 28: GTE 7B: all five measured bilateral exits, on three datasets. Cells give macro nDCG@10 (mean per-dataset retention, %). Bold marks the highest unrounded retention per row. Each calibrated readout has its own map. MaxSim was not evaluated.

Exit	Native pooling no calibration	Attention no calibration	Attention + calibration	Native pooling + calibration
10	0.004 (1.0)	0.038 (6.4)	0.411 (**68.8**)	0.126 (20.3)
12	0.004 (1.0)	0.035 (6.5)	0.412 (**69.1**)	0.113 (18.2)
16	0.004 (0.8)	0.023 (4.1)	0.470 (**79.3**)	0.192 (30.2)
22	0.256 (40.4)	0.223 (36.7)	0.517 (**86.3**)	0.473 (77.6)
28	0.575 (**96.6**)	0.567 (95.4)	0.573 (96.5)	0.574 (96.4)

The same fixed full-depth references are used across methods. Boldface does not imply statistical significance.

Scope of the independent readout result. Across the 80 original bilateral configurations, uncalibrated Attention^2 exceeds native pooling at 28 exits: 8/29 Qwen 4B, 7/25 Mistral, 10/21 Qwen 0.6B and 3/5 GTE. It exceeds calibrated native pooling at none and leads at none of the 63 exits with all three independent interventions. These are descriptive layer counts. At Mistral L20, retention is 40.9% for native pooling, 67.0% for uncalibrated attention, 90.6% for calibrated native pooling and 96.5% for calibrated attention. Attention's largest additional gains therefore occur with readout-specific calibration; the winning ranges overlap.

Readout and geometry. Native calibration probes geometric recovery; calibrated attention tests the added benefit of changing the readout under the same fitting recipe. Token selection, attention weights and the readout-specific map jointly define that change (mean-pooling controls: Appendix B); the comparison neither isolates attention weighting nor rules out other maps. With both sides shallow, a shared non-orthogonal map also changes similarity, as derived in Appendix C.6.

E.2 Deep-index reuse: readouts with and without calibration

The fixed-index tables distinguish readout changes from geometric alignment. Native and Attention^2 are evaluated both without a map and with readout-specific analytic-scale calibration. All columns within a model use the same seven-dataset study and native full-depth index. MaxSim requires a token index and is evaluated in the preceding bilateral tables.

Table 29: Qwen 4B: all measured query exits against the native full-depth document index, on seven datasets. Cells give macro nDCG@10 (mean per-dataset retention, %). Each calibrated readout has its own map.

Exit	Native pooling no calibration	Attention no calibration	Attention + calibration	Native pooling + calibration
12	0.002 (0.5)	0.002 (0.6)	0.206 (37.5)	0.128 (23.2)
13	0.003 (0.7)	0.002 (0.6)	0.280 (50.9)	0.158 (27.1)
14	0.003 (0.6)	0.002 (0.6)	0.214 (38.7)	0.146 (25.4)
15	0.003 (0.6)	0.002 (0.5)	0.361 (65.2)	0.170 (29.6)

Table 29 (continued)

Exit	Native pooling no calibration	Attention no calibration	Attention + calibration	Native pooling + calibration
16	0.004 (0.8)	0.002 (0.5)	0.328 (58.0)	0.154 (26.8)
17	0.003 (0.8)	0.002 (0.5)	0.360 (64.4)	0.144 (25.1)
18	0.003 (0.7)	0.002 (0.6)	0.387 (68.6)	0.154 (26.3)
19	0.003 (0.7)	0.002 (0.6)	0.370 (66.0)	0.160 (27.7)
20	0.003 (0.6)	0.002 (0.5)	0.410 (73.3)	0.152 (26.2)
21	0.003 (0.8)	0.002 (0.6)	0.431 (76.7)	0.174 (29.5)
22	0.003 (0.8)	0.003 (0.6)	0.455 (80.8)	0.196 (33.4)
23	0.006 (1.3)	0.002 (0.6)	0.448 (80.0)	0.232 (39.4)
24	0.031 (5.9)	0.003 (0.7)	0.493 (89.0)	0.435 (77.2)
25	0.121 (22.1)	0.004 (0.8)	0.506 (91.6)	0.475 (85.2)
26	0.138 (25.4)	0.004 (0.8)	0.506 (91.8)	0.480 (86.0)
27	0.228 (41.1)	0.004 (0.8)	0.521 (94.7)	0.497 (90.1)
28	0.285 (51.8)	0.004 (0.9)	0.521 (95.0)	0.502 (91.2)
29	0.315 (57.2)	0.009 (1.9)	0.528 (96.6)	0.508 (92.5)
30	0.418 (77.1)	0.009 (2.0)	0.532 (97.5)	0.525 (96.0)
31	0.442 (81.4)	0.016 (3.2)	0.534 (97.9)	0.533 (97.4)
32	0.479 (88.3)	0.047 (8.9)	0.537 (98.2)	0.535 (97.7)
33	0.505 (93.1)	0.006 (1.2)	0.540 (98.7)	0.539 (98.1)
34	0.524 (96.2)	0.144 (26.1)	0.543 (99.3)	0.540 (98.7)
35	0.532 (97.4)	0.371 (64.6)	0.548 (100.2)	0.544 (99.6)
36	0.545 (99.8)	0.444 (77.5)	0.546 (99.9)	0.546 (99.9)

Table 30: Mistral 7B: all measured query exits against the native full-depth document index, on seven datasets. Cells give macro nDCG@10 (mean per-dataset retention, %). Each calibrated readout has its own map.

Exit	Native pooling no calibration	Attention no calibration	Attention + calibration	Native pooling + calibration
12	0.005 (1.2)	0.002 (0.5)	0.255 (51.8)	0.128 (25.4)
13	0.005 (1.1)	0.002 (0.5)	0.202 (41.5)	0.117 (23.5)
14	0.005 (1.2)	0.002 (0.4)	0.309 (60.9)	0.118 (23.1)
15	0.006 (1.2)	0.001 (0.4)	0.272 (53.4)	0.111 (21.3)
16	0.021 (5.1)	0.002 (0.5)	0.349 (67.7)	0.217 (42.7)
17	0.023 (5.3)	0.002 (0.5)	0.302 (59.2)	0.238 (46.2)
18	0.016 (3.8)	0.002 (0.5)	0.435 (83.7)	0.262 (50.0)
19	0.044 (9.7)	0.002 (0.7)	0.422 (81.6)	0.332 (63.5)
20	0.095 (20.6)	0.005 (1.3)	0.461 (89.2)	0.395 (75.8)
21	0.167 (36.0)	0.005 (1.2)	0.456 (89.4)	0.445 (85.6)
22	0.245 (52.1)	0.004 (1.1)	0.466 (91.2)	0.472 (91.5)
23	0.269 (56.6)	0.002 (0.6)	0.458 (89.9)	0.478 (93.3)
24	0.275 (58.1)	0.002 (0.6)	0.478 (93.3)	0.482 (94.0)
25	0.322 (66.4)	0.003 (1.0)	0.476 (93.7)	0.486 (94.9)
26	0.342 (69.6)	0.002 (0.7)	0.479 (93.7)	0.488 (95.3)
27	0.363 (73.5)	0.004 (1.1)	0.459 (90.0)	0.493 (96.7)
28	0.370 (75.0)	0.004 (1.0)	0.488 (95.4)	0.494 (96.6)
29	0.396 (79.9)	0.026 (5.8)	0.483 (94.7)	0.497 (97.4)
30	0.440 (88.3)	0.046 (9.8)	0.490 (96.3)	0.506 (99.5)
31	0.469 (93.5)	0.097 (19.7)	0.490 (95.7)	0.506 (99.4)
32	0.507 (99.9)	0.440 (88.3)	0.487 (96.0)	0.507 (99.9)

Table 31: Qwen 0.6B: all measured query exits against the native full-depth document index, on seven datasets. Cells give macro nDCG@10 (mean per-dataset retention, %). Each calibrated readout has its own map.

Exit	Native pooling no calibration	Attention no calibration	Attention + calibration	Native pooling + calibration
12	0.003 (0.6)	0.002 (0.5)	0.247 (44.4)	0.100 (16.6)
13	0.001 (0.4)	0.002 (0.5)	0.312 (56.6)	0.089 (15.8)
14	0.001 (0.4)	0.002 (0.5)	0.210 (36.0)	0.077 (13.0)
15	0.005 (0.9)	0.002 (0.6)	0.315 (57.8)	0.086 (15.0)

Table 31 (continued)

Exit	Native pooling no calibration	Attention no calibration	Attention + calibration	Native pooling + calibration
16	0.007 (1.2)	0.002 (0.6)	0.360 (65.6)	0.117 (20.3)
17	0.004 (0.9)	0.002 (0.6)	0.393 (72.4)	0.156 (27.7)
18	0.048 (7.6)	0.002 (0.6)	0.422 (77.7)	0.271 (48.4)
19	0.050 (7.9)	0.002 (0.6)	0.439 (81.3)	0.275 (49.7)
20	0.053 (8.2)	0.002 (0.6)	0.467 (86.2)	0.362 (64.9)
21	0.150 (26.3)	0.002 (0.6)	0.481 (89.5)	0.431 (79.7)
22	0.293 (52.7)	0.003 (0.7)	0.489 (91.7)	0.466 (86.3)
23	0.330 (60.1)	0.003 (0.7)	0.498 (93.9)	0.472 (87.9)
24	0.363 (65.9)	0.003 (0.8)	0.491 (91.9)	0.485 (90.3)
25	0.436 (80.4)	0.003 (0.9)	0.500 (94.0)	0.506 (94.8)
26	0.494 (92.0)	0.008 (1.7)	0.512 (97.0)	0.517 (97.7)
27	0.523 (98.9)	0.018 (3.7)	0.519 (98.7)	0.526 (99.9)
28	0.527 (100.1)	0.254 (41.1)	0.529 (100.9)	0.526 (99.9)

E.3 Detailed scale, neighboring-method and encoder results

Transfer to a larger index. Full-depth scores for dev MRR@10, DL19 nDCG@10 and DL20 nDCG@10 are 0.3590, 0.7564 and 0.7296 for Qwen 4B, and 0.3167, 0.6692 and 0.6754 for Qwen 0.6B, respectively. Table 32 compares the two readouts at each scale on the 8,841,823-passage index. These quality measurements have no matched large-index latency measurement; the A800 timings use the separate workload in Appendix B.6.

Table 32: **The recovery persists on an 8.84-million-passage index.** MS MARCO and TREC DL quality retained (%). Each pair compares native pooling and Attention^2 at the same scale recipe on the same full-depth index.

		Ridge, $\alpha = 1$		Ridge + analytic scale	
Model and exit	Evaluation	Native	Attention^2	Native	Attention^2
Qwen 4B L24	Dev MRR@10	46.6	80.6	52.0	83.1
	DL19 nDCG@10	69.1	87.2	74.1	89.1
	DL20 nDCG@10	57.3	86.7	60.4	88.0
Qwen 0.6B L20	Dev MRR@10	35.7	78.3	38.4	81.1
	DL19 nDCG@10	50.8	85.3	50.6	86.9
	DL20 nDCG@10	45.0	84.6	47.6	87.4

This scale study uses 6,980 dev, 43 DL19 and 54 DL20 queries. Dev reports MRR@10; DL19/DL20 report nDCG@10. Anchors are the original core-corpus documents, with no MS MARCO relevance labels used for fitting. These fixed-exit results do not use the optional readout selector. Full-depth references are in Appendix E.3.

Table 33: Qwen 4B, layer 27: Quora test scores for calibrated readouts selected on development queries. TP: Token Prepending; HTP: Hierarchical Token Prepending. Full-depth nDCG@10 is 0.8812; gain is attention minus each alternative.

Readout	Calibration scale α	nDCG@10	Quality retained (%)	Attention^2 gain (pp) [95% interval]
Ditto	$\hat{\alpha}$	0.5811	65.9	+30.53 [29.46, 31.61]
TP	$\hat{\alpha}$	0.8361	94.9	+1.59 [1.08, 2.11]
HTP	1	0.8418	95.5	+0.95 [0.49, 1.41]
Attention^2	$\hat{\alpha}$	0.8501	**96.5**	–

Gain is Attention^2 minus the readout in that row. Configuration selection uses development queries and the declared grid. The original intervals adjust over the three selected-family comparisons; selected variants are in Appendix B.5.

CLS-pooled encoder transfer. The CLS-row readout uses unprompted anchors, a separate unit-scale ridge map at each exit, and the native L24 document index. Across L4–L23, it improves over calibrated CLS in the pointwise paired analysis. Table 34 gives all 21 measured exits.

Table 34: **BGE-large: all 21 measured exits, L4–L24.** Three datasets, full-depth documents. Cells are macro nDCG@10 (retained %). Both readouts use their own unit-scale ridge calibration; this is an encoder composition study.

Exit layer	Native pooling + calibration	Attention^2 + calibration
4	0.147 (26.9)	0.189 (34.3)
5	0.155 (28.6)	0.231 (42.7)
6	0.170 (31.7)	0.232 (43.4)
7	0.178 (33.3)	0.245 (45.5)
8	0.218 (40.6)	0.269 (50.0)
9	0.232 (43.5)	0.297 (55.0)
10	0.223 (40.8)	0.304 (56.5)
11	0.213 (39.5)	0.327 (61.8)
12	0.154 (28.0)	0.254 (46.8)
13	0.105 (18.7)	0.210 (39.8)
14	0.055 (10.7)	0.163 (33.2)
15	0.047 (8.9)	0.249 (47.0)
16	0.026 (5.7)	0.117 (24.3)
17	0.027 (5.7)	0.145 (27.8)
18	0.029 (6.3)	0.214 (41.2)
19	0.035 (7.5)	0.299 (57.8)
20	0.066 (13.4)	0.321 (62.3)
21	0.264 (50.9)	0.417 (79.9)
22	0.401 (75.9)	0.467 (89.8)
23	0.489 (94.6)	0.501 (97.0)
24	0.517 (100.0)	0.520 (100.6)

E.4 Recovering encoder representations with a frozen readout

FTAE tests whether encoder states also retain retrieval evidence that their native pooling fails to expose. Unlike current-layer Attention^2, Frozen Terminal Attention Exit applies the frozen last-layer attention, output projection and LayerNorm to H_ℓ, omits the residual and feed-forward sublayer, then pools and L2-normalizes. It uses no fitted map or calibration anchors.

Table 35 separates bilateral retrieval from reuse of the native full-depth index. At L10 with full-depth documents, BGE-base rises from 0.046 to 0.459 nDCG@10 and e5-base-v2 from 0.255 to 0.421. The recovery depends on the readout and index: e5-base-v2 L6 declines under bilateral retrieval. This extends the evidence for accessible intermediate information beyond the particular calibrated readout used in the main experiments.

Table 35: **FTAE: all three measured exits per encoder.** Absolute macro nDCG@10 on SciFact, NFCorpus and FiQA. Full depth is L12. No alignment map or calibration anchors are used.

		Both sides shallow		Deep documents		
Model	Exit layer	Native	Frozen terminal attention	Native	Frozen terminal attention	Native full depth
BGE-base	6/12	0.005	0.256	0.003	0.301	0.489
	8/12	0.003	0.281	0.005	0.378	0.489
	10/12	0.041	0.400	0.046	0.459	0.489
e5-base-v2	6/12	0.136	0.089	0.139	0.292	0.468
	8/12	0.028	0.108	0.165	0.391	0.468
	10/12	0.038	0.216	0.255	0.421	0.468

BGE-base-en-v1.5 uses CLS pooling and e5-base-v2 uses mean pooling, both with maximum length 128 and padding masked. BGE prefixes queries with `Represent this sentence for`

`searching relevant passages:` and leaves documents unprefixed; E5 uses `query:` / `passage:` . At BGE L10 on the three core datasets, token mean obtains 0.162 nDCG, uniform mixing with the same terminal value/output projections and LayerNorm obtains 0.332, and FTAE obtains 0.459. The uniform-mixing control separates the contribution of attention weighting from the terminal transformation.

The seven-dataset extension below keeps native full-depth documents and uses its own full-depth references. It includes ArguAna instead of Touché and therefore differs from the main seven-dataset study. These quality results do not establish an encoding-time advantage; FTAE requires an additional attention computation.

