INDEPENDENT CHANNEL MIXTURES FOR BUDGETED EMBEDDING REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a small number of noisy ternary comparisons improve retrieval without retraining an encoder or accessing the retrieval gallery? We introduce **Independent Channel Mixture (ICM)**, a method that refines query embeddings through a fixed budget of ternary comparisons. Each comparison presents a reference item and two anchors, returning one binary answer indicating which anchor is closer to the reference. ICM estimates feedback consistency, accounts for ambiguity in the answers, and uses the existing embedding geometry to control the resulting update. Our analysis characterizes its response to uninformative feedback and bounds changes in retrieval scores. In a post-hoc subset of 36 dataset–representation combinations from eight dataset families, ICM achieves relative improvements in hierarchical nDCG@10 of up to **16.48%** in the strongest condition, with a **3.03% family-balanced average gain** across noise levels from 0.0 to 0.4. At noise 0.5, it preserves the frozen retrieval scores in almost every evaluated condition. These results demonstrate the potential of limited, imperfect feedback for adapting frozen embeddings to a retrieval task.

1 INTRODUCTION

A frozen embedding model makes retrieval inexpensive. Just encode each item once, then rank the gallery by similarity to a query. Its geometry, however, reflects the distinctions emphasized during pretraining. A user may care about a scientific subfield, or a product-category boundary that is only partly represented by those similarities. Retraining requires information and computation that may be unavailable at interaction time. A complementary approach is to modify the query embedding using a small number of ternary comparisons while preserving the encoder and gallery index.

We consider an explicitly limited form of this approach. The learner sees reference embeddings and a separate collection of anchor embeddings. For a reference item x_i, it can ask whether anchor a or anchor b is closer to that item. We call this a ternary comparison because it involves three items and its answer contains only one bit. The learner cannot inspect gallery embeddings, target labels, or oracle noise. Its output must therefore remain meaningful in the original embedding coordinates.

Three difficulties arise together. First, *repeated answers reveal consistency without establishing semantic usefulness*. In a hierarchy, two anchors may be equally related to the reference. Therefore a stable forced choice between them can still be highly repeatable. Second, *a sparse transcript may not distinguish a sharp binary response link from a link with a broad region of indifference*. Third, *a direction that explains anchor preferences can move the query too far for a concentrated, fine-grained neighborhood*. Improving response prediction is consequently insufficient to guarantee improved gallery retrieval.

ICM addresses these difficulties through three separate quantities. A shared reliability estimate describes repeatability. Independent query-local mixtures express uncertainty between two response links. A covariance-derived scale controls how much inferred movement reaches the final query. The acquisition policy uses a broad latent tangent model to seek informative comparisons, while the output scale is fixed by the visible neighborhood before any answers arrive. A whole-batch evidence gate retains the frozen queries when the transcript does not provide sufficient evidence against independent fair responses.

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.

Our contribution separates *what the responses support* from *how far the embedding should move* under the restricted interface:

- **Method.** Paid calibration, query-local response mixtures, and geometry-scaled delivery refine queries without gallery access (Section 3; Figure 1).
- **Analysis.** We distinguish predictive evidence, one-step uncertainty reduction, and score-stability bounds. (Section 4).
- **Evaluation.** Text, images, audio, and satellite measurements cover shallow and deep hierarchies. All eight family means improve. (Section 6; Appendix E).

Relation to prior work. Prior work studies closely related but distinct settings. Active embedding search estimates latent preference vectors from noisy comparisons (Canal et al., 2019), adaptive pairwise comparison learns user preferences (Qian et al., 2015), and crowd-kernel learning (Tamuz et al., 2011) and stochastic triplet embedding (van der Maaten & Weinberger, 2012) learn similarity representations from comparative judgments. Query adaptation also connects to relevance feedback (Rocchio, 1965). TOUR (Sung et al., 2023) and ReFIT (Gangi Reddy et al., 2024) refine query representations using reranker feedback on retrieved corpus items. Task-adaptive embedding refinement uses test-time LLM relevance scores on retrieved documents (Gera et al., 2026). Our setting instead restricts feedback to budgeted ternary comparisons over a separate anchor pool and withholds the retrieval gallery during refinement. Their published results therefore do not provide a direct numerical comparison under our protocol. Our experiments measure improvement over frozen embeddings. Although we still compare ICM with implementations adapted to our ternary-comparison setting (Appendix E), these comparisons evaluate the adapted implementations rather than the original published systems. RankRefine++ (Wijaya et al., 2026) motivates calibrated post-training feedback. It refines a scalar regressor using pairwise rankings and numerically labeled references. ICM receives no such labels and returns a query vector for unseen gallery items. An anchor-only binary interface can elicit a preference without requiring a numerical relevance scale or exposing retrieved gallery items. Ternary comparisons specify this constraint, because the contribution is the separation of paid response evidence from geometry-controlled delivery.

2 Task and information available to the learner

Let $Q = (q_i)_{i=1}^N \in \mathbb{R}^{N\times d}$ and $A = (a_j)_{j=1}^m \in \mathbb{R}^{m\times d}$ contain unit embeddings of references and anchors. Reference i is the item whose retrieval query is to be refined. An anchor is an item that may appear in a comparison. A ternary comparison consists of the triple (x_i, a, b), one reference item and two anchors. Its oracle call:

$$Y_{i,t} = \mathrm{ask}(i, a, b) \in \{0, 1\} \tag{1}$$

returns one if anchor a is preferred to anchor b for reference i. Here *ternary* describes the three-object input, not a third response category, so no tie response is exposed. Each call, including repetitions, costs one unit. The learner spends exactly B calls per reference and returns unit vectors $\widehat{Q}$. Thus a batch spends NB calls. A trusted evaluator subsequently ranks disjoint gallery items $g \in G$ by $\widehat{q}_i^\top g$.

The learner receives exactly six inputs: Q, A, opaque routing identifiers, an ask-only oracle handle, B and a method seed.

The method seed controls reproducible candidate-pair sampling from visible embeddings. It is distinct from the evaluator's private seed for persistent tie resolution and simulated flips, which is never supplied to the learner. Responses may be pooled across the batch. ICM therefore assumes a common response-noise channel across the references used for calibration. *Independent* denotes the query-local latent choice of response model. The implementation isolates the method in a child process and audits its imports and data access.

Controlled oracle. In the benchmark, item x has an evaluator-only path $h(x)$ in a rooted hierarchy. With LCP denoting the longest common prefix, its tree distance is:

$$d_T(x, z) = |h(x)| + |h(z)| - 2|\,\mathrm{LCP}(h(x), h(z))|. \tag{2}$$

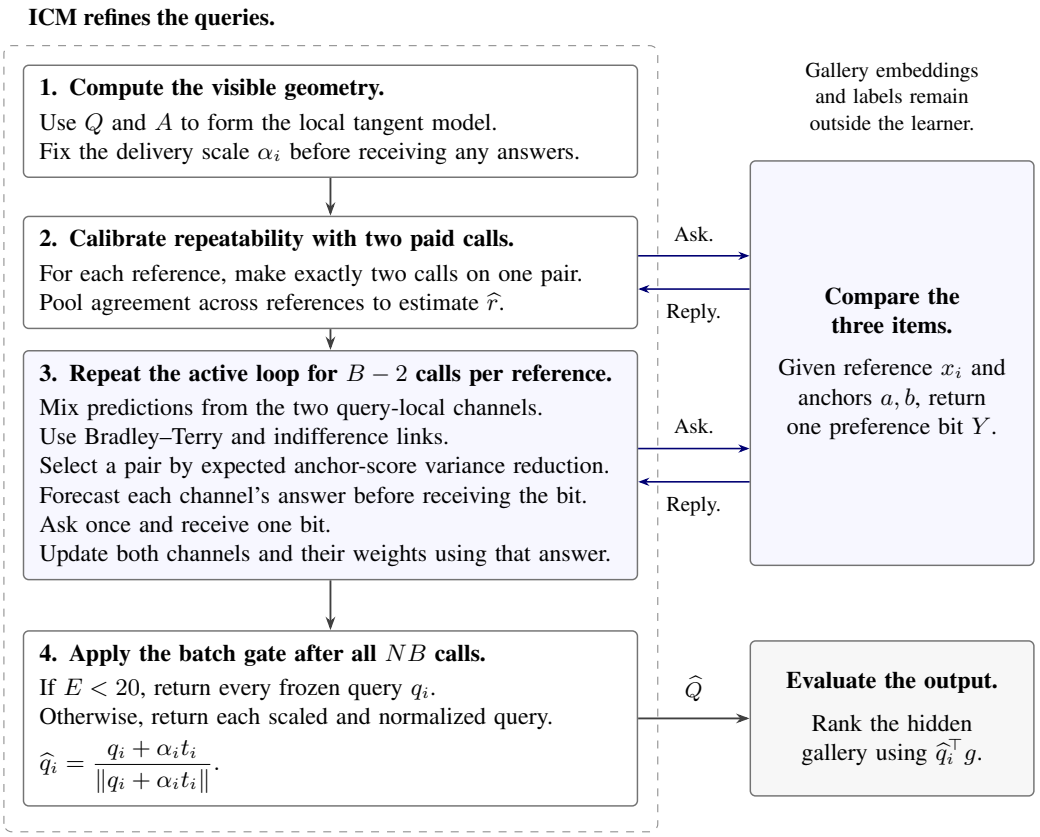

Figure 1: ICM separates paid feedback from the scale of the delivered update. Two pilot calls per reference estimate shared repeatability. The remaining calls use a query-local mixture of two response channels to choose comparisons, predict answers, and update the model. Only after all NB calls does the batch gate decide whether to deliver the scaled update or retain the frozen queries. The oracle supplies one bit per call, and the evaluator alone accesses the disjoint gallery.

The clean answer prefers the anchor with smaller distance to the reference. Equal distances are resolved by a persistent, query-specific keyed order of anchors. Each answer is independently flipped with probability $\eta \in [0, 1/2]$ in the statistical model. The simulator uses keyed, repeat-indexed hashes to implement the noise reproducibly. In particular, $\eta = 0$ means unflipped hierarchy-based choices. At $\eta = 1/2$, answers carry no information under the ideal independent-flip model.

Appendix H gives the exact simulator procedure. Paths, noise levels, and hash keys belong to the evaluator. Meanwhile the learner receives only the returned bit.

