Arc Science
Technical Paper

Cammuteur-R1 and Library-1M: Calibrated Bilingual Memory Retrieval, Multimodal Funnel Indexing, and Machine-Checked Routing

Formal Ranking Geometry, Two-Fold Out-of-Fold Calibration, and Verified Policy Invariants Across Local Hardware

Danil Kotelnikov

Abstract. Personal memory archives combine bilingual notes and scientific figures in growing stores. Naive text scanning exceeds latency limits, leaks raw text past exit gