CELLCERT: EXACT VISUAL LATE-INTERACTION RETRIEVAL BY ADAPTIVE CELLWISE CERTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual late-interaction retrieval captures fine-grained page information, but exhaustive MaxSim scoring compares every query token with every page patch. Compression and shortlisting reduce this cost without guaranteeing the original ranking. Page-level exact search preserves that ranking, yet may score an entire page because only a few query-token interactions remain uncertain. We present CELLCERT, a deterministic search algorithm that refines document–query-token maxima, or cells, independently. MaxSim's additive structure allows the search to eliminate a page without resolving all of its cells. Quantization-aware geometric summaries bound each cell; a cost-normalized policy selects cells to refine, and group bounds exclude patches that cannot attain the cell maximum. The stopping rule certifies the original ordered top-k under the declared numerical and tie contract. Across 1,174 queries from five ViDoRe domains, including an untouched Pharmaceuticals domain, CELLCERT certifies and matches exhaustive ranking in all 3,522 query/k cases for $k \in \{1, 5, 10\}$. At $k = 5$, it reduces the fused multiply–accumulate (FMA) proxy by 51.4–57.9% across domains (55.4% query-weighted), and by 18.3% relative to a matched page-level certificate on Pharmaceuticals. The tested contrastive variants do not reduce work further. The current executor remains slower than fused exhaustive Flash-MaxSim: the demonstrated gain is in score-evaluation arithmetic, not wall-clock latency.

1 INTRODUCTION

Retrieval over visually rich PDFs requires matching a query to evidence in text, tables, charts, and page layout. Visual late-interaction models represent each page with hundreds or thousands of patch vectors and compare them with individual query-token vectors (Khattab & Zaharia, 2020; Faysse et al., 2025). Their MaxSim score sums, over query tokens, the largest similarity to any patch on the page. This fine-grained representation has a substantial query-time cost even when page embeddings are precomputed: exhaustive scoring of N pages with T query vectors and P patch vectors per page requires $O(NTPd)$ multiply–accumulates in embedding dimension d.

Existing approaches reduce this cost through representation compression or selective scoring. Token pooling and pruning reduce the stored page representation (Ma et al., 2025; Zong & Piwowarski, 2025; Kankanampati et al., 2026; Veneroso et al., 2025); candidate generators restrict exact MaxSim evaluation to a shortlist (Santhanam et al., 2022a; Dhulipala et al., 2024); learned routing selects which interactions to evaluate (Li et al., 2023). These approaches can offer favorable quality–cost trade-offs, but similar aggregate retrieval metrics need not imply the same retrieved pages. In our untouched-domain control, 128-centroid pooling improves nDCG@5 by 0.98 points while retaining only 78.4% of the exhaustive top-5 pages. A shortlist baseline matches the reported nDCG and recall values but still changes some top-k lists. When a downstream pipeline or regression test relies on a fixed scorer, such changes alter the retrieved evidence even if average quality is unchanged. We therefore study how to reduce MaxSim scoring work while certifying the original ordered top-k for each query.

Deterministic branch-and-bound can preserve the scorer by proving that omitted candidates cannot enter the top-k. Centroid–radius bounds are classical in maximum inner-product search (Ram & Gray, 2012) and have also been used for exact pruning in large-output computations (Liu et al., 2026). Applying them at page granularity, however, does not exploit the full structure of MaxSim. A page

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.

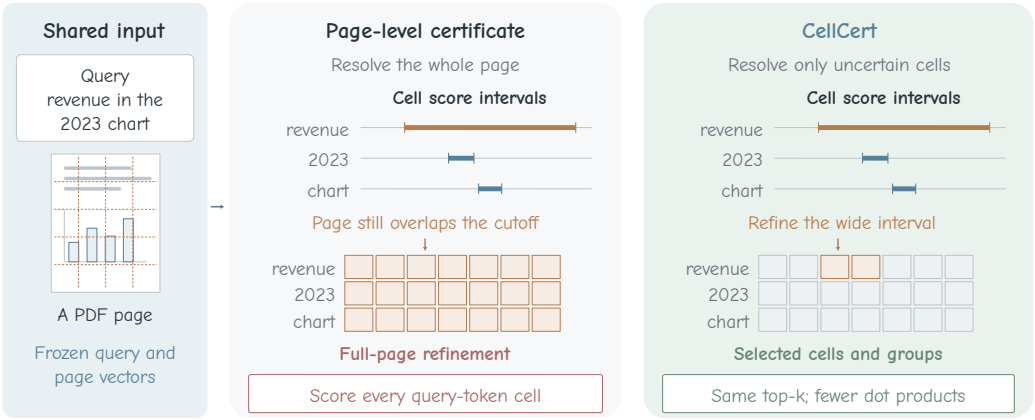


Figure 1: Pagewise versus cellwise refinement (schematic). Both methods receive the same frozen query and page vectors. A wide cell interval can keep a page competitive, triggering evaluation of every cell in page-level search. CELLCERT retains separate cell intervals and refines selected cells using only admissible patch groups. Orange and gray matrix entries denote evaluated and skipped interactions, respectively. The illustrated intervals and patch counts are not measured results.

upper bound sums a separate upper bound for every query-token maximum. If this sum overlaps the current cutoff, page-level refinement evaluates the entire page, even when most query-token contributions are already tightly bounded. Uncertainty in a few terms can thus trigger computation of all T terms. Figure 1 illustrates this mismatch between the source of uncertainty and the unit of refinement.

The central observation is that a valid bound on the sum does not require its terms to be resolved together. For document n and query token t, we call the maximum similarity to the page's patches a *cell*. MaxSim is the sum of these cells. Maintaining a separate interval for each cell allows the page bound to tighten one term at a time. The search can therefore certify a page's exclusion without computing its full score.

CELLCERT implements this principle using quantization-aware summaries of per-page patch groups. The summaries bound every cell, and a cost-normalized policy selects cells on pages that remain competitive. Refinement also operates within a cell: groups whose upper bounds lie below the cell's lower bound cannot contain its maximizing patch and need not be scored. The resulting procedure reduces work both by leaving cells unresolved and by restricting the patches evaluated in each selected cell. Valid bounds and the declared numerical contract determine correctness; the reveal policy determines only how much work precedes certification. Neither the retriever nor its MaxSim score is changed.