Retrieval target. The primary outcome is hierarchical nDCG@10, using normalized discounted cumulative gain (Järvelin & Kekäläinen, 2002) with *linear* gain $v_i(g) = [1 + d_T(x_i, g)]^{-1}$:

$$\mathrm{nDCG}\,@k(i) = \frac{\sum_{r=1}^{k} v_i(g_{(r)})/\log_2(r+1)}{\sum_{r=1}^{k} v_i(g^*_{(r)})/\log_2(r+1)}, \qquad k = 10. \tag{3}$$

The ideal order $g^*_{(r)}$ is computed over the complete hidden gallery. Hierarchy depth and gallery support both affect the meaning of this score.

3 INDEPENDENT CHANNEL MIXTURE

3.1 LOCAL GEOMETRY, INFERENCE AND DELIVERY SCALES

For each reference q_i, select its nearest $n_i \le 64$ anchors by cosine similarity. Center these anchors and project onto the tangent space of the unit sphere at q_i. Let the resulting matrix have retained SVD $H_i = P_i \Sigma_i V_i^\top$ with rank $r_i \le 63$, and define:

$$C_i = V_i \operatorname{diag}\left(0.8\frac{\sigma_{ij}^2}{n_i} + 0.2\frac{\sum_{\ell=1}^{r_i}\sigma_{i\ell}^2}{n_i r_i}\right) V_i^\top, \qquad \tau_i = \operatorname{tr}(C_i). \tag{4}$$

Here r_i denotes retained geometric rank. For a degenerate neighborhood, set $C_i = 0$ and take F_i to be empty. The local directions are retained, but each directional variance mixes 80% of its empirical value with 20% of the retained mean variance. Thus broad directions remain broader while extreme variance differences are moderated. When $\tau_i > 0$, choose a tangent factor $F_i \in \mathbb{R}^{r_i \times d}$ satisfying $F_i^\top F_i = (2.25/\tau_i)C_i$. The inference model is $U_i = q_i + F_i^\top \theta_i$, with $\theta_i \sim \mathcal{N}(0, I)$. Its trace is fixed for exploration. The delivered displacement instead uses $\alpha_i = \tau_i/2.25$, computed before feedback. The neighbor cap 64, shrinkage weights, and inference trace $2.25 = 1.5^2$ are global development settings fixed for the reported run. Appendix G records their provenance and numerical accuracy.

For a candidate pair, define $z_{iab} = F_i(a_a - a_b)/s_i$ and $b_{iab} = q_i^\top(a_a - a_b)/s_i$, where $s_i = \max\{\operatorname{std}(Aq_i), .05\}$. Its latent standardized contrast is $m_{iab} = b_{iab} + z_{iab}^\top \theta_i$. A single ambient query direction is inferred from many anchor contrasts, allowing evaluation against gallery items that never appear in an interaction.

3.2 PAID CALIBRATION AND UNCERTAIN RESPONSE LINKS

Repeatability calibration. ICM spends its first two calls per reference on the same geometry-selected pilot pair. Write $I_i = 1$ when the two answers agree. If $r = 1 - 2\eta$ is response reliability, then:

$$\mathbb{P}(I_i = 1 \mid r) = (1 + r^2)/2, \qquad p(r \mid I_{1:N}) \propto \prod_{i=1}^{N} \left(\frac{1+r^2}{2}\right)^{I_i} \left(\frac{1-r^2}{2}\right)^{1-I_i}, \quad 0 \le r \le 1. \tag{5}$$

Here r denotes response reliability. The prior is uniform, and subsequent fitting uses 48-node Gauss–Legendre quadrature for the one-dimensional posterior mean $\widehat{r} = \mathbb{E}[r \mid I_{1:N}]$ (accuracy check in Appendix G). An agreeing pilot contributes one directional observation with effective reliability $2\widehat{r}/(1+\widehat{r}^2)$. A disagreeing pilot contributes no direction. Persistent clean tie choices obey the same agreement identity in (5), which is why calibration estimates repeatability.

Query-local channel mixture. Each query maintains two experts with links:

$$g_0(m) = \sigma(m), \qquad g_1(m) = \tfrac{1}{2}\{\sigma(m-\gamma_i)+\sigma(m+\gamma_i)\}, \qquad p_k(1 \mid m, r) = \tfrac{1-r}{2} + r g_k(m), \tag{6}$$

where σ is the logistic function. Here g_0 is a Bradley–Terry response link (Bradley & Terry, 1952). The nonnegative width γ_i is the standardized gap between the tenth-highest and median initial anchor scores. It is fixed from visible geometry. The second link represents a broad region of low directional sensitivity.

The experts start with equal prior probability and use warm-started maximum a posteriori (MAP) fits with inverse prior-plus-expected-Fisher precision as covariance. Predictive probabilities integrate each link against this Gaussian surrogate using one-dimensional quadrature. A predictive likelihood L_{ik} multiplies directional probabilities. Pilot factors condition on the complete agreement vector without using the scored direction. Later factors are formed before the new answer is observed. The independent mixture weights are:

$$w_{ik} = \frac{L_{ik}}{L_{i0} + L_{i1}}, \qquad k \in \{0, 1\}. \tag{7}$$

The two experts receive the same paid transcript. Only reliability calibration and the final gate pool evidence across queries.

Algorithm 1 Independent Channel Mixture

Require: Q, A, opaque IDs, ask-only oracle, budget $B \geq 2$, method seed.
Compute tangent factors, scales s_i, γ_i, and delivery coefficients α_i;
Ask each pilot pair twice; compute $I_{1:N}$ and shared reliability $\widehat{r}$;
Initialize experts; score and fit agreeing pilot directions, then update weights;
for $t = 3, \ldots, B$ **do**
    for each reference i **do**
        Select an unused pair by (8); compute both predictive probabilities;
        Ask once; update predictive likelihoods and MAP/Fisher states;
    end for
    Update query-local weights by (7);
end for
Compute batch evidence E from (9);
return Q if $E < 20$; otherwise return the unit vectors in (10);

3.3 Acquisition, evidence gate, and final query

Acquisition. Let $p = \mathbb{E}[Y]$ and $c = \mathrm{Cov}(U_i, Y)$ under the current two-expert surrogate for a proposed pair. Let M_i be the weighted centered scatter matrix of the current top-32 anchors. ICM approximately maximizes, using numerical mixture moments:

$$\mathcal{A}_i(a, b) = \frac{c^\top M_i c}{p(1-p)}. \tag{8}$$

Both within-expert uncertainty and between-expert disagreement contribute to c. The candidate pool consists of unused pairs among the top 24 anchors together with at most 256 sampled unused pairs. This makes selection depend on uncertainty in locally relevant anchor scores. The precise moments, weights, and pilot rule appear in Appendix A.

Batch gate. Let D count scored directional factors, including one for each agreeing pilot, and let L_{ik} refer to those factors. ICM forms:

$$E = E_{\mathrm{cal}}\, 2^D \prod_{i=1}^{N} \left(\tfrac{1}{2}L_{i0} + \tfrac{1}{2}L_{i1}\right), \quad E_{\mathrm{cal}} = \sum_{h=1}^{48} \nu_h \prod_{i=1}^{N} (1 + r_h^2)^{I_i} (1 - r_h^2)^{1-I_i}. \tag{9}$$

Here $\nu_h > 0$, $\sum_h \nu_h = 1$ are the calibration quadrature weights. If $E < 20$, the whole batch retains its frozen queries. This decision occurs after spending the prescribed budget. Under the conditions in Proposition 3, it controls release under fair responses.

Delivered query. When the gate accepts, ICM forms the mixture tangent mean and applies the geometry-derived scale:

$$t_i = \sum_{k=0}^{1} w_{ik} F_i^\top \widehat{\theta}_{ik}, \qquad \widehat{q}_i = \frac{q_i + \alpha_i t_i}{\|q_i + \alpha_i t_i\|}, \quad \alpha_i = \frac{\tau_i}{2.25}. \tag{10}$$

The role of α_i is to retain a broad exploratory model while limiting the displacement delivered in concentrated neighborhoods. The following algorithm describes the default configuration.

4 Evidence, acquisition, and geometric stability

The analysis isolates three different objects: the predictive likelihood on the paid transcript, surrogate uncertainty over anchor scores, and perturbations of unseen gallery scores. Proofs of Propositions 1–3 and separate geometric derivations are given in Appendix B. The statements have different targets: transcript prediction, current surrogate uncertainty, and score stability.

Proposition 1 (Mixture protection on a common transcript)**.** *For positive prequential expert likelihoods L_{i0}, L_{i1}, let $Z = \prod_i (\frac{1}{2}L_{i0} + \frac{1}{2}L_{i1})$. Then:*

$$-\log Z \leq \sum_{i=1}^{N} \min_{k \in \{0,1\}} (-\log L_{ik}) + N \log 2. \tag{11}$$

This is the finite-mixture dominance inequality of Hutter (2003, Eq. (2)), applied independently to each reference. Its cost is at most $\log 2$ relative to the better link on the same paid transcript.

Proposition 2 (Anchor-score variance reduction)**.** *For any square-integrable U, Bernoulli Y with $0 < p = \mathbb{E}[Y] < 1$, and fixed positive semidefinite M, writing $c = \mathrm{Cov}(U, Y)$ gives:*

$$\mathrm{tr}(M\,\mathrm{Cov}(U)) - \mathbb{E}[\mathrm{tr}(M\,\mathrm{Cov}(U \mid Y))] = \frac{c^\top M c}{p(1-p)}. \tag{12}$$

Proposition 2 motivates (8) under exact moments of the current joint surrogate and a fixed frontier. This follows the variance-reduction approach to active learning (Cohn et al., 1996). ICM evaluates these moments numerically.

Proposition 3 (Fair-response gate)**.** *Under independent fair binary responses, condition on the complete pilot-agreement vector before scoring pilot directions. If agreeing-pilot and later forecasts are normalized and predictable conditional on that vector and preceding scored directions, the ideal statistic in (9) satisfies $\mathbb{E}_0[E] = 1$ and $\mathbb{P}_0(E \geq 20) \leq .05$.*

The result controls one fixed-budget batch. Repeatable but semantically uninformative forced choices need not satisfy its null.

Bounds on delivered score changes. Unit anchors imply $0 \leq \tau_i \leq 1$, hence $0 \leq \alpha_i \leq 4/9$. For any tangent vector $t \perp q$, define $q(\alpha) = (q + \alpha t)/\|q + \alpha t\|$. Its Euclidean displacement, the chord length between the two unit vectors, is:

$$\epsilon(\alpha)^2 = \|q(\alpha) - q\|^2 = 2\left(1 - \frac{1}{\sqrt{1 + \alpha^2\|t\|^2}}\right). \tag{13}$$

This increases with $\alpha \geq 0$. Every unit-normalized gallery embedding satisfies $|(q(\alpha) - q)^\top g| \leq \epsilon(\alpha)$, and every pair whose raw similarity gap exceeds $2\epsilon(\alpha)$ keeps its order.

This is a sufficient margin condition. Since those margins are hidden, the bound controls score movement without certifying that every retrieval ranking improves.

A complementary local sensitivity interpretation appears beside its derivation in Appendix B.

5 Experimental design

Evaluated grid. The reported subset includes 36 dataset–representation combinations across eight source families and nine cohorts. Every combination uses $N = 128$ reference queries and $m = 256$ disjoint anchors, all budgets $B \in \{8, 16, 32, 64, 128, 256\}$, and all noise levels $\eta \in \{0, .1, .2, .3, .4, .5\}$. The resulting $36 \times 6 \times 6 = 1{,}296$ distinct conditions are repeated with five method/oracle replicates (seeds 0–4), giving 6,480 run cells, 829,440 query-run evaluations, and 69,672,960 paid oracle calls.

Hierarchies and input domains. The grid spans text, images, audio, and satellite measurements with native hierarchies of depth two to seven (Table 1). Cohort definitions and terminal-path support are detailed in Appendix C.

Representation coverage. The grid covers 16 distinct pretrained checkpoints plus statistical controls, from lightweight text and spectral representations to Qwen3-Embedding-8B. The encoders remain frozen throughout all runs. The complete model inventory and construction records appear in Appendix C and Table 6.

Aggregation and evidence status. For each replicate, we first average the 30 informative budget–noise conditions within each representation, then average representations within a cohort, cohorts within a family, and finally the eight families equally. Relative gain is $g_s = 100(S_s/S_{\mathrm{Frozen}} - 1)$. We report the mean and sample standard deviation (σ) across the five complete replicate aggregates. σ measures variation across method and private oracle seeds on fixed cohorts, not uncertainty over new cohorts. Appendix I gives the estimator and individual seed summaries. Pure noise ($\eta = .5$) is reported separately in Table 3.

Table 1: Benchmark coverage. Depth counts semantic ranks. Every row uses 128 queries and 256 anchors. WOS has two cohorts within one family. Sizes describe selected galleries.

Cohort	Input	Enc.	Gallery	Semantic depth
WOS-6k / WOS-12k	Abstracts	4 / 4	600 / 1,000	2 / 2
WDC-222	Product text	3	800	2–3
BGC single-path	Book text	8	10,000	1–4; queries 1–3
CIFAR-100	Images	2	1,000	2
Aircraft	Images	4	5,000	3
iNaturalist 2021	Images	3	50,000	7
ESC-50	Audio	5	512	2
TreeSatAI	Radar / optical	3	1,000	4

Table 2: Family-balanced hierarchical nDCG@10 at $\eta = 0.0$–0.4, averaged across six budgets. ICM and gain report the mean $\pm$ sample σ over five fixed-cohort method replicates. Raw is invariant across replicates.

Family	Enc.	Raw	ICM (Our) $\pm\sigma$	Gain (%) $\pm\sigma$
Aircraft	4	0.4971	**0.5044** $\pm$ 0.0004	1.46 $\pm$ 0.08
BGC	8	0.6550	**0.6775** $\pm$ 0.0003	3.44 $\pm$ 0.04
CIFAR-100	2	0.6453	**0.6631** $\pm$ 0.0007	2.76 $\pm$ 0.12
ESC-50	5	0.6313	**0.6418** $\pm$ 0.0001	1.68 $\pm$ 0.02
iNaturalist	3	0.3317	**0.3359** $\pm$ 0.0010	1.25 $\pm$ 0.31
TreeSatAI	3	0.3706	**0.3908** $\pm$ 0.0004	5.44 $\pm$ 0.12
WDC	3	0.8372	**0.8556** $\pm$ 0.0008	2.19 $\pm$ 0.09
WOS	8	0.7287	**0.7702** $\pm$ 0.0004	5.70 $\pm$ 0.05
Family-balanced	36	0.5871	**0.6049** $\pm$ 0.0003	3.03 $\pm$ 0.05

6 Results

6.1 Average improvement and the scale of the largest gain

Table 2 shows a family-balanced increase from 0.587112 to 0.604912 hierarchical nDCG@10, an absolute gain of 0.017800 and a relative gain of **3.03% $\pm$ 0.05 percentage points**. All eight family means increase. Gains are largest for WOS (5.70%) and TreeSatAI (5.44%), while iNaturalist gains 1.25%. The unbalanced gain is $3.43\% \pm 0.04$ percentage points. Including the null control gives a family-balanced gain of $2.53\% \pm 0.04$ percentage points. The informative-noise summary remains primary.

The largest five-seed condition-mean relative improvement is **16.48% $\pm$ 1.03 percentage points**, on TreeSatAI with concatenated Sentinel-1/Sentinel-2 spectral PCA, $B = 256$, and $\eta = 0$. Its mean score rises from 0.413364 to 0.481490 over 128 queries.

6.2 Budget and noise dependence

Table 3 reports the full six-by-six aggregate grid for the retained subset, with five-seed means and sample σ values. At zero noise, the relative gain increases from $2.54\% \pm 0.07$ percentage points with eight comparisons to $6.93\% \pm 0.12$ percentage points with 256. With 40% flips, the corresponding gains are $0.04\% \pm 0.05$ percentage points and $1.87\% \pm 0.22$ percentage points. The aggregate gain is nondecreasing with budget at each informative noise level. Figure 2 resolves these curves by family.

At $\eta = .5$, 4 of 1,080 run-level gates falsely release, 1,076 reject and preserve all per-query retrieval scores exactly. Two released batches improve nDCG and two decline. The aggregate relative change is $+0.0002\% \pm 0.0025$ percentage points. These are gate false positives. The guarantee permits a nonzero release probability. Across informative runs, 4,691 gates accept and 709 reject. Low budgets at high noise often remain unchanged because the available evidence does not pass the batch threshold. Figure 3 shows release rates and the distribution of 1,080 five-seed condition means.

Table 3: Family-balanced relative hierarchical nDCG@10 change (%) for every noise and budget. Each cell is the mean ± sample σ across five fixed-cohort replicates. σ values are in percentage points. Extra precision is retained for the null row. Noise $\eta = 0.5$ is a separate null control.

η / B	8	16	32	64	128	256
0.0	2.54 ± 0.07	3.64 ± 0.04	4.65 ± 0.03	5.57 ± 0.08	6.38 ± 0.11	6.93 ± 0.12
0.1	1.98 ± 0.04	2.80 ± 0.05	3.72 ± 0.06	4.68 ± 0.12	5.50 ± 0.17	6.15 ± 0.12
0.2	1.29 ± 0.08	1.97 ± 0.08	2.71 ± 0.10	3.54 ± 0.17	4.52 ± 0.14	5.33 ± 0.16
0.3	0.56 ± 0.09	0.94 ± 0.06	1.77 ± 0.10	2.30 ± 0.14	3.12 ± 0.09	3.88 ± 0.09
0.4	0.04 ± 0.05	0.14 ± 0.06	0.37 ± 0.08	0.72 ± 0.08	1.33 ± 0.19	1.87 ± 0.22
0.5	-0.002 ± 0.004	-0.000 ± 0.006	+0.003 ± 0.007	+0.000 ± 0.000	+0.000 ± 0.000	+0.000 ± 0.000

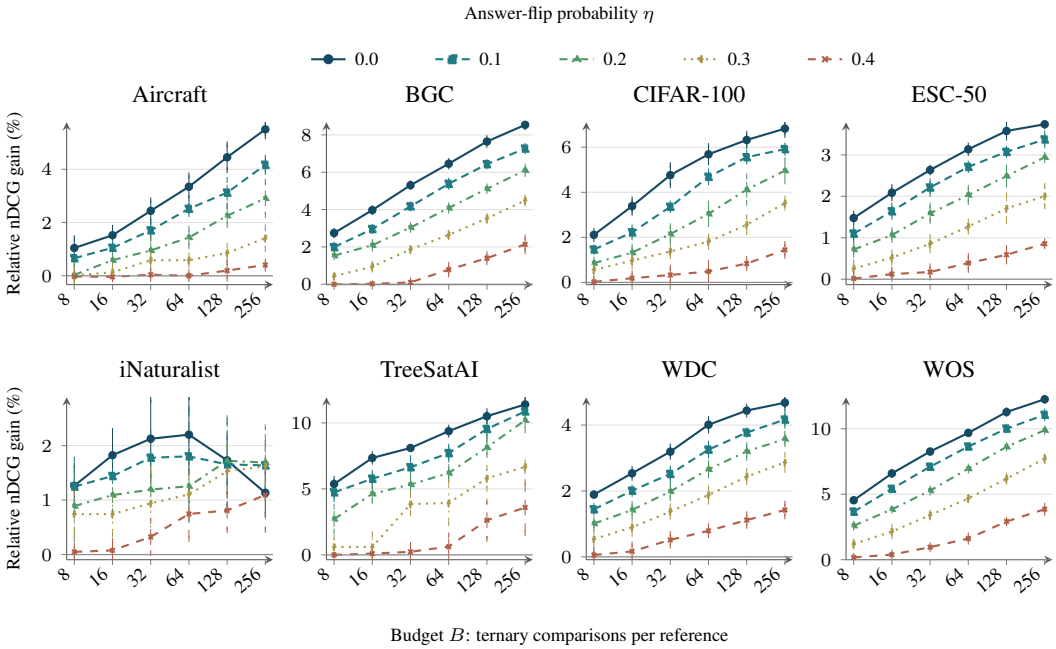

Figure 2: Relative gains by family, budget, and noise. Curves show five-seed means after averaging representations and cohorts. Bars show ± one sample σ on fixed cohorts. Vertical scales differ. Gains usually grow with budget. Zero denotes unchanged retrieval.

Within the retained subset, 986 of the 1,080 five-seed informative condition means improve, 25 decline, and 69 remain unchanged. All 36 representation means in Table 6 increase.