	BGE-base		e5-base-v2	
Dataset	Frozen terminal attention, L10	Native full depth, L12	Frozen terminal attention, L10	Native full depth, L12
SciFact	0.682	0.722	0.632	0.676
NFCorpus	0.359	0.373	0.334	0.357
FiQA	0.335	0.371	0.296	0.370
ArguAna	0.396	0.420	0.326	0.295
SciDocs	0.201	0.219	0.176	0.185
Quora	0.869	0.887	0.787	0.862
TREC-COVID	0.647	0.732	0.694	0.654
Macro nDCG@10	0.498	0.532	0.463	0.486

E.5 Readout gains across document settings

Changing the readout and changing the document index address distinct parts of the retrieval problem. Figure 5 shows the attention gain within each study. Table 36 supplies the three-dataset full-depth-document scores corresponding to the bilateral attention queries and maps; input-matched controls are in Appendix C.4. Bilateral scores and the separate seven-dataset results appear only in their respective depth tables.

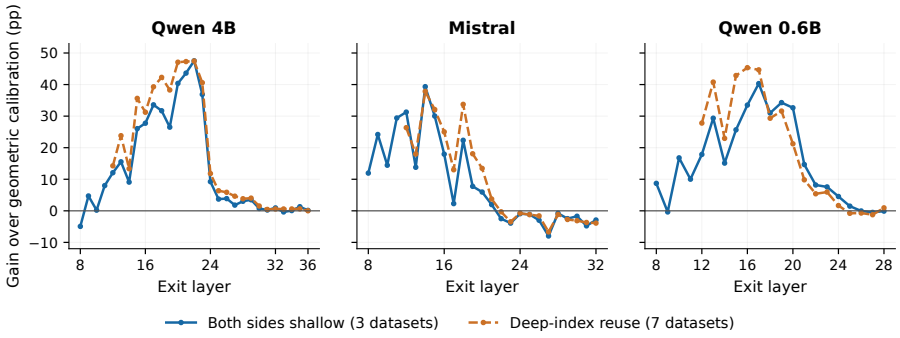


Figure 5: **Readout gain across depth.** Retention of Attention^2 + calibration minus retention of calibrated native pooling, in percentage points. Solid: both sides shallow, three datasets; dashed: the separate seven-dataset deep-index study. Each difference stays within its own study, and each readout has its own fitted map. All measured exits are plotted.

Table 36: Calibrated attention queries against the original full-depth document index, on three datasets. Cells give macro nDCG@10 (retention, %) relative to each model's own reference. Queries and maps match the bilateral study; document inputs are controlled in Table 21. Dashes mark unmeasured exits or exits beyond the model's depth.

Exit layer	Qwen 4B	Mistral 7B	Qwen 0.6B	GTE 7B
10	–	–	–	0.487 (81.6)
12	0.224 (40.8)	0.280 (56.4)	0.218 (45.4)	0.484 (81.4)
13	0.279 (50.7)	0.234 (48.4)	0.275 (57.3)	–
14	0.219 (40.2)	0.327 (66.3)	0.194 (39.8)	–

Table 36 (continued)

Exit layer	Qwen 4B	Mistral 7B	Qwen 0.6B	GTE 7B
15	0.369 (65.3)	0.299 (59.0)	0.261 (55.0)	–
16	0.348 (60.9)	0.377 (73.3)	0.306 (63.7)	0.527 (89.0)
17	0.393 (69.1)	0.352 (69.3)	0.340 (70.1)	–
18	0.408 (71.4)	0.453 (88.6)	0.379 (77.0)	–
19	0.391 (68.8)	0.449 (86.4)	0.404 (81.5)	–
20	0.438 (76.5)	0.488 (94.0)	0.440 (88.8)	–
21	0.466 (80.8)	0.489 (93.8)	0.449 (90.7)	–
22	0.493 (85.2)	0.482 (92.4)	0.461 (92.9)	0.552 (92.2)
23	0.478 (82.6)	0.490 (94.3)	0.468 (94.9)	–
24	0.534 (92.3)	0.504 (96.9)	0.458 (92.7)	–
25	0.548 (95.0)	0.493 (95.1)	0.475 (96.5)	–
26	0.544 (94.4)	0.501 (96.7)	0.480 (97.4)	–
27	0.550 (95.5)	0.488 (93.6)	0.490 (99.5)	–
28	0.555 (96.5)	0.507 (97.7)	0.496 (100.6)	0.593 (99.7)
29	0.558 (96.8)	0.500 (96.4)	–	–
30	0.566 (98.1)	0.511 (98.6)	–	–
31	0.566 (98.5)	0.506 (97.5)	–	–
32	0.569 (98.9)	0.505 (97.2)	–	–
33	0.569 (98.8)	–	–	–
34	0.572 (99.2)	–	–	–
35	0.578 (100.2)	–	–	–
36	0.577 (100.1)	–	–	–