Our contributions are threefold.

- **Cellwise certification of visual MaxSim retrieval.** We formulate exact ordered top-k search as refinement of persistent document–query-token intervals and establish a stopping rule that separates ranking correctness from the choice of reveal policy.
- **Refinement across cells and within cells.** Quantization-aware group bounds support both partial page evaluation and exact scoring over selected patch groups. On Pharmaceuticals, this combination reduces top-5 FMA by 18.3% relative to a matched direct page-level certificate using the same group geometry.
- **Evidence on the exactness–cost trade-off.** Across five ViDoRe domains, including an untouched domain and two post-freeze robustness domains, CELLCERT certifies all 3,522 query/k cases and reduces the top-5 FMA proxy by 51.4–57.9%. The tested contrastive variants do not improve on the non-learned baselines, and the current executor remains slower than fused exhaustive scoring.

2 RELATED WORK

Visual late interaction and compression. ColBERT established token-level late interaction for text retrieval (Khattab & Zaharia, 2020), and ColPali extended it to page images (Faysse et al., 2025). MaxSim retains fine-grained interactions that a single fixed-dimensional vector can lose (Killingback et al., 2026), at the cost of storing and scoring many patch vectors. Light-ColPali, CRISP, learned pooling, and spatial Matryoshka training reduce the number of stored vectors (Ma et al., 2025; Veneroso et al., 2025; Josef, 2026; Roy et al., 2026); PULSAR combines pooled candidate generation with pooled late-interaction reranking in a deployed visual system (Constable et al., 2026). These methods change the representation to trade retrieval quality against cost. CELLCERT instead keeps the representation fixed and certifies which score computations can be omitted for each query. Its summaries supplement the original index rather than replace it, so arithmetic savings come with additional storage.

Efficient multi-vector search. ColBERTv2 compresses residual token vectors (Santhanam et al., 2022b); PLAID cascades centroid interaction, pruning, decompression, and reranking (Santhanam et al., 2022a); CITADEL learns sparse lexical routes (Li et al., 2023). DESSERT uses locality-sensitive sketches for set search (Engels et al., 2023), while MUVERA maps vector sets into fixed-dimensional encodings for candidate generation (Dhulipala et al., 2024). These methods generally trade exact agreement with original-space MaxSim for efficiency. Col-Bandit is more directly related to our approach: it adaptively evaluates query-token MaxSim values and eliminates documents using finite-population confidence intervals (Pony et al., 2026b). Its practical configurations provide empirical overlap or probabilistic guarantees. CELLCERT instead constructs deterministic intervals from per-document geometry and prunes patch groups within each selected cell. The distinction is the ranking guarantee and the granularity of exact refinement, rather than a claim of lower latency.

Exact geometric pruning and learned routing. Centroid–radius branch-and-bound is classical in exact MIPS (Ram & Gray, 2012). CSV-Decode applies related bounds to exact vocabulary top-k search with custom sparse kernels (Liu et al., 2026); OPTIMUS shows that indexed exact MIPS can be slower than blocked matrix multiplication and motivates hardware-aware fallback (Abuzaid et al., 2019). CELLCERT builds on these geometric inequalities. Its contribution is to compose them for a sum of query-token maxima, allowing partial cell refinement to tighten document intervals without computing a full page score. Query-aware quantization and learned cluster routing have also been studied for approximate MIPS (Guo et al., 2016; Vecchiato et al., 2024). Our contrastive controls test whether learned choices can reduce work while retaining an analytic certificate.

Contrastive hard-negative learning. Hard-negative mining trains dense retrievers to distinguish confusable query–document pairs (Xiong et al., 2021; Qu et al., 2021; Zhang & Stratos, 2021). In our setting, the target representation and score are fixed: learning can change the reveal order or the geometry used to compute bounds, but not the ranking being certified. We construct certificate-hard pairs from cells near the active winner–outsider boundary and test both reveal policies and scorer-preserving transforms. Unlike representation learning for relevance, their objective is to reduce the work needed to certify an unchanged result. The tested variants do not improve on the corresponding non-learned baselines.

3 ADAPTIVE CELLWISE CERTIFICATION

3.1 TARGET SCORER AND EXACTNESS CONTRACT

Let $Q = (q_1, \ldots, q_T)$ be the unit-normalized query vectors and let page $D_n = (x_{n1}, \ldots, x_{nP_n})$ contain unit-normalized patch vectors, all in $\mathbb{R}^d$. The frozen visual late-interaction score is

$$S_n \equiv S(Q, D_n) = \sum_{t=1}^{T} s_{nt}, \qquad s_{nt} = \max_{1 \le j \le P_n} q_t^\top x_{nj}. \tag{1}$$

The target is the ordered top-k produced by exhaustive evaluation of Eq. (1), with corpus-manifest order as the secondary key for equal scores. CELLCERT never changes Q, D_n, or Eq. (1). Its summaries decide which dot products must be evaluated; they are not replacement page embeddings. Figure 2 summarizes the offline index and query-time loop.

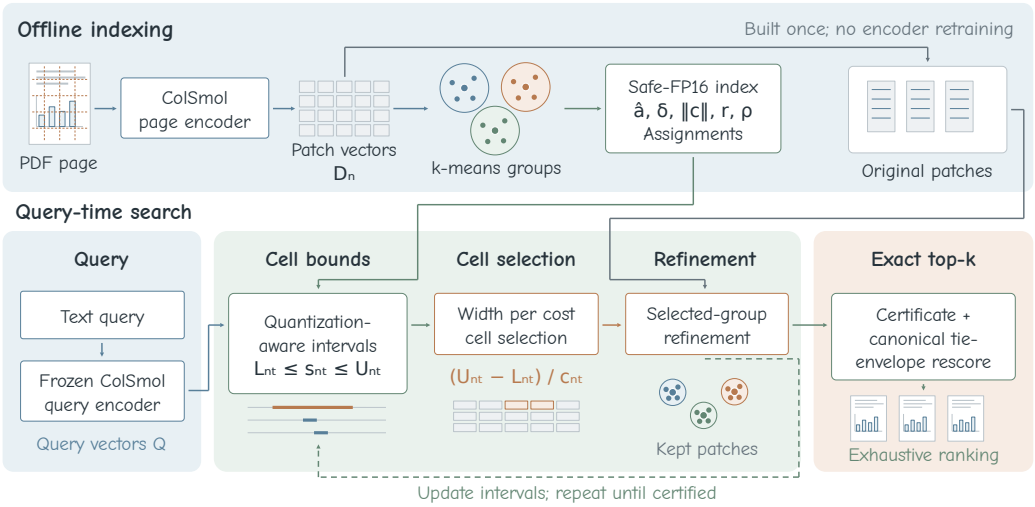


Figure 2: CELLCERT separates offline indexing (top) from query-time search (bottom). A frozen ColSmol encoder produces page vectors, which are grouped into quantization-aware summaries; the original patches are retained. Query vectors and summaries initialize cell bounds. Cost-normalized selection and selected-group refinement update these bounds until the stopping condition is met. Canonical rescoring then returns the same ordered top-k as exhaustive scoring under the declared numerical contract. The small page, vector, and group diagrams are schematic.

3.2 A QUANTIZATION-AWARE INTERVAL FOR EVERY CELL

Offline, we partition each page's patches into C nonempty groups using deterministic Euclidean k-means. Let c_{ng} be the mean of group G_{ng}, and define its radius $r_{ng} = \max_{x \in G_{ng}} \|x - c_{ng}\|_2$, unit centroid direction $a_{ng} = c_{ng}/\|c_{ng}\|_2$, and cap cosine $\rho_{ng} = \min_{x \in G_{ng}} a_{ng}^\top x$. A query's inner product with the group mean cannot exceed its maximum inner product with a group member:

$$\ell_{ntg} = q_t^\top c_{ng} \le \max_{x \in G_{ng}} q_t^\top x. \tag{2}$$

Two geometric bounds give upper bounds on the group maximum. Cauchy–Schwarz yields the Euclidean-ball bound $u_{ntg}^{\text{ball}} = q_t^\top c_{ng} + r_{ng}$. Writing $z = q_t^\top a_{ng}$, the spherical-cap bound is

$$u_{ntg}^{\text{cap}} = \begin{cases} 1, & z \ge \rho_{ng}, \\ z\rho_{ng} + \sqrt{1 - z^2}\sqrt{1 - \rho_{ng}^2}, & z < \rho_{ng}. \end{cases} \tag{3}$$

We take $u_{ntg} = \min(u_{ntg}^{\text{ball}}, u_{ntg}^{\text{cap}})$ and combine group bounds into a cell interval:

$$L_{nt} = \max_g \ell_{ntg}, \qquad U_{nt} = \max_g u_{ntg}, \qquad L_{nt} \le s_{nt} \le U_{nt}. \tag{4}$$

We store centroid directions in FP16 to reduce auxiliary storage. To account for quantization in an FP32 target scorer, the index also stores $\delta_{ng} \ge \|a_{ng} - \widehat{a}_{ng}\|_2$. We evaluate $\widehat{z} = q_t^\top \widehat{a}_{ng}$ and substitute $z^- = \max(-1, \widehat{z} - \delta_{ng})$ in the lower bound and $z^+ = \min(1, \widehat{z} + \delta_{ng})$ in both upper bounds. A fixed floating-point guard ε_{fp} is subtracted from group lower bounds and added to group upper bounds. The direction-error envelope is derived from the stored quantization error rather than fitted to test scores. The arithmetic guard is execution-stack specific: we validate it against materialized cell scores, and a change in precision or reduction kernel requires revalidation.

Summing cells yields a document interval

$$L_n = \sum_t L_{nt} \le S_n \le \sum_t U_{nt} = U_n. \tag{5}$$

Retaining the individual cell intervals allows one term in Eq. (5) to be tightened without evaluating the others.

Algorithm 1 CELLCERT query-time search

Require: Query Q, grouped pages and summaries, target k
1: Compute all $[L_{nt}, U_{nt}]$ and c_{nt}
2: **while** Eq. (9) fails and a threatening unresolved cell exists **do**
3: Mark unresolved cells of threatening documents
4: Select a batch by Eq. (8)
5: Evaluate selected cells by Eq. (7)
6: Replace their intervals and update L_n, U_n
7: **end while**
8: Resolve remaining winner cells; canonical-rescore the tie envelope
9: **return** stable ordered top-k and certificate

3.3 EXACT REFINEMENT INSIDE ONE CELL

Group bounds can also reduce the cost of refining a selected cell (n, t). Define the admissible groups as

$$\mathcal{G}_{nt} = \{g : u_{ntg} \geq L_{nt} - \eta\}, \tag{6}$$

where $\eta \geq 0$ is the numerical certificate margin. Every omitted group has maximum score below a valid lower bound on the cell maximum. Therefore at least one maximizing patch remains in $\cup_{g \in \mathcal{G}_{nt}} G_{ng}$, and

$$s_{nt} = \max_{g \in \mathcal{G}_{nt}} \max_{x \in G_{ng}} q_t^\top x \tag{7}$$

is exact. This direct scorer accesses the original patch vectors only in admissible groups. Its estimated cost is $c_{nt} = d \sum_{g \in \mathcal{G}_{nt}} |G_{ng}|$ FMAs.

3.4 COST-NORMALIZED ADAPTIVE REVEAL

The search state consists of all cell intervals and their document sums. Let W be the stable top-k pages under current lower bounds and $\tau = \min_{n \in W} L_n$. A document remains competitive, or *threatening*, if $U_n \geq \tau - \eta$. Across unresolved cells of all threatening documents, the scheduler prioritizes interval width relative to refinement cost:

$$\pi_{nt} = \frac{U_{nt} - L_{nt}}{\max(c_{nt}, 1)}. \tag{8}$$

We reveal the 256 highest-priority available cells as a batch, replace their intervals by the exact values from Eq. (7), and update only the affected document sums. The procedure stops early when

$$\min_{n \in W} L_n > \max_{m \notin W} U_m + \eta. \tag{9}$$

If strict separation is impossible because the cutoff contains an exact tie, the reveal loop instead ends when no unresolved threatening cell remains; all pages in the boundary envelope then enter the canonical path. Any unresolved cells of the winners are evaluated to return exact scores and order. Finally, to absorb reduction-order differences, we apply the canonical dense scorer to the winners and any page whose upper bound lies within 5×10^{-5} of the cutoff. Exact ties are ordered by manifest position.