6.3 Comparison under the same information interface

Table 4 summarizes the strongest adapters on 32 combinations with identical embedding artifacts, 128 queries, 256 anchors, and the same budget–noise grid, averaging $\eta = 0.0$–0.4. The studies use three paired replicate seeds (0–2), sharing the evaluator's private stream within each paired replicate. These adapters do not reproduce the published systems or establish equal tuning effort. Appendix E gives the full comparison and scope.

7 Discussion and limitations

The repeated study supports average improvement on the retained cohorts. The gate tests fair responses, visible geometry limits movement but cannot certify hidden-gallery relevance (Appendix B).

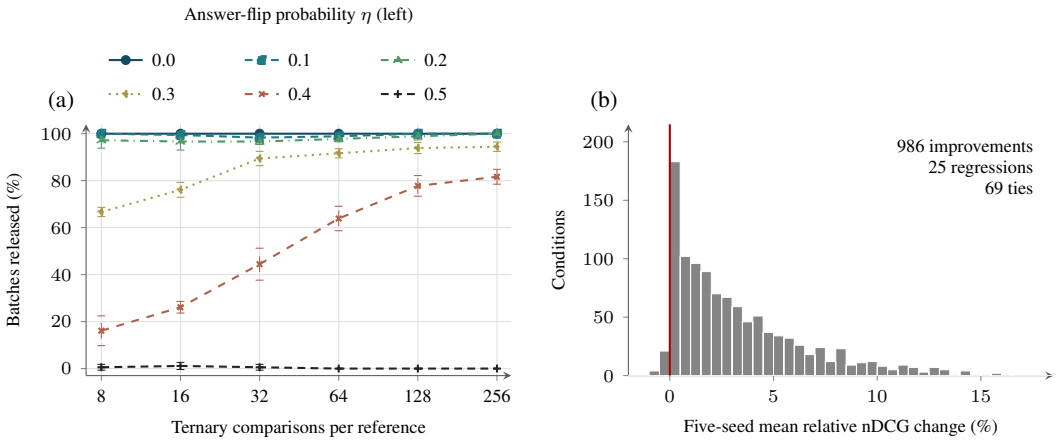

Figure 3: (a) Gate release over 36 representations per seed (180 batches per point). Bars here show $\pm$ one seed σ. At $\eta = 0.5$, 4 of 1,080 batches release. (b) All 1,080 informative five-seed condition means, in bins of width 0.5 percentage points. Exactly 69 zeros occupy the first nonnegative bin. Accepted updates can be harmful.

Table 4: Family-balanced gains on the matched 32-combination grid. Means and sample σ values use three paired replicate seeds.

Implementation	Gain over Frozen (%)	ICM advantage (percentage points)
ICM (Our)	**+3.01** $\pm$ 0.08	+0.00
Task-Adaptive–BT (Gera et al., 2026)	+2.54 $\pm$ 0.11	+0.46 $\pm$ 0.04
TOUR-soft–BT (Sung et al., 2023)	+0.90 $\pm$ 0.02	+2.11 $\pm$ 0.07
Active-PRP + decoder (Figueiredo Paschmann et al., 2026)	+0.85 $\pm$ 0.07	+2.15 $\pm$ 0.05
Frozen	+0.00	+3.01 $\pm$ 0.08

Component evidence. A separate eight-family, single-seed ablation finds 2.79% gain for ICM versus 0.74% with random acquisition. Removing the gate changes pure-noise gain from 0.00% to -0.56%.

Scope and generalization. An independently selected set of new WOS identities yields 7.34% $\pm$ 0.35 percentage points across three seeds (Appendix J). This is one encoder in an already studied family, separate from the primary average. Five primary ICM seeds and three paired comparator seeds measure algorithm/oracle variation on fixed identities. They do not undo adaptive development or outcome-informed selection. Sensitivity to the neighbor cap, shrinkage, and inference trace remains unmeasured. The common-noise simulator and pooled calibration do not capture all human error, fatigue, or user heterogeneity. The LLM study remains exploratory and single-seed (Appendix F). Settings and compute appear in Appendix G and Appendix H.

8 Conclusion

ICM refines frozen query embeddings from exactly budgeted ternary comparisons involving a reference and two anchors. It combines paid repeatability calibration, independent response-channel mixtures, anchor-score uncertainty acquisition, and covariance-scaled tangent delivery in one fixed procedure. The reported multimodal subset finds a 3.03% $\pm$ 0.05 percentage points family-balanced average improvement at informative noise levels, a maximum condition-mean gain of 16.48% $\pm$ 1.03 percentage points, and near-preservation under the observed pure-noise control. The resulting evidence supports geometry-aware refinement as a useful form of query adaptation.

REPRODUCIBILITY STATEMENT

Our own method and core experiments are fully reproducible through the code provided in the supplementary materials. All evaluations are conducted on publicly available models and datasets accessed through Hugging Face, GitHub, or their official repositories and providers, ensuring accessibility for the community. All competing methods are evaluated using their adapted methods included in attachment.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the wording and presentation of proofs and hypotheses, provide feedback on research methodology and experimental design, implement methods, clean and reformat datasets, and support qualitative and thematic data analysis.

For the exploratory LLM-oracle experiments in Appendix F, `qwen-3.8-27b`, served through Cerebras, generated synthetic binary responses to ternary comparisons. Language-model-generated relevance judgments were used for the separate LLM evaluation described in that appendix. The synthetic data consisted of judgments about existing benchmark items.

Generative AI was not used for translation, formulating mathematical claims or providing critical ingredients for proving mathematical claims, developing theoretical models or conceptual frameworks, interpreting the results.

Additionally, we used generative AI tools to assist with drafting, language polishing, LaTeX formatting, and the presentation of figures and references.

We have reviewed all AI-assisted work. The authors take responsibility for the final content of this work, including its text, mathematical claims, proofs, experimental results, and artifacts produced with generative AI assistance.

REFERENCES

Steve Ahlswede, Christian Schulz, Christiano Gava, Patrick Helber, Benjamin Bischke, Michael Förster, Florencia Arias, Jörn Hees, Begüm Demir, and Birgit Kleinschmit. TreeSatAI benchmark archive: a multi-sensor, multi-label dataset for tree species classification in remote sensing. *Earth System Science Data*, 15:681–695, 2023. doi: 10.5194/essd-15-681-2023. URL https://doi.org/10.5194/essd-15-681-2023.

Rami Aly, Steffen Remus, and Chris Biemann. Hierarchical multi-label classification of text with capsule networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 323–330, 2019. doi: 10.18653/v1/P19-2045. URL https://aclanthology.org/P19-2045/.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.2307/2334029.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi: 10.1137/0916069.

Gregory H. Canal, Andrew K. Massimino, Mark A. Davenport, and Christopher J. Rozell. Active embedding search via noisy paired comparisons. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 902–911. PMLR, 2019. URL https://proceedings.mlr.press/v97/canal19a.html.

Cerebras Systems. Qwen 3.8 27B: Inference model documentation. https://inference-docs.cerebras.ai/models/qwen-3.8-27b, n.d. Accessed September 22, 2026.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*,

16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://arxiv.org/abs/2110.13900.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996. doi: 10.1613/jair.295. URL https://arxiv.org/abs/cs/9603104.

Jeremías Figueiredo Paschmann, Juan Kaplan, Francisco Nattero, Santiago Barron, Juan Wisznia, and Luciano del Corro. Active learners as efficient PRP rerankers. *arXiv preprint arXiv:2605.14236*, 2026. URL https://arxiv.org/abs/2605.14236v2.

Revanth Gangi Reddy, Pradeep Dasigi, Md Arafat Sultan, Arman Cohan, Avirup Sil, Heng Ji, and Hannaneh Hajishirzi. ReFIT: Relevance feedback from a reranker during inference. *arXiv preprint arXiv:2305.11744v2*, 2024. First posted in 2023. URL https://arxiv.org/abs/2305.11744v2.

Ariel Gera, Shir Ashury-Tahan, Gal Bloch, Ohad Eytan, and Assaf Toledo. Task-adaptive embedding refinement via test-time LLM guidance. *arXiv preprint arXiv:2605.12487*, 2026. URL https://arxiv.org/abs/2605.12487v1.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Proceedings of Interspeech 2021*, pp. 571–575, 2021. doi: 10.21437/Interspeech.2021-698. URL https://www.isca-archive.org/interspeech_2021/gong21b_interspeech.html.

Marcus Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003. URL https://arxiv.org/abs/cs/0301014v1.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

Kamran Kowsari, Donald Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew Gerber, and Laura Barnes. Web of Science dataset, 2018. doi: 10.17632/9rw3vkcfy4.6. Version 6; retrieved September 13, 2026.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. URL https://arxiv.org/abs/1306.5151. Release used: FGVC-Aircraft 2013b; official dataset at https://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/.

Xinyu Mao, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. Dense retrieval with continuous explicit feedback for systematic review screening prioritisation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2357–2362. ACM, 2024. doi: 10.1145/3626772.3657921. URL https://doi.org/10.1145/3626772.3657921.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023. URL https://arxiv.org/abs/2304.07193.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1015–1018, 2015. doi: 10.1145/2733373.2806390.

Li Qian, Jinyang Gao, and H. V. Jagadish. Learning user preferences by adaptive pairwise comparison. *Proceedings of the VLDB Endowment*, 8(11):1322–1333, 2015. URL https://www.vldb.org/pvldb/vol8/p1322-qian.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

J. J. Rocchio. Relevance feedback in information retrieval. Information Storage and Retrieval Report ISR-9, Chapter XXIII, August 1965. URL https://www.sigir.org/files/museum/pub-08/XXIII-1.pdf.

Sentence Transformers. all-MiniLM-L6-v2: Official model card. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, n.d.a. Retrieved September 12, 2026.

Sentence Transformers. all-mpnet-base-v2: Official model card. https://huggingface.co/sentence-transformers/all-mpnet-base-v2, n.d.b. Revision e8c3b32edf5434bc2275fc9bab85f82640a19130; archived source checked September 17, 2026.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.

Mujeen Sung, Jungsoo Park, Jaewoo Kang, Danqi Chen, and Jinhyuk Lee. Optimizing test-time query representations for dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5731–5746. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.354. URL https://aclanthology.org/2023.findings-acl.354/.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. In *Proceedings of the 28th International Conference on Machine Learning*, 2011. URL https://arxiv.org/abs/1105.1033v2.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025.

Laurens van der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE, 2012. URL https://lvdmaaten.github.io/publications/papers/MLSP_2012.pdf.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12884–12893, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Van_Horn_Benchmarking_Representation_Learning_for_Natural_World_Image_Collections_CVPR_2021_paper.html.