E.6 Effect of the regularization floor at shallow exits

At Qwen 4B L8/L12, the floor in Equation 3 imposes a strong penalty relative to the variance of the attention readout. We test its effect on seven datasets with early-exit queries and the original full-depth document index fixed. We refit the same readouts with $\lambda = 0.01\,\mathrm{tr}(S_{xx})/d_x$, removing the floor, and recompute the analytic scale after each fit.

Table 37: Qwen 4B: effect of the ridge floor on calibrated Attention^2. Seven-dataset query-only nDCG@10 retention (%); documents remain at full depth.

Exit layer	Original recipe	Without ridge floor
8	12.0	51.4
12	37.5	65.4

The gains in Table 37 identify a calibration-induced loss in this query-only setting. In the same study, removing the first token from the attention readout also changes its variance. Under matched relative regularization, the remaining gain is small and the bilateral L20 difference is unresolved. First-token removal is not part of Attention^2.

Matched bilateral controls. We repeat the three-dataset setting of Table 1 at Qwen 4B L8/L12/L22/L30, changing only the ridge floor within each readout. Inputs, anchors, normalization and fitting arithmetic are fixed, and each fit uses the same analytic-scale rule. The additional content-mean control averages tokens overlapping the input text, excluding prompt-only and special tokens; empty text uses the last valid token. All maps use the same 45,000 unlabeled anchors and closed-form fit. These controls and their full-depth references are measured together on an RTX 5080. Intervals use 10,000 paired query resamples within datasets, with fitted maps fixed.

Table 38: Matched Qwen 4B readout and ridge controls. Queries and documents exit together on the three core datasets. Cells give macro nDCG@10 (retention, %), using full-depth references recomputed in the same run. Gain is the retention change after removing the floor, computed before rounding; brackets give pointwise 95% paired-bootstrap intervals.

Exit layer	Readout	No calibration		Calibrated: Original floor		Calibrated: Without floor		Gain (pp) [pointwise 95% CI]
8	Native	0.006	(1.4)	0.070	(11.8)	0.070	(11.8)	+0.00 [+0.00, +0.00]
	Content mean	0.260	(43.2)	0.377	(64.2)	0.377	(64.2)	+0.00 [+0.00, +0.00]
	Attention^2	0.004	(0.9)	0.041	(6.9)	0.189	(32.8)	+25.87 [+23.30, +28.54]
12	Native	0.006	(1.6)	0.097	(16.7)	0.097	(16.7)	+0.00 [+0.00, +0.00]
	Content mean	0.293	(49.6)	0.399	(69.1)	0.399	(69.1)	+0.00 [+0.00, +0.00]
	Attention^2	0.007	(1.6)	0.165	(28.6)	0.305	(54.1)	+25.52 [+23.13, +27.87]
22	Native	0.010	(2.2)	0.197	(32.9)	0.197	(32.9)	+0.00 [+0.00, +0.00]
	Content mean	0.112	(17.9)	0.409	(69.8)	0.409	(69.8)	+0.00 [+0.00, +0.00]
	Attention^2	0.063	(10.7)	0.464	(80.3)	0.467	(80.7)	+0.37 [+0.08, +0.68]
30	Native	0.486	(82.8)	0.535	(92.5)	0.535	(92.5)	+0.00 [+0.00, +0.00]
	Content mean	0.405	(68.5)	0.461	(78.8)	0.461	(78.8)	+0.00 [+0.00, +0.00]
	Attention^2	0.365	(62.3)	0.540	(93.3)	0.540	(93.3)	+0.00 [+0.00, +0.00]

Removing the floor raises Attention^2 retention by 25.9 and 25.5 pp at L8 and L12. The change is 0.4 pp at L22; all other entries are unchanged because the floor is inactive. Under either ridge rule, calibrated content mean outperforms Attention^2 at L8/L12, and this ordering reverses at L22/L30. Thus regularization explains part of the shallow attention gap, while simple aggregation also recovers substantial shallow retrieval evidence.