Correctness. **Theorem 1.** *If the group intervals contain their exact group maxima and the final scorer follows the declared tie rule, Algorithm 1 returns the same ordered top-k as exhaustive Eq.* (1)*.*

Proof sketch. Equations (4) and (5) hold initially. Equation (6) retains a global cell maximizer, so refinement replaces an interval by its contained exact value and preserves the invariant. There are two termination cases. Under Eq. (9), each winner has $S_n \geq L_n$, whereas every outsider has $S_m \leq U_m < \min_{n \in W} L_n$ after the numerical margin. Otherwise, every threatening page is exact because it has no unresolved cell; every nonthreatening page has $U_m < \tau - \eta$, below the exact boundary candidates. Thus only exact pages can occupy the cutoff envelope. Canonical rescoring and the manifest secondary key then recover their reference order. Appendix A gives the full argument. □

3.5 WORK AND STORAGE

Let A_{nt} be the number of original patches scored when cell (n, t) is revealed, and let R denote the set of revealed cells. Ignoring lower-order scalar bound operations, the query FMA count is

$$\mathrm{FMA}_{\textsc{CellCert}} = NTCd + d \sum_{(n,t)\in R} A_{nt} + \mathrm{FMA}_{\mathrm{tie}}, \quad \mathrm{FMA}_{\mathrm{full}} = Td \sum_{n} P_n. \tag{10}$$

We report their ratio per query, charging the initial summary scan, all dot products with selected patches, and the canonical rescore. For each page, the compact auxiliary index stores C groups, each with one FP16 direction and four FP32 scalars, plus one uint16 assignment per original patch when $C < 2^{16}$. Its size is $NC(2d + 16) + 2\sum_n P_n$ bytes. Original patches remain available because exactness may require them. Index construction is independent across pages and performed once.

4 EXPERIMENTS

We evaluate four aspects of CELLCERT: agreement with the exhaustive scorer, arithmetic reduction across domains and values of k, the roles of refinement granularity and reveal policy, and runtime on current GPU executors.

4.1 EXPERIMENTAL PROTOCOL

Data and scorer. We evaluate five public English visual-document retrieval tasks: Economics Reports v2 (452 pages, 58 queries), Biomedical Lectures v2 (1,016 pages, 160 queries), and the ViDoRe v3 Finance (2,942 pages, 309 queries), Industrial (5,244 pages, 283 queries), and Pharmaceuticals (2,313 pages, 364 queries) domains (Faysse et al., 2025; Loison et al., 2026). The unmodified ColSmol-256M checkpoint uses ColPali-style late interaction and produces 128-dimensional query and page vectors.

Economics and Biomedical serve as development domains. We selected Pharmaceuticals by a fixed hash rule before downloading, encoding, or scoring it, and froze $C = 384$, seed 17, 12 Lloyd iterations, batch size 256, $\eta = 10^{-5}$, and $k \in \{1, 5, 10\}$. Finance and Industrial were evaluated without adaptation after the method and hyperparameters were fixed. They are robustness domains rather than untouched domains because earlier rejected method families had inspected them. All three v3 domains use the specified FP32 scorer. A protocol correction restored that precision for Pharmaceuticals; the earlier FP16 run was retained as a diagnostic. Appendices B and C document dataset revisions and corrections.

Metrics. Exact ordered agreement compares each returned list with exhaustive MaxSim, including manifest-order tie breaking. Certificate rate is the fraction of queries that terminate with a valid certificate; we also check every lower and upper bound against materialized cell scores. We measure retrieval quality with nDCG@k and Recall@k, and arithmetic work with the per-query ratio in Eq. (10). Confidence intervals for the mean use 10,000 query-level bootstrap resamples with seed 1701. The FMA ratio is a hardware-independent work proxy, not a latency measurement. We report auxiliary storage using the compact layout in Section 3.5 and measure runtime with both GPU-resident and pinned-host data.

Comparisons. Exhaustive MaxSim defines the reference scores and ranking. The flat exact certificate uses the same $C = 384$ group geometry but refines each surviving page as a unit, using either direct selected-group scoring or a centroid-complement identity. Approximate controls on Pharmaceuticals comprise 128-vector k-means and recursive-bisection pooling, an 80-group plus 48-singleton routed pool, MUVERA-style query ball carving at a fixed cosine threshold of 0.8, and 128-vector pooling followed by exact rescoring of a 128-page shortlist. Flash-MaxSim provides the fused exhaustive runtime baseline (Pony et al., 2026a). We do not compare directly with published PLAID, MUVERA, or PULSAR latencies, which use different models, corpora, and hardware. The controlled comparisons instead measure fidelity to the same frozen scorer. Appendix C documents the sequence of method selection and evaluation.

Table 1: CELLCERT results on all evaluated queries. FMA is relative to exhaustive MaxSim; lower is better. CIs are 95% query-bootstrap intervals. Ordered agreement and certificate rate are 100% throughout. * marks the untouched domain; † marks post-freeze robustness domains.

Domain	k	nDCG	Recall	FMA ratio (95% CI)	Reduction
Economics	1	0.5172	0.0528	0.4163 [0.4113, 0.4217]	58.4%
	5	0.5281	0.2977	0.4526 [0.4455, 0.4601]	54.7%
	10	0.5192	0.4399	0.4910 [0.4804, 0.5019]	50.9%
Biomedical	1	0.4750	0.2933	0.3965 [0.3947, 0.3985]	60.4%
	5	0.5268	0.5726	0.4209 [0.4179, 0.4239]	57.9%
	10	0.5596	0.6749	0.4367 [0.4336, 0.4399]	56.3%
Finance†	1	0.5620	0.2884	0.3919 [0.3882, 0.3957]	60.8%
	5	0.5555	0.5193	0.4279 [0.4235, 0.4325]	57.2%
	10	0.5836	0.6230	0.4478 [0.4428, 0.4528]	55.2%
Industrial†	1	0.3993	0.2251	0.3938 [0.3902, 0.3975]	60.6%
	5	0.4320	0.4181	0.4259 [0.4214, 0.4305]	57.4%
	10	0.4479	0.4860	0.4421 [0.4373, 0.4470]	55.8%
Pharmaceuticals*	1	0.4799	0.2622	0.4613 [0.4596, 0.4632]	53.9%
	5	0.5362	0.5133	0.4859 [0.4831, 0.4887]	51.4%
	10	0.5640	0.6284	0.5005 [0.4974, 0.5036]	49.9%

Table 2: Fixed approximate controls on Pharmaceuticals. ΔnDCG@5 is relative to exhaustive original MaxSim. Top-5 overlap measures set overlap with the exhaustive result, irrespective of order. Equal aggregate quality need not imply ranking identity.

Method	Relative work/storage	ΔnDCG@5	Top-5 overlap	Exact?
128 k-means centroids	14.7% storage	+0.0098	78.41%	No
128 bisection centroids	14.7% storage	-0.0240	67.20%	No
80 core + 48 singleton	14.7% storage	+0.0044	76.54%	No
Query ball carving	45.0% query tokens	+0.0056	93.74%	No
Pool-128, shortlist-128	20.2% FMA	+0.0000	99.89%	No
CELLCERT	48.59% FMA	0.0000	100.00%	Yes

4.2 MAIN RESULT: EXACTNESS WITH FEWER INTERACTIONS

Table 1 summarizes 1,174 queries and 3,522 query/k decisions over 11,967 indexed pages. Every returned list agrees with exhaustive scoring, and every search terminates with a valid certificate. The smallest FMA reduction is 49.9%, on Pharmaceuticals at $k = 10$. At $k = 5$, reductions range from 51.4% to 57.9%, with a query-weighted reduction of 55.4% [55.2%, 55.7%]. Work increases with k, as more pages remain competitive and more winner cells require exact evaluation. The bootstrap intervals reflect uncertainty across queries, not retraining variance: the scorer and search configuration are fixed.

We check bound validity in addition to ranking agreement. Across the five domains, the largest discrepancy between the sum of materialized cells and the cached exhaustive page score is 7.63×10^{-6}. The minimum upper-bound slack is 1.99×10^{-5}; lower-bound slack is nonnegative, with equality in feasible-centroid cases on the three FP32 domains. At $k = 5$, the mean number of canonically rescored pages is 5.000 in each domain, so the tie envelope adds no pages beyond the winners on average. Duplicate or score-tied Biomedical pages also exercise the manifest-order tie rule, with all returned lists matching the reference order.

4.3 FIDELITY OF APPROXIMATE ALTERNATIVES

When ranking fidelity is not required, approximate methods can offer lower scoring or storage costs. Table 2 shows why exactness remains a separate evaluation criterion. On Pharmaceuticals, k-means pooling improves nDCG@5 while replacing more than one fifth of the exhaustive top-5 membership. Conversely, shortlist-128 matches the reported retrieval metrics but misses a top-5 page in a small

Table 3: Refinement statistics at $k = 5$. Cells and Pages are mean revealed fractions. Med./P95 denote the median and 95th percentile of query-level FMA ratios; Launches estimates the number of refinement batches.

Domain	Cells	Pages	Launches	Med.	P95
Economics	35.7%	78.4%	20.2	0.4515	0.4970
Biomedical	23.9%	56.5%	30.6	0.4181	0.4590
Finance†	32.0%	82.3%	135.6	0.4275	0.5035
Industrial†	30.9%	77.3%	219.7	0.4201	0.4946
Pharmaceuticals*	25.6%	67.9%	83.5	0.4806	0.5396

number of queries. CELLCERT preserves both the aggregate metrics and every ordered list under its exactness contract, at the cost of more arithmetic and additional storage.

4.4 WHAT PRODUCES THE REDUCTION?

Cellwise versus pagewise decisions. On Pharmaceuticals at $k = 5$, the flat direct certificate has an FMA ratio of 0.5946 and evaluates 1,369 of 2,313 pages exactly on average. Centroid-complement reconstruction lowers this ratio to 0.5454 but requires FP32 centroids and scalar reconstruction. Cellwise direct refinement reaches 0.4859, reducing FMA by 18.3% relative to the matched direct pagewise baseline and by 11.0% relative to complement reconstruction. Because the direct pagewise baseline uses the same group geometry, this comparison isolates the benefit of retaining query-token-specific uncertainty: a surviving page need not have every cell evaluated.

Table 3 quantifies how much of a surviving page is evaluated. Although 56.5–82.3% of pages receive at least one reveal, only 23.9–35.7% of all document–query-token cells are resolved. The 95th-percentile FMA ratio is at most 0.5396 across domains, indicating that the arithmetic reduction is not confined to a small set of easy queries. However, the estimated number of refinement launches rises from 20.2 on Economics to 219.7 on Industrial. These small, repeated operations help explain the runtime results in Section 4.5.

Reveal policy and contrastive learning. We trained static and state-dependent MLP policies with contrastive supervision favoring cells that reduce the certificate gap. On the first 32 queries of both development domains, every policy remains exact because only reveal order is learned. The analytic width-per-cost rule attains FMA ratios 0.4195/0.4101 (Economics/Biomedical). Averaged over three seeds, the static policy gives 0.4328/0.4141 and the dynamic policy 0.4294/0.4147; width-only is 0.4592/0.4265. A scorer-preserving dual transform trained against certificate-hard negative upper envelopes also uses 0.28% more work than identity on the first development domain and fails its prespecified evaluation gate. Neither tested policy improves on the analytic scheduler under this training budget. The cost-normalized rule therefore remains the default. Appendix B gives the policy-training settings.

Partition robustness and group budget. Across three partition seeds at $C = 384$, the sample standard deviation of the mean FMA ratio is 1.5×10^{-4} on Economics and 2.3×10^{-4} on Biomedical; every run is exact. Varying $C \in \{64, 128, 192, 256, 384, 512\}$ reveals a non-monotonic trade-off: more groups tighten intervals but increase the mandatory $NTCd$ summary scan. Appendix D reports the full curve. This diagnostic was specified after the untouched-domain run and was not used to revise the primary $C = 384$ configuration.

Storage. The safe summary keeps FP16 directions but retains FP32 error and geometry scalars. Including compact uint16 patch-to-group assignments, its auxiliary bytes are 36.7–47.7% of a standardized FP16 token index across the five domains (18.4–23.8% of the FP32 token indexes used by the three v3 scorers). The original patch index cannot be removed: exact fallback is part of the contract. These byte counts assume compact arrays and exclude the additional overhead of Python-list serialization.

4.5 RUNTIME AND FAILURE ANALYSIS

All runtime experiments use one RTX 5090, three interleaved forward/reverse repeats, CUDA-event synchronization, and a warm-up covering every query shape. Exhaustive Flash-MaxSim scores all 2,313 Pharmaceuticals pages with median latency 0.421 ms for GPU-resident data and 9.51 ms for data in pinned host memory. The adaptive Python/PyTorch executor is slower. On Economics/Biomedical, its GPU-resident medians are 11.45/18.76 ms, compared with 0.191/0.298 ms for Flash-MaxSim; the pinned-host medians are 78.46/69.07 ms, compared with 2.35/4.86 ms. A more regular one-shot survivor portfolio and a union-pruned Flash variant are also slower in both storage regimes.

The gap is consistent with prior evidence that exact MIPS with fewer dot products can still be slower than blocked dense computation (Abuzaid et al., 2019). CELLCERT replaces a fused operation with repeated reductions, masks, gathers, and host-visible stopping decisions. On Pharmaceuticals, query-token count has Spearman correlation 0.35 with the top-5 FMA ratio, whereas the exhaustive $k/k{+}1$ score margin has correlation -0.07. These correlations suggest that the cutoff margin alone does not explain the observed variation in work. A fused persistent kernel and group-contiguous storage are potential ways to reduce execution overhead, but remain untested here. The evaluated portfolio selects the Flash-MaxSim fallback in every measured regime; the current results do not establish an end-to-end speedup.

Limitations. The largest corpus has 5,244 pages, and all five tasks share one benchmark family and one checkpoint; results do not establish web-scale or cross-model behavior. Auxiliary summaries add storage, per-page k-means adds offline build cost, and the present executor is substantially slower than dense GPU kernels. Exactness is relative to a declared checkpoint, precision, and tie rule; changing any of them requires rebuilding or revalidating the summaries. Finally, preserving the scorer also preserves its relevance errors and biases. The experiments establish deterministic fidelity and arithmetic reduction, not better retrieval quality or downstream answer generation.

5 CONCLUSION

CELLCERT performs exact visual late-interaction retrieval by refining individual query-token maxima rather than complete page scores. Per-cell intervals and group bounds allow the search to omit work both across cells and within a selected cell while preserving the original ordered top-k. With a frozen ColSmol scorer on five ViDoRe domains, the method matches every exhaustive result and reduces the top-5 FMA proxy by 51.4–57.9%. The tested contrastive variants do not improve on the non-learned baselines, and the current executor remains slower than fused exhaustive scoring. The results demonstrate the arithmetic benefit of cellwise certification; realizing a latency benefit requires further work on kernels and memory layout.

AI USE STATEMENT

Generative AI tools were used primarily to improve the readability and language of the manuscript.

ETHICS STATEMENT

This work studies retrieval efficiency on public benchmark pages and does not involve human subjects or the release of personal data. Exact ranking fidelity does not make a retriever's relevance judgments unbiased: biases in the underlying visual-language model and benchmark remain unchanged. The additional summary index also increases storage and should be considered when deploying on sensitive document collections.

REPRODUCIBILITY STATEMENT

Section 3 defines the scorer, interval construction, stopping rule, and work accounting. Section 4.1 specifies the evaluation setup. Appendices A–C provide the correctness argument, dataset revisions, model and precision settings, and the sequence of protocol decisions and corrections.

REFERENCES

Firas Abuzaid, Geet Sethi, Peter Bailis, and Matei Zaharia. To index or not to index: Optimizing exact maximum inner product search. In *IEEE 35th International Conference on Data Engineering*, pp. 1250–1261, 2019. doi: 10.1109/ICDE.2019.00114. URL https://arxiv.org/abs/1706.01449.

Benjamin Constable, Anup Roy, Vishal Sharma, Rishabh Gyanendra Upadhyay, Robin Mills, and Aidan Philip Millar. PULSAR: Pooled unified late-interaction search and retrieval for enterprise visual document RAG. *arXiv preprint arXiv:2608.28572*, 2026. URL https://arxiv.org/abs/2608.28572.

Laxman Dhulipala, Majid Hadian, Rajesh Jayaram, Jason Lee, and Vahab Mirrokni. MUVERA: Multi-vector retrieval via fixed dimensional encodings. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/hash/b71cfefae46909178603b5bc6c11d3ae-Abstract-Conference.html.

Joshua Engels, Benjamin Coleman, Vihan Lakshman, and Anshumali Shrivastava. DESSERT: An efficient algorithm for vector set search with vector set queries. In *Advances in Neural Information Processing Systems*, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/d6cc45de2e2dea14b96c1eba88fd8ef7-Abstract-Conference.html.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/99e9e141aafc314f76b0ca3dd66898b3-Abstract-Conference.html.

Ruiqi Guo, Sanjiv Kumar, Krzysztof Choromanski, and David Simcha. Quantization based fast inner product search. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pp. 482–490. PMLR, 2016. URL https://proceedings.mlr.press/v51/guo16a.html.

Stefan Josef. Learn to pool: Lightweight fine-tuning for flexible multi-vector compression. *arXiv preprint arXiv:2607.06036*, 2026. URL https://arxiv.org/abs/2607.06036.