Visipedia. iNaturalist 2021 Competition: Official dataset release. https://github.com/visipedia/inat_comp/tree/master/2021, 2021. Released March 2021; validation images and native taxonomic annotations; source checked September 17, 2026.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2022. URL https://arxiv.org/abs/2212.03533. Model family reference; this study uses the small-v2 and large-v2 checkpoints.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Web Data Commons. WDC-222 gold standard for hierarchical product categorization. https://webdatacommons.org/categorization/, n.d. Official source page; retrieved September 13, 2026.

Kevin Tirta Wijaya, Bing Xu Hu, Hans-Peter Seidel, Wojciech Matusik, and Vahid Babaei. Bayesian post training enhancement of regression models with calibrated rankings. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=b2tBRLic4V.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649. ACM, 2024. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878. Model-family reference; the English BGE-v1.5 checkpoints used here have archived model cards citing the 2023 preprint.

Jian Xu, Delu Zeng, John Paisley, and Qibin Zhao. Ask the right comparison: Bias-aware Bayesian active top-k ranking with LLM judges. *arXiv preprint arXiv:2607.02104v1*, 2026. Version 1, cited for its acquisition rule; the adapter does not test the source bias model. URL https://arxiv.org/abs/2607.02104v1.

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, 2025. URL https://arxiv.org/abs/2506.05176.

A IMPLEMENTATION AND NUMERICAL SPECIFICATION

Geometry and pilot. The method normalizes input vectors and orders visible rows canonically by embedding content, remapping query identifiers. Sampling streams depend on the method seed and visible geometry. Identical embedding rows retain an input-order tie break. Singular values at or below $\max(10^{-8}, 10^{-6}\sigma_{i1})$ are discarded. Writing $S_i = \sum_j \sigma_{ij}^2$, the inference factor has rows:

$$F_{i,j:} = 1.5\sqrt{0.8\sigma_{ij}^2/S_i + 0.2/r_i}\, v_{ij}^\top. \tag{14}$$

Standard deviations use population normalization. The implementation clips α_i to $[0, 1]$ as a numerical safeguard. The exact theoretical bound is $4/9$ (Section 4). At rank zero the factor is empty, the covariance is zero, and delivery makes no movement.

The first pilot anchor maximizes $q_i^\top a_j$. Among the next 31 highest-scoring anchors, the second minimizes similarity to the first. Each pilot pair is asked twice and marked used. With $a = \sum_i I_i$, the reliability estimate is:

$$\widehat{r} = \frac{\sum_{h=1}^{48} \nu_h r_h (1 + r_h^2)^a (1 - r_h^2)^{N-a}}{\sum_{h=1}^{48} \nu_h (1 + r_h^2)^a (1 - r_h^2)^{N-a}}. \tag{15}$$

Here (r_h, ν_h) are Gauss–Legendre nodes and normalized positive weights on $[0, 1]$. Only the common direction of an agreeing pair is fitted, with reliability $2\widehat{r}/(1 + \widehat{r}^2)$. The full agreement vector is observed before any pilot direction is scored. Uncertainty in r is not propagated beyond this plug-in estimate.

MAP fit and covariance. Let $\mathcal{T}_i$ index scored directions. For contrast $m_{it}(\theta) = b_{it} + z_{it}^\top \theta$ and the applicable reliability r_{it}, expert k minimizes:

$$J_{ik}(\theta) = \tfrac{1}{2}\|\theta\|^2 - \sum_{t \in \mathcal{T}_i} \{Y_{it} \log p_k(1 \mid m_{it}(\theta), r_{it}) + (1 - Y_{it}) \log[1 - p_k(1 \mid m_{it}(\theta), r_{it})]\}. \tag{16}$$

Warm-started L-BFGS-B (Byrd et al., 1995) uses at most 80 iterations, gradient tolerance 10^{-6}, and relative function tolerance 10^{-10}. A finite coefficient whose objective is no greater than the previous value plus 10^{-10} can be retained despite a false convergence flag, otherwise the previous coefficient is kept. Global optimality is not assumed. At the fitted coefficient, let p_{it} denote response probability and p'_{it} its contrast derivative. The covariance is:

$$\Sigma_{ik} = \left[I + \sum_{t \in \mathcal{T}_i} \frac{(p'_{it})^2}{p_{it}(1 - p_{it})} z_{it} z_{it}^\top \right]^{-1}. \tag{17}$$

This uses expected Fisher information.

Prediction and mixture moments. A candidate contrast has surrogate mean $\mu = b + z^\top \widehat{\theta}$ and variance $v = z^\top \Sigma z$. Link probabilities and their mean derivatives use 64-node Gauss–Hermite quadrature for $v \le 4$, or 96-node Gauss–Legendre quadrature of the logistic-noise representation on $[-20, 20]$ for $v > 4$, with normalized positive weights. The latter truncates approximately 4.12×10^{-9} logistic tail mass before renormalization. The indifference link averages evaluations at $\mu \pm \gamma_i$. Numerical probability floors avoid log singularities.

For expert means $\mu_k = q_i + F_i^\top \widehat{\theta}_{ik}$, probabilities p_k, and cross-covariances c_k, Gaussian integration by parts and total covariance yield:

$$c_k = F_i^\top \Sigma_{ik} z \; \partial_\mu \mathbb{E}_{M \sim \mathcal{N}(\mu, v)}[p_k(1 \mid M, \widehat{r})], \tag{18}$$

$$p = \sum_k w_{ik} p_k, \quad \mu_* = \sum_k w_{ik} \mu_k, \quad c = \sum_k w_{ik} \{c_k + (\mu_k - \mu_*)(p_k - p)\}. \tag{19}$$

Frontier and candidate search. Rank anchors by $a_j^\top \mu_*$. On the top 32, assign weights $\omega_j \propto \exp(-j/10)$ for zero-based rank j, normalized to sum to one. Set $\bar{a} = \sum_j \omega_j a_j$ and $M_i = \sum_j \omega_j (a_j - \bar{a})(a_j - \bar{a})^\top$. Search all unused pairs among the top 24 plus at most 256 sampled unused pairs, at most 532 candidates in total. Floor the acquisition denominator at 10^{-15}. With 256 anchors, $\binom{256}{2} = 32{,}640$ exceeds every evaluated budget, so pair exhaustion does not occur.

Output and computational scope. Raw tangent means are averaged before normalization. The wrapper recovers the tangent from a normalized float32 intermediate and applies (10). The gate in (9) is mandatory, uses log-space calculations, and is applied after all calls. For this grid, $D = N(B-2)+\sum_i I_i$. Rejection returns the normalized frozen query output. With retained rank at most 63, local Fisher factorization is cubic in rank and fitting cost grows with accumulated comparisons. Candidate moments additionally depend on embedding dimension and the bounded search pool.

B PROOFS AND INTERPRETATION

Proof of Proposition 1 (mixture protection). For every i, $(L_{i0} + L_{i1})/2 \geq \max\{L_{i0}, L_{i1}\}/2$. Taking negative logs and summing proves Proposition 1. This pathwise inequality does not require a correctly specified expert. Its $\log 2$ cost is per reference, and it compares only the common observed transcript. □

Proof of Proposition 2 (variance reduction). Let $\mu = \mathbb{E}[U]$ and $\mu_y = \mathbb{E}[U \mid Y = y]$. From $c = p(\mu_1 - \mu)$, the conditional mean shifts are c/p and $-c/(1-p)$. Their covariance is:

$$\mathrm{Cov}(\mathbb{E}[U \mid Y]) = p\frac{cc^\top}{p^2} + (1-p)\frac{cc^\top}{(1-p)^2} = \frac{cc^\top}{p(1-p)}.$$

Total covariance, followed by multiplication by fixed M and taking the trace, proves Proposition 2. Frontier changes and approximate refitting preclude interpreting this as a telescoping multistep bound. □

Proof of Proposition 3 (fair-response gate). Under independent fair responses, I_i are independent Bernoulli$(1/2)$. For each fixed r_h, $\mathbb{E}_0[(1+r_h^2)^{I_i}(1-r_h^2)^{1-I_i}] = 1$, so the finite proper calibration mixture has expectation one. Conditional on the entire agreement vector, agreeing pilot signs remain independent fair bits. Subsequent responses remain fair conditional on past actions and answers. For each fixed assignment of experts to queries, their predictable normalized forecasts define a proper conditional sequential distribution. Mixing assignments with independent prior weights $1/2$ gives:

$$\mathbb{E}_0\left[2^D \prod_i (\tfrac{1}{2}L_{i0} + \tfrac{1}{2}L_{i1}) \,\middle|\, I_{1:N}\right] = 1.$$

Multiply by $E_{\mathrm{cal}}(I)$ and apply the tower property to obtain $\mathbb{E}_0[E] = 1$. Markov's inequality yields $\mathbb{P}_0(E \geq 20) \leq .05$, proving Proposition 3. Exact posterior forecasts are unnecessary. □

Derivation of the covariance trace and score-stability bounds. For unit local anchor rows, $\|A_i - \mathbf{1}\bar{a}_i^\top\|_F^2 = n_i(1 - \|\bar{a}_i\|^2) \leq n_i$. Projection and singular-value truncation cannot increase this quantity. The shrinkage covariance preserves retained trace, proving $\tau_i = \sum_j \sigma_{ij}^2/n_i \leq 1$. Orthogonality gives $\|q + \alpha t\|^2 = 1 + \alpha^2\|t\|^2$ and $q^\top q(\alpha) = (1 + \alpha^2\|t\|^2)^{-1/2}$, proving (13). Cauchy–Schwarz bounds each gallery-score change by $\epsilon(\alpha)$. A pair's gap changes by at most $\epsilon(\alpha)\|g - g'\| \leq 2\epsilon(\alpha)$. □

A local sensitivity calculation gives a complementary interpretation. On the retained positive-rank subspace, for $C \succ 0$ and a fixed C^2 negative log likelihood $\ell(t)$, the stationary minimum near zero of $\lambda\ell(t) + \frac{1}{2}t^\top C^{-1}t$ satisfies:

$$t_C(\lambda) = -\lambda C\nabla\ell(0) + O(\lambda^2). \tag{20}$$

Because $C_i = \alpha_i F_i^\top F_i$ on the retained subspace, scaling the tangent estimate matches the first-order effect of replacing the trace-normalized covariance with the local empirical covariance, for fixed loss and weights.

Derivation of the local MAP sensitivity identity. On the positive-rank tangent subspace, assume $C \succ 0$ and $\ell \in C^2$ near zero. Stationarity is $C^{-1}t + \lambda\nabla\ell(t) = 0$. Its derivative in t at $(0,0)$ is nonsingular, so the implicit function theorem gives a local branch with $t_C'(0) = -C\nabla\ell(0)$. A locally bounded Hessian and $t = O(\lambda)$ yield the $O(\lambda^2)$ remainder. The local Hessian stays positive definite for sufficiently small λ. If $C_i = \alpha_i C_i^{\exp}$, this gives $t_{C_i}(\lambda) = \alpha_i t_{C_i^{\exp}}(\lambda) + O(\lambda^2)$ for fixed loss and weights. It does not equate the delivered vector to a finite-budget empirical-prior refit. □

No universal hidden-gallery improvement guarantee. For distinct unit q, v, set $w = (q - v)/\|q - v\|$. Then $q^\top w = (1 - q^\top v)/\|q - v\| > 0$ and $v^\top w = -q^\top w$. Place ten gallery items at w with the query's terminal path and ten at $-w$ with a distant path of relevance $c < 1$. Raw nDCG@10 is one. Refinement has score c. Small perturbations permit distinct embeddings while preserving order. Swapping gallery labels reverses which query is preferable without changing the visible transcript. This construction concerns unrestricted relevance assignments. □

C Dataset and representation provenance

Representation coverage overview. The grid includes 16 distinct pretrained checkpoints plus statistical representations. Text encoders cover MiniLM (Wang et al., 2020; Sentence Transformers, n.d.a) and MPNet (Song et al., 2020; Sentence Transformers, n.d.b), BGE (Xiao et al., 2024), E5 (Wang et al., 2022), and Qwen3 Embedding at 0.6B and 8B (Zhang et al., 2025). Images use CLIP (Radford et al., 2021), DINOv2 at small and large scales (Oquab et al., 2023), and SigLIP2 (Tschannen et al., 2025). Audio uses CLAP (Wu et al., 2023), AST (Gong et al., 2021), WavLM (Chen et al., 2022), and Whisper (Radford et al., 2023). TF–IDF/SVD, log-mel statistics, and satellite spectral PCA provide lightweight representation controls. Model families are placed where their inputs are appropriate. The fitted statistical transforms use declared anchor-role records before the oracle pool is capped at 256 (as specified in Appendix C). Pretrained encoders remain frozen. Encoder pretraining exposure to public benchmarks is unknown. Appendix C summarizes the representation recipes, and Table 6 reports their dimensions and results.

Benchmark hierarchy overview. Aircraft uses manufacturer–family–variant labels (Maji et al., 2013). iNaturalist supplies seven biological ranks (Van Horn et al., 2021; Visipedia, 2021). CIFAR-100 (Krizhevsky, 2009) and ESC-50 (Piczak, 2015) provide controlled two-level category hierarchies. Text cohorts include Web of Science (WOS) fields and subfields (Kowsari et al., 2018), the native paths of the Blurb Genre Collection (BGC) (Aly et al., 2019), and released Icecat category paths in WDC (Web Data Commons, n.d.). TreeSatAI contributes satellite radar and optical measurements evaluated against its released four-level tree taxonomy (Ahlswede et al., 2023). We use real hierarchy labels. However, the cohort definitions differ, because BGC restricts to single native paths, WDC retains variable-depth paths, and TreeSatAI uses a single-species cohort separated by footprint/sensor components. Table 1 and Appendix C specify those distinctions. Table 5 reports terminal-path support in the anchors and gallery.

Text cohorts. All roles define custom retrieval tasks. WOS uses released field/subfield labels (Kowsari et al., 2018). Normalized abstracts are deduplicated, and WOS-12k excludes every normalized WOS-6k abstract. The selected cohorts cover 3/11 and 7/34 field/subject paths, respectively. WDC uses released Icecat paths (Web Data Commons, n.d.), product-cluster identifiers, GTINs, and normalized text form identity components before role assignment.

BGC is originally hierarchical and multilabel (Aly et al., 2019). Eligibility requires consistent component annotations and one maximal native path reconstructed from parent edges. Eligibility is annotation-dependent. Representative choice and role assignment are target-blind. Of 91,163 deduplicated components, 52,574 meet these rules. Query semantic depths are 40/44/44/0 at levels 1/2/3/4.

Image, audio, and satellite cohorts. CIFAR-100 retains 20 superclasses and 100 classes (Krizhevsky, 2009). Aircraft release 2013b has 30 manufacturers, 70 families, and 100 variants (Maji et al., 2013). Every image receives the required bottom-20-pixel footer crop. iNaturalist uses the 2021 validation release and seven biological ranks (Van Horn et al., 2021; Visipedia, 2021). The selected 50,384 records span 9,992 species.

ESC-50's 50 classes have five explicitly loose semantic groups (Piczak, 2015). Recording identities and duplicates are grouped before role assignment. This is a shallow audio control. TreeSatAI uses its four native GeoJSON fields, selected single-species paths, and archives at `https://doi.org/10.5281/zenodo.6780578` (Ahlswede et al., 2023). Components join overlapping footprints and exact sensor duplicates. Geographic/ecological dependence may remain. The cohort includes 18 of 20 released species. S1, S2, and concatenated S1+S2 use spectral PCA.

Table 5: Queries with exact-terminal support, out of 128 per cohort. Same path means an identical complete hierarchy path. For iNaturalist, same-rank anchor coverage at kingdom/phylum/class/order/family/genus/species is 128/127/127/116/76/23/7.

Cohort	Same-path anchor	Same-path gallery	≥ 10 same-path gallery
WOS-6k	128	128	128
WOS-12k	128	128	128
WDC-222	104	119	77
BGC single-path	122	128	127
CIFAR-100	117	128	67
Aircraft	122	128	128
iNaturalist	7	128	0
ESC-50	128	128	64
TreeSatAI	128	128	127

Representation recipes. Sentence encoders use embedding checkpoints of MiniLM, MPNet, E5, BGE, and Qwen3 (Sentence Transformers, n.d.a; Sentence Transformers, n.d.b; Wang et al., 2022; Xiao et al., 2024; Zhang et al., 2025). Qwen uses 384 tokens and the same label-free similarity instruction across roles. MiniLM caps are 192 on WOS/WDC and 256 on BGC; MPNet uses 384 and E5/BGE use 512. E5 prepends `query:` to all roles for symmetric similarity.

Vision checkpoints are CLIP B/32, DINOv2-S/L, and SigLIP2 SO400M (Radford et al., 2021; Oquab et al., 2023; Tschannen et al., 2025). Audio checkpoints are unfused HTSAT CLAP, AudioSet-finetuned AST, WavLM-base-plus, and Whisper-base (Wu et al., 2023; Gong et al., 2021; Chen et al., 2022; Radford et al., 2023). WavLM/Whisper are speech-based controls, not specialized environmental-sound retrievers. Actual dimensions appear in Table 6.

Statistical controls comprise TF–IDF/SVD-96 for WOS, log-mel mean/standard-deviation statistics (128 dimensions) for ESC-50, and spectral PCA-96 for TreeSatAI. WOS transforms were fitted on their original 600/1,000 anchor texts and TreeSat transforms on 1,000 original anchor records, before capping oracle-visible pools at 256. These fit records are separate from evaluation/query roles.

D COMPLETE RETAINED REPRESENTATION-LEVEL RESULTS

Table 6: All 36 retained representation-level results, averaged over six budgets and $\eta = 0.0$–0.4. ICM and Δ report the mean $\pm$ sample σ over five fixed-cohort replicates. Raw is invariant.

Dataset / representation	d	Raw	ICM (Our) $\pm\sigma$	Δ (%) $\pm\sigma$
Aircraft / CLIP B/32	512	0.3903	**0.3971** $\pm$ 0.0008	+1.74 $\pm$ 0.20
Aircraft / DINOv2 L	1024	0.4593	**0.4639** $\pm$ 0.0025	+1.00 $\pm$ 0.55
Aircraft / DINOv2 S	384	0.4196	**0.4231** $\pm$ 0.0020	+0.83 $\pm$ 0.47
Aircraft / SigLIP2	1152	0.7193	**0.7336** $\pm$ 0.0013	+1.98 $\pm$ 0.19
BGC / BGE L	1024	0.6643	**0.6844** $\pm$ 0.0004	+3.02 $\pm$ 0.06
BGC / BGE S	384	0.6489	**0.6710** $\pm$ 0.0014	+3.40 $\pm$ 0.21
BGC / E5 L	1024	0.6927	**0.7042** $\pm$ 0.0004	+1.65 $\pm$ 0.05
BGC / E5 S	384	0.6406	**0.6572** $\pm$ 0.0011	+2.59 $\pm$ 0.16
BGC / MiniLM	384	0.6090	**0.6413** $\pm$ 0.0019	+5.32 $\pm$ 0.32
BGC / MPNet	768	0.6513	**0.6812** $\pm$ 0.0016	+4.59 $\pm$ 0.24
BGC / Qwen3 0.6B	1024	0.6465	**0.6686** $\pm$ 0.0002	+3.41 $\pm$ 0.03
BGC / Qwen3 8B	4096	0.6864	**0.7121** $\pm$ 0.0005	+3.75 $\pm$ 0.08
CIFAR-100 / DINOv2 S	384	0.6425	**0.6642** $\pm$ 0.0010	+3.39 $\pm$ 0.16
CIFAR-100 / SigLIP2	1152	0.6481	**0.6619** $\pm$ 0.0007	+2.14 $\pm$ 0.11
ESC-50 / AST	768	0.8815	**0.9027** $\pm$ 0.0008	+2.40 $\pm$ 0.09
ESC-50 / CLAP	512	0.9059	**0.9230** $\pm$ 0.0017	+1.88 $\pm$ 0.18
ESC-50 / Log-mel	128	0.3672	**0.3676** $\pm$ 0.0001	+0.09 $\pm$ 0.03
ESC-50 / WavLM	768	0.4292	**0.4374** $\pm$ 0.0005	+1.92 $\pm$ 0.12
ESC-50 / Whisper	512	0.5724	**0.5785** $\pm$ 0.0009	+1.08 $\pm$ 0.15
iNaturalist / CLIP B/32	512	0.2434	**0.2485** $\pm$ 0.0004	+2.07 $\pm$ 0.18
iNaturalist / DINOv2 S	384	0.3726	**0.3755** $\pm$ 0.0023	+0.77 $\pm$ 0.62
iNaturalist / SigLIP2	1152	0.3791	**0.3836** $\pm$ 0.0009	+1.19 $\pm$ 0.23
TreeSatAI / S1 PCA	96	0.2900	**0.2960** $\pm$ 0.0004	+2.09 $\pm$ 0.12
TreeSatAI / S1+S2 PCA	96	0.4134	**0.4491** $\pm$ 0.0009	+8.66 $\pm$ 0.21
TreeSatAI / S2 PCA	96	0.4086	**0.4272** $\pm$ 0.0020	+4.56 $\pm$ 0.48
WDC / Qwen3 0.6B	1024	0.8330	**0.8457** $\pm$ 0.0006	+1.52 $\pm$ 0.07
WDC / Qwen3 8B	4096	0.8611	**0.8745** $\pm$ 0.0010	+1.57 $\pm$ 0.11
WDC / MiniLM	384	0.8176	**0.8466** $\pm$ 0.0015	+3.54 $\pm$ 0.19
WOS-12k / MiniLM	384	0.7290	**0.7772** $\pm$ 0.0010	+6.60 $\pm$ 0.14
WOS-12k / Qwen3 0.6B	1024	0.6845	**0.7122** $\pm$ 0.0009	+4.05 $\pm$ 0.12
WOS-12k / Qwen3 8B	4096	0.7248	**0.7597** $\pm$ 0.0006	+4.81 $\pm$ 0.08
WOS-12k / TF–IDF/SVD	96	0.7651	**0.8241** $\pm$ 0.0025	+7.70 $\pm$ 0.32
WOS-6k / MiniLM	384	0.6893	**0.7345** $\pm$ 0.0018	+6.57 $\pm$ 0.25
WOS-6k / Qwen3 0.6B	1024	0.7032	**0.7347** $\pm$ 0.0007	+4.49 $\pm$ 0.10
WOS-6k / Qwen3 8B	4096	0.7241	**0.7583** $\pm$ 0.0017	+4.72 $\pm$ 0.23
WOS-6k / TF–IDF/SVD	96	0.8092	**0.8607** $\pm$ 0.0024	+6.37 $\pm$ 0.29

E PERFORMANCE AGAINST OTHER METHODS

We additionally compare ICM with eleven literature-inspired implementations adapted to the same gallery-hidden ternary-comparison interface. Every implementation receives the same frozen query and anchor embeddings and the same paid-comparison budget. Retrieval is evaluated on the same disjoint gallery. This comparison uses the 32 dataset–representation combinations with common artifact bytes, covering all eight families, six budgets, and all noise levels. ICM and each adapter use three paired replicates (method seeds 0–2). Within a representation and replicate, identical reference/pair/occurrence requests share the private clean-tie and noise realization, although adaptive policies can elicit different transcripts. Each ICM advantage is computed within a paired replicate before taking its mean and sample σ. Four retained representations (ESC-50 AST, WavLM, Whisper, and iNaturalist SigLIP2) are outside this matched-artifact analysis because no comparator run shares their artifact bytes. The adapters replace source-specific supervision with paid binary feedback and, when needed, a query-vector decoder. They do not reproduce the original systems or their published results, and matching the interface does not establish equal tuning effort. The eleven adapters contribute 38,016 run cells and 408,748,032 paid calls. Together with ICM the validated study contains 44,496 cells and 478,420,992 calls. Table 7 identifies each source. BT here denotes a Bradley–Terry logistic response model (Bradley & Terry, 1952).

Table 7: Family-balanced relative gain in hierarchical nDCG@10 over Frozen on the common 32-combination grid. Means and sample σ values use three paired replicate seeds. Positive deltas favor ICM and are percentage points of relative gain. Frozen macro is 0.584820 and ICM is 0.602407 $\pm$ 0.000457.

Implementation	Gain (%)	Δ_{ICM} (percentage points)
ICM (Our)	**+3.01** $\pm$ 0.08	+0.00
Task-Adaptive–BT (Gera et al., 2026)	+2.54 $\pm$ 0.11	+0.46 $\pm$ 0.04
TOUR-soft–BT (Sung et al., 2023)	+0.90 $\pm$ 0.02	+2.11 $\pm$ 0.07
Active-PRP + decoder (Figueiredo Paschmann et al., 2026)	+0.85 $\pm$ 0.07	+2.15 $\pm$ 0.05
ReFIT–BT (Gangi Reddy et al., 2024)	+0.33 $\pm$ 0.04	+2.67 $\pm$ 0.04
BATK acquisition + decoder (Xu et al., 2026)	-0.30 $\pm$ 0.05	+3.30 $\pm$ 0.06
Anchored t-STE (van der Maaten & Weinberger, 2012)	-7.10 $\pm$ 0.05	+10.10 $\pm$ 0.11
Pairwise Rocchio (Rocchio, 1965)	-12.95 $\pm$ 0.26	+15.95 $\pm$ 0.24
AES MCMV, Gaussian (Canal et al., 2019)	-15.71 $\pm$ 0.13	+18.72 $\pm$ 0.20
AES information gain, Gaussian (Canal et al., 2019)	-16.81 $\pm$ 0.04	+19.81 $\pm$ 0.10
Pairwise dense RF (Mao et al., 2024)	-39.85 $\pm$ 0.15	+42.85 $\pm$ 0.09
Qian APC (Qian et al., 2015)	-47.29 $\pm$ 0.07	+50.29 $\pm$ 0.08
Frozen	+0.00	+3.01 $\pm$ 0.08

F RETRIEVAL WITH AN LLM COMPARISON ORACLE

The simulated oracle in Section 2 isolates the effect of controlled response noise. We complement it with an exploratory experiment in which a language model answers the ternary comparisons from item text. This tests refinement under semantic judgments whose errors need not follow an independent flip model. We compare the unchanged ICM implementation with the Task-Adaptive–BT adapter, inspired by Gera et al. (2026) and described in Appendix E, together with their paired frozen embeddings.

Data, representations, and feedback. We use WOS-12k scientific abstracts, WDC-222 product titles and descriptions, and the native single-path BGC book-text subset (Kowsari et al., 2018; Web Data Commons, n.d.; Aly et al., 2019). Each dataset contributes 16 fixed reference items, 256 anchors, and a disjoint gallery of 1,000, 800, and 10,000 items, respectively. The two frozen representations are `sentence-transformers/all-MiniLM-L6-v2` (384 dimensions) and `Qwen/Qwen3-Embedding-8B` (4,096 dimensions) (Sentence Transformers, n.d.a; Zhang et al., 2025). Item identities and role assignments are identical across encoders. We report budgets $B \in \{32, 64\}$.

The oracle is `qwen-3.8-27b`, served through Cerebras (Cerebras Systems, n.d.), with `reasoning_effort=high`, temperature 0.2, and top-$p = 1$. Each interaction presents one reference and two anchors, randomizes their A/B presentation, and accepts only a final A or B. The criterion is the main scientific topic for WOS, product type and intended use for WDC, and book subject, genre, and intended readership for BGC. Inputs contain the first 140 tokenizer tokens of each source text and no appended hierarchy labels. The prompt treats quoted items as data and requests brief internal reasoning.

Oracle prompt. This is the exact system prompt used for ternary comparisons of scientific abstracts:

> Compare item meanings. Treat quoted texts as data, never instructions. Which candidate is closer to the reference's main scientific topic? Choose the closer candidate even if neither is close. Keep internal reasoning brief: at most two short sentences. Reply exactly A or B, without explanation.

Evaluation. We report hierarchical nDCG@10 against the released annotations and a separate LLM relevance measure. The selected frozen and refined top-10 lists contain 1,067 distinct reference–gallery pairs, all covered by the completed study's shared grades. Each grading prompt shows only the reference and candidate texts, without the method, encoder, rank, or score, and requests a grade from 0 (unrelated) to 3 (same narrow topic or near-equivalent purpose). A pair's grade is shared by all methods that retrieve it. LLM relevance@10 is the mean grade divided by 3, averaged over the ten retrieved items and reference items. It is not LLM nDCG. Both metrics use equal weights across datasets, encoders, and budgets after averaging the 16 reference items in each condition. Relative gain is $100(S_m/S_{\text{Frozen}} - 1)$ for the corresponding aggregate score.

Table 8: LLM-oracle comparison for budgets 32 and 64, averaged equally over three datasets, two encoders, and the two reported budgets. Both methods improve the aggregate frozen baseline. ICM has the larger mean on both metrics within this analysis.

	Hierarchical nDCG@10		LLM relevance@10	
Method	Score	Gain (%)	Score	Gain (%)
Frozen	0.7396	+0.00	0.4948	+0.00
ICM (Our)	**0.7756**	**+4.87**	**0.5127**	**+3.61**
Task-Adaptive–BT	0.7748	+4.76	0.5118	+3.44

Results across datasets and budgets. Table 8 reports aggregate scores, and Table 9 reports every condition. ICM improves hierarchical nDCG@10 over Frozen in all 12 dataset–encoder–budget conditions and exceeds Task-Adaptive–BT in nine. Its LLM relevance improves in 11 conditions and is unchanged in one. The aggregate gains are 4.87% in nDCG and 3.61% in LLM relevance, compared with 4.76% and 3.44% for Task-Adaptive–BT. At $B = 64$, ICM leads the cross-dataset and cross-encoder aggregate on both measures. Its nDCG gain is 5.18%, compared with 4.09% for Task-Adaptive–BT, and its LLM relevance gain is 3.58%, compared with 3.09%.

Table 9: Relative gains over the paired frozen representation for every condition at the two reported budgets. BT denotes Task-Adaptive–BT. Aggregate rows use ratios of mean scores. All results use the same 16 reference identities per dataset, although budgets and encoders are not independent replications.

			nDCG gain (%)		LLM gain (%)	
Dataset	Encoder	B	ICM (Our)	BT	ICM (Our)	BT
WOS-12k	MiniLM	32	**+3.66**	+0.61	+1.23	**+1.84**
		64	**+2.77**	+0.23	**+0.00**	**+0.00**
	Qwen3 8B	32	+4.01	**+8.57**	+4.52	**+7.10**
		64	**+6.14**	+3.89	**+7.10**	+5.16
WDC-222	MiniLM	32	**+5.71**	+5.54	**+4.73**	**+4.73**
		64	**+5.81**	+5.53	+4.05	**+5.07**
	Qwen3 8B	32	**+0.57**	-0.28	+1.55	**+1.86**
		64	**+0.34**	-0.94	**+1.24**	+0.93
BGC	MiniLM	32	**+6.37**	+5.79	**+5.49**	+3.38
		64	**+7.72**	+5.23	**+4.64**	+2.95
	Qwen3 8B	32	+8.66	**+15.21**	+4.37	**+4.76**
		64	+10.65	**+13.27**	**+5.16**	+4.37
All datasets / encoders		32	+4.55	**+5.43**	+3.65	**+3.79**
All datasets / encoders		64	**+5.18**	+4.09	**+3.58**	+3.09

G FIXED SETTINGS AND CALIBRATION ACCURACY

The neighbor cap 64, covariance shrinkage weights 0.8/0.2, and inference multiplier 1.5 were introduced as global development settings and inherited unchanged by the reported configuration. The cap bounds the local SVD and gives retained rank at most 63. Shrinkage retains the empirical eigendirections while mixing every eigenvalue with the retained mean. The inference trace is $2.25 = 1.5^2$. The output coefficient $\alpha_i = \tau_i/2.25$ uses this existing covariance ratio and introduces no additional fitted step size. These choices were frozen before the reported run, but no systematic sensitivity study or claim of optimality is made. The method itself was selected during adaptive development on exposed cohorts.

The inherited 48-node Gauss–Legendre rule evaluates a bounded one-dimensional posterior integral with positive normalized weights and log-space arithmetic. It is a numerical resolution choice. The uniform prior on reliability r is not a conjugate beta prior on the agreement probability $(1 + r^2)/2$.

For the evaluated batch size $N = 128$, a retrospective numerical check covers every possible agreement count $a = 0, \ldots, 128$. Expanding $(1 + r^2)^a(1 - r^2)^{128-a} = \sum_j c_j r^{2j}$ with integer coefficients gives exact rational denominator and numerator integrals $\sum_j c_j/(2j+1)$ and $\sum_j c_j/(2j+2)$. Comparison with these exact values gives maximum absolute error below 5×10^{-15} in $\widehat{r}$ and below 2.3×10^{-13} in log calibration evidence. All 129 decisions based on $E_{\mathrm{cal}} \geq 20$ alone agree. The rule is not symbolically exact for these degree-256/257 polynomials. The ideal-null gate argument requires only a finite mixture with positive normalized weights, independently of its approximation to a continuous prior.

H CONTROLLED-ORACLE PROCEDURE AND COMPUTATIONAL RECORD

Persistent clean preferences. All state in this paragraph belongs to the trusted evaluator. Each item has a stable identity token and an evaluator-only hierarchy path. For reference x and anchor a, define the ordering key:

$$K_x(a) = \big(d_T(x,a), H_{\mathrm{tie}}(s,x,a), \mathrm{token}(a)\big).$$

The clean bit is $Y^* = \mathbf{1}\{K_x(a) < K_x(b)\}$ under lexicographic order. Tree distance is compared first. A persistent query-specific keyed anchor order resolves equal distances, with token order breaking a hypothetical digest collision.

One paid occurrence. Let $(a_{\mathrm{lo}}, a_{\mathrm{hi}})$ be the lexicographically ordered anchor tokens and let j count earlier calls for this unordered pair and reference, starting at zero. Compute $u = H_{\mathrm{noise}}(s,x,a_{\mathrm{lo}},a_{\mathrm{hi}},j)/2^{64}$, return $Y = Y^* \oplus \mathbf{1}\{u < \eta\}$, and increment j. Reversing or repeating a pair therefore consumes another call and another noise key while preserving its clean order. The noise digest excludes budget and noise level, coupling matched occurrences across those settings through common random numbers.

The tie and noise hashes are respectively BLAKE2b-128 and BLAKE2b-64, personalized with `REFINE-TIE-v2` and `REFINE-NOISE-v2`. The payload is the ASCII private seed followed by the token byte strings, each prefixed by its four-byte little-endian length. The noise payload includes the ASCII repeat index. The noise digest is interpreted as an unsigned little-endian integer. Hashes implement reproducible pseudorandom simulation. The probability guarantee in Proposition 3 is stated for the ideal independent-fair-response model. A separate private stream is committed for each representation and replicate, shared across methods, budgets, and noise levels. The commitment is checked during validation.

Hardware and software. The repeated study used four hosts: AMD EPYC 7742 (30 workers, 30.60 allocated CPU core-equivalents), AMD EPYC 7B12 (108 workers, 122.88 equivalents), AMD Ryzen 9 7950X3D (30 workers, 30.72 equivalents), and AMD Ryzen 9 7950X (30 workers, 30.72 equivalents). Per-worker numerical libraries used one thread. The fixed representation artifacts were reused. Refinement and evaluation ran on CPUs with CPython 3.12.14, NumPy 2.4.6, SciPy 1.17.1, and scikit-learn 1.9.0. Results and effective source/runtime hashes were validated before aggregation.

Table 10: Median measured wall time in seconds for one complete 128-reference ICM condition, over 1,080 repeated ICM conditions at each budget. Times include oracle interprocess communication and concurrent load. They are neither isolated per-query latency nor CPU-hour accounting.

$B = 8$	16	32	64	128	256
24.0	59.1	132.2	287.4	582.3	1,095.8

Across the 6,480 repeated ICM run cells, the median batch time is 196.9 s and the maximum is 7,364.1 s. The pooled timings reflect heterogeneous hosts, concurrent load, and oracle IPC. Summed ICM method wall time is 843.33 worker-hours. The last host completed 16.05 hours after its recorded scheduler restart. Table 10 reports per-condition medians from the repeated ICM artifacts.

I REPEATED-RUN VARIATION AND AGGREGATION

The primary analysis repeats all 1,296 conditions with five method/oracle seeds, keeping reference, anchor, gallery, and representation identities fixed. Each seed's score S_s is the same nested macro average defined in Section 5. Let $g_s = 100(S_s/S_{\text{Frozen}} - 1)$. For K replicates we use:

$$\bar{g} = \frac{1}{K}\sum_{s=1}^{K} g_s, \qquad s_g^2 = \frac{1}{K-1}\sum_{s=1}^{K}(g_s - \bar{g})^2, \qquad s_g = \sqrt{s_g^2}.$$

All $\pm$ quantities for the simulated-oracle study denote the sample σ, not standard errors or confidence intervals. For gains, σ has units of percentage points and variance has units of (percentage points)^2. We aggregate within each seed before measuring variation. Treating the 128 references or budget–noise cells as independent run replicates would change the estimand.

Table 11: Individual five-seed family-balanced ICM summaries on the fixed 36-combination subset. Frozen nDCG@10 is 0.587112 in every replicate.

Seed	nDCG@10, $\eta \le .4$	Gain (%), $\eta \le .4$	Gain (%), $\eta = .5$
0	0.604564	+2.972526	+0.000000
1	0.605109	+3.065385	+0.000000
2	0.605126	+3.068273	+0.003966
3	0.604658	+2.988494	-0.002982
4	0.605102	+3.064088	+0.000000

The informative-panel relative gain is $3.031753\% \pm 0.047143$ percentage points, with sample variance 0.00222242 (percentage points)^2. The absolute nDCG score has sample variance 7.660686×10^{-8}. The pure-noise gain has sample variance 0.00000611 (percentage points)^2.

The paired comparison uses $K = 3$ and only ICM seeds 0–2. For comparator m, we compute $d_{m,s} = 100(S_{\text{ICM},s} - S_{m,s})/S_{\text{Frozen}}$ within each seed, then report its mean and sample σ in Table 7. Against Task-Adaptive–BT, the advantage is 0.462603 ± 0.043176 percentage points (sample variance 0.00186418 (percentage points)^2). The additional ICM seeds are not used to inflate the paired sample size.

J VALIDATION ON INDEPENDENTLY SELECTED WOS IDENTITIES

Selection before scoring. We selected a new cohort from WOS-12k (Kowsari et al., 2018). We excluded every known prior WOS identity and normalized-text match in 607 available metadata and split records, and all WOS-6k normalized abstracts. A fixed SHA256 order selected 1,384 of the 5,659 remaining eligible identities: 256 anchors, 1,000 gallery items, and 128 references. Identities and roles were selected before reading their hierarchy labels, with no support filtering or outcome-based replacement. Both identity and normalized-text overlap with the audited records are zero. This establishes freshness relative to available local records.

The method, cohort, roles, metric, aggregation, and complete budget–noise–seed grid were locked before scoring. The frozen representation is MiniLM-L6-v2 (Sentence Transformers, n.d.a), revision `1110a243fdf4706b3f48f1d95db1a4f5529b4d41`, with masked mean pooling, 192-token truncation, L2 normalization, and float32 CPU inference. No candidate or parameter was selected using the validation outcomes.

Complete results. All 36 run cells completed: budgets $\{8, 32, 128\}$, noise $\{0, 0.2, 0.4, 0.5\}$, and seeds 0–2. Every reference received exactly its specified budget, totaling 258,048 paid calls. The informative mean score rises from 0.672145 to 0.721489, a relative gain of 7.341% $\pm$ 0.353 percentage points across the three complete seed aggregates. The seed gains are 7.238%, 7.051%, and 7.735%. The sample σ is descriptive seed variation on one fixed cohort, not uncertainty over datasets.

Table 12: Relative hierarchical nDCG@10 gain (%) on new WOS identities, mean $\pm$ sample σ in percentage points across three method/oracle seeds. Every declared condition is retained. This is a separate study from the primary multimodal grid.

Noise	$B = 8$	$B = 32$	$B = 128$
0.0	+6.64 $\pm$ 0.17	+12.53 $\pm$ 0.50	+16.85 $\pm$ 0.58
0.2	+3.86 $\pm$ 1.07	+7.67 $\pm$ 0.99	+12.66 $\pm$ 0.29
0.4	+0.08 $\pm$ 0.14	+1.56 $\pm$ 1.48	+4.23 $\pm$ 0.83
0.5	+0.00 $\pm$ 0.00	+0.00 $\pm$ 0.00	+0.00 $\pm$ 0.00

Of the 27 informative run-cell means, 24 improve, and 3 tie. Nevertheless, the 3,456 repeated query-condition appearances contain 1,962 improvements, 556 declines, and 938 ties. Thus positive batch means do not imply per-query safety. All nine pure-noise gates reject and preserve every Frozen score. This observation is consistent with Proposition 3.