E.7 Retrieval gains beyond input embeddings

To measure what contextualization adds, we apply the same MaxSim scorer to input embeddings (L0) and L8/L12 states, retaining the prompted 192-token inputs and token masks. BM25 uses fixed regex tokenization, stopwords and no stemming ($k_1 = 0.9$, $b = 0.4$), on either full text or text decoded from the tokens visible to MaxSim. All methods use the same test queries and full corpora.

Table 39: Input-embedding controls with auxiliary lexical references, on the same test queries and full corpora. Cells give nDCG@10; Macro weights datasets equally. Qwen 4B L0 is the embedding lookup before any transformer block. Visible-text BM25 uses the text retained by the MaxSim input mask.

Readout or baseline	SciFact	NFCorpus	FiQA	Macro
Input embeddings (L0 MaxSim)	0.542	0.168	0.115	0.275
L8 MaxSim	0.665	0.280	0.221	0.389
L12 MaxSim	0.681	0.322	0.262	0.422
BM25 full text	0.662	0.312	0.244	0.406
BM25 visible text	0.561	0.155	0.203	0.306

L8/L12 MaxSim adds 0.114/0.146 macro nDCG@10 over L0 (pointwise 95% paired intervals [0.099, 0.128]/[0.130, 0.163]). Input embeddings already support retrieval, and contextual layers add measurable quality. The BM25 rows provide an auxiliary lexical reference.

Mistral token-matching inputs. The L1–L12 token comparison removes the instruction prefix from matching positions and retains the appended EOS. Query and document token vectors are normalized individually; intermediate states precede terminal normalization. Results and uncertainty are summarized in Appendix E.8.

E.8 PAIRED SHALLOW-LAYER COMPARISONS

The three-model comparison uses SciFact, NFCorpus and FiQA, with 240, 258 and 518 test queries. The preceding depth tables contain all MaxSim measurements; the paired comparisons below use the established L8–L12 calibrated results.

Paired uncertainty. Intervals use 5,000 paired query resamples within each corpus and fixed full-depth denominators. The simultaneous adjustment retains the full planned family of 24 contrasts per model (12 layers and two competitors), although only the validated L8–L12 comparisons are shown.

Table 40: MaxSim minus calibrated single-vector retrieval in nDCG@10 retention percentage points (pp). Positive values favor MaxSim. Intervals are simultaneous 95% for the recorded recipes.

Model	Exit layer	MaxSim − calibrated attention gain (pp) [95% interval]	MaxSim − calibrated native gain (pp) [95% interval]
Qwen 4B	8	+58.4 [+52.8, +64.1]	+53.5 [+47.9, +59.1]
	9	+52.0 [+46.4, +57.7]	+56.8 [+51.1, +62.4]
	10	+58.0 [+52.4, +63.6]	+58.2 [+52.6, +63.8]
	11	+49.1 [+43.5, +54.7]	+57.1 [+51.5, +62.7]
	12	+43.1 [+37.5, +48.8]	+55.2 [+49.6, +60.8]
Mistral	8	+30.1 [+23.9, +36.2]	+42.0 [+35.9, +48.1]
	9	+15.4 [+9.3, +21.5]	+39.6 [+33.5, +45.7]
	10	+22.0 [+15.9, +28.1]	+36.5 [+30.3, +42.6]
	11	-0.7 [-6.8, +5.4]	+28.7 [+22.6, +34.8]
	12	-3.6 [-9.7, +2.5]	+27.6 [+21.5, +33.7]
Qwen 0.6B	8	+59.7 [+53.6, +65.8]	+68.4 [+62.3, +74.5]
	9	+69.1 [+63.0, +75.2]	+68.7 [+62.7, +74.8]
	10	+46.6 [+40.6, +52.7]	+63.4 [+57.4, +69.5]
	11	+54.8 [+48.7, +60.9]	+64.8 [+58.7, +70.9]
	12	+39.3 [+33.3, +45.4]	+57.2 [+51.1, +63.3]

MaxSim exceeds both calibrated alternatives at Qwen 4B and 0.6B L8–L12 and Mistral L8–L10 under these simultaneous intervals. At Mistral L11–L12, the contrast with calibrated Attention^2 is unresolved. These comparisons apply to the stated recipes; the regularization dependence is discussed in Appendix E.6.