Yash Kankanampati, Yuxuan Zong, Nadi Tomeh, Benjamin Piwowarski, and Joseph Le Roux. A voronoi cell formulation for principled token pruning in late-interaction retrieval models. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 756–766, 2026. doi: 10.1145/3805712.3809726. URL https://arxiv.org/abs/2603.09933.

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, 2020. doi: 10.1145/3397271.3401075. URL https://arxiv.org/abs/2004.12832.

Julian Killingback, Varad Ingale, Hamed Zamani, and Cameron Musco. Quantifying and expanding the theoretical capacity of late-interaction retrieval models. *arXiv preprint arXiv:2607.05803*, 2026. URL https://arxiv.org/abs/2607.05803.

Minghan Li, Sheng-Chieh Lin, Barlas Oguz, Asish Ghoshal, Jimmy Lin, Yashar Mehdad, Wen tau Yih, and Xilun Chen. CITADEL: Conditional token interaction via dynamic lexical routing for efficient and effective multi-vector retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://aclanthology.org/2023.acl-long.663/.

Dong Liu, Shu Wang, Yanxuan Yu, Haisheng Wang, and Ben Lengerich. CSV-Decode: Certifiable sub-vocabulary decoding for efficient large language model inference. *arXiv preprint arXiv:2511.21702*, 2026. URL https://arxiv.org/abs/2511.21702.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. ViDoRe v3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. *arXiv preprint arXiv:2601.08620*, 2026. URL https://arxiv.org/abs/2601.08620.

Yubo Ma, Jinsong Li, Yuhang Zang, Xiaobao Wu, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Jiaqi Wang, Yixin Cao, and Aixin Sun. Towards storage-efficient visual document retrieval: An empirical study on reducing patch-level embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. URL https://aclanthology.org/2025.findings-acl.1003/.

Roi Pony, Daniel Ezer, Adi Raz Goldfarb, Idan Friedman, Oshri Naparstek, and Udi Barzelay. FLASH-MAXSIM: IO-aware fused kernels for late-interaction retrieval. *arXiv preprint arXiv:2605.29517*, 2026a. URL https://arxiv.org/abs/2605.29517.

Roi Pony, Adi Raz Goldfarb, Oshri Naparstek, Idan Friedman, Udi Barzelay, and Eli Schwartz. ColBandit: Query-time top-k estimation for late-interaction retrieval. *arXiv preprint arXiv:2602.02827*, 2026b. URL https://arxiv.org/abs/2602.02827.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. URL https://aclanthology.org/2021.naacl-main.466/.

Parikshit Ram and Alexander G. Gray. Maximum inner-product search using tree data-structures. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 931–939, 2012. doi: 10.1145/2339530.2339677. URL https://arxiv.org/abs/1202.6101.

Trishan Singha Roy, Arkadeep Acharya, Vishwajeet Kumar, Jaydeep Sen, and Sachindra Joshi. ColSNAP: Spatial matryoshka training for multi-granularity visual document retrieval. *arXiv preprint arXiv:2608.29951*, 2026. URL https://arxiv.org/abs/2608.29951.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. PLAID: An efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, 2022a. URL https://arxiv.org/abs/2205.09707.

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, 2022b. doi: 10.18653/v1/2022.naacl-main.272. URL https://aclanthology.org/2022.naacl-main.272/.

Thomas Vecchiato, Claudio Lucchese, Franco Maria Nardini, and Sebastian Bruch. A learning-to-rank formulation of clustering-based approximate nearest neighbor search. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. doi: 10.1145/3626772.3657931. URL https://arxiv.org/abs/2404.11731.

João Veneroso, Rajesh Jayaram, Jinmeng Rao, Gustavo Hernández Ábrego, Majid Hadian, and Daniel Cer. CRISP: Clustering multi-vector representations for denoising and pruning. *arXiv preprint arXiv:2505.11471*, 2025. URL https://arxiv.org/abs/2505.11471.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=zeFrfgyZln.

Wenzheng Zhang and Karl Stratos. Understanding hard negatives in noise contrastive estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. URL https://aclanthology.org/2021.naacl-main.86/.

Yuxuan Zong and Benjamin Piwowarski. Towards lossless token pruning in late-interaction retrieval models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2407–2417, 2025. doi: 10.1145/3726302.3730100. URL https://arxiv.org/abs/2504.12778.

A Complete correctness argument

For every document n, query token t, and group g, let $m_{ntg} = \max_{x \in G_{ng}} q_t^\top x$. By construction, $\ell_{ntg} \le m_{ntg} \le u_{ntg}$. The lower inequality follows because a linear functional at the group mean equals the mean of its values on the group and cannot exceed their maximum. The ball inequality follows from $q_t^\top (x - c_{ng}) \le \|q_t\|_2 \|x - c_{ng}\|_2 \le r_{ng}$. Equation (3) is the support function of the intersection of the unit sphere with $a_{ng}^\top x \ge \rho_{ng}$. Taking the smaller of two upper bounds preserves validity. The stored direction error satisfies $|q_t^\top (a_{ng} - \widehat{a}_{ng})| \le \delta_{ng}$, so the quantized substitutions enclose the unquantized inner product before the arithmetic guard is applied.

Because $s_{nt} = \max_g m_{ntg}$, maxima over the group lower and upper bounds give Eq. (4); summation gives Eq. (5). For Eq. (6), let g^* contain a maximizing page patch. If g^* were omitted, then $m_{ntg^*} \le u_{ntg^*} < L_{nt} - \eta \le s_{nt} - \eta \le s_{nt}$, contradicting $m_{ntg^*} = s_{nt}$. Direct refinement therefore returns s_{nt}. By induction, replacing a cell interval with this value preserves the validity of every document interval.

At termination, let W be the stable lower-bound top-k and $\tau = \min_{i \in W} L_i$. If Eq. (9) holds, then every $n \in W$ satisfies $S_n \ge L_n \ge \tau$, whereas every $m \notin W$ satisfies $S_m \le U_m < \tau - \eta$. Hence no outsider can enter W.

It remains to cover the exact-tie case, for which strict separation may be impossible. If the loop exits because no threatening unresolved cell exists, every page with $U_n \ge \tau - \eta$ has all of its cells materialized and therefore $L_n = U_n = S_n$. This set includes W, since $U_n \ge L_n \ge \tau$ for $n \in W$. Every page outside this threatening set satisfies $S_m \le U_m < \tau - \eta$ and cannot reach the boundary. Thus the only pages that can occupy or tie the cutoff are already exact. In either termination case, pages within the numerical tie envelope are evaluated through the same dense path as the reference, and stable sorting applies the same manifest secondary key. The returned ordered list therefore equals exhaustive scoring.

B Reproducibility details

Environment. Experiments use Python 3.12.9, PyTorch 2.8.0 with CUDA 12.8, Sentence Transformers 6.0.1, one NVIDIA GeForce RTX 5090 (32,607 MiB reported by the driver), and fixed seeds. Embeddings retain the number of vectors produced for each page; padding vectors are never included in scoring. Development scores use the stored FP16 embeddings; all three v3 reports normalize their stored FP32 embeddings in FP32. Flash runtime uses FP16 query/document inputs because that is the supported fused path and is reported separately from the FP32 certification results.

Checkpoint and preprocessing. We use the public ColSmol-256M checkpoint, a SmolVLM-based multi-vector retriever. Page images are converted to RGB and passed to its document encoder; query strings are passed to its query encoder using the checkpoint's default processor. We retain English queries and their relevance judgments while keeping the full page corpus. Table 4 identifies the frozen dataset revisions; the repository names and revision prefixes uniquely identify the evaluated data. No retriever fine-tuning is performed for CellCert.

Numerical policy. For each quantized group direction, we store its realized ℓ_2 error plus a 2×10^{-7} guard. Group upper bounds are increased and lower bounds decreased by 2×10^{-5}; termination uses $\eta = 10^{-5}$; the final canonical rescore envelope is 5×10^{-5}. These constants were fixed before the untouched evaluation. Tests cover bound validity, ragged pages, quantized directions, direct group masks, cell batching, stable ties, and agreement with exhaustive top-k.

Contrastive-policy controls. Policy training uses fixed 512-page subsets from ViDoRe v3 Computer Science and Energy, with deterministic training/validation splits. Economics and Biomedical

Table 4: Public evaluation datasets and frozen revision prefixes. All repositories are under the `vidore` organization on Hugging Face; links resolve to the full revisions.

Dataset repository	Revision
`economics_reports_v2`	`76fe40166b`
`biomedical_lectures_v2`	`c475466573`
`vidore_v3_finance_en`	`7f432c176d`
`vidore_v3_industrial`	`e26c864724`
`vidore_v3_pharmaceuticals`	`3abd4aa8a9`

are excluded from optimization and early stopping. The static and dynamic policies are $8 \to 32 \to 1$ and $12 \to 32 \to 1$ MLPs with GELU activations. Both use AdamW (learning rate 10^{-3}, weight decay 10^{-4}), at most 40 epochs, early-stopping patience 6, and seeds 17, 29, and 43. A listwise contrastive cross-entropy matches policy probabilities to max-normalized oracle utilities at temperature 0.1. The static target rewards raising winner lower bounds or lowering competitive outsider upper bounds per unit cost. The dynamic target uses the current lower-bound winner set and reranks up to 1,024 heuristic candidates to choose 256 reveals. These controls use centroid-complement costs for training features and oracle utilities, with learned-policy computation charged in evaluation. Their ordering never changes the analytic stopping condition.

C Frozen protocols and correction ledger

Method selection preceded the untouched evaluation. The cellwise policy and $C = 384$ were selected on development queries. The untouched domain was chosen by sorting eligible dataset identifiers with SHA-256 of a fixed salt and taking the smallest digest. Pharmaceuticals was then fixed at a public dataset commit before download. Direct selected-group scoring replaced centroid-complement reconstruction only after passing a two-domain development criterion and before any Pharmaceuticals metric was observed.

Two subsequent changes corrected implementation inconsistencies rather than retuning the method. The first Pharmaceuticals run incorrectly cast the declared FP32 scorer input to FP16; it was retained as a diagnostic, the correction was recorded, and the complete evaluation was rerun. A later audit found that the evaluator still serialized the old FP32 complement index although the finalized direct-scoring configuration specified the safe-FP16 summary. The corrected run changed only the summary representation and its conservative error envelope. All partitions, queries, scorer values, and hyperparameters remained fixed for this second correction.

After the method and hyperparameters were frozen, Finance and Industrial were designated as additional robustness domains before either CellCert result was inspected. Both reuse the untouched-run hyperparameters and evaluate every query without adaptation. They are reported as robustness rather than untouched domains because earlier rejected method families had inspected those datasets.

D Group-budget diagnostic

Table 5: Post-freeze C sensitivity on the first 32 development queries at $k = 5$. Entries are FMA ratios relative to exhaustive MaxSim; lower is better. Every configuration preserves ordered top-5 and terminates with a certificate. These diagnostics did not alter the primary configuration.

Domain	64	128	192	256	384	512
Economics	0.6632	0.5660	0.4992	0.4568	0.4447	0.5089
Biomedical	0.5691	0.4314	0.3721	0.3635	0.4208	0.5222

Increasing C raises the mandatory summary cost linearly but can reduce refinement work by tightening cell intervals. The best trade-off therefore depends on page lengths and group geometry, and need not

occur at the same C across domains. The full curve makes this dependence explicit; a larger group budget is not uniformly better.

E ADDITIONAL NEGATIVE RESULTS

We also evaluated several scorer-preserving or certificate-preserving variants. Multiple partitions tightened individual token bounds but increased the summary scan cost and did not meet the work-reduction criterion. Second-moment and mean-conservation bounds did not offset their scalar and storage costs. A centroid-complement identity saves one vector dot product per selected group but requires scalar reconstruction and an FP32 mean; it reduces work in the flat arithmetic simulator but is less suitable for direct selected-group execution. Static and dynamic contrastive reveal policies outperform random and width-only controls in work reduction, but not width divided by reveal cost. Paired transforms $x' = xA$, $q' = qA^{-\top}$, with vectors written as rows and A invertible, preserve every token dot product; neither the covariance-based nor the certificate-contrastive choice reduced work relative to identity. None of these variants was used for the primary five-domain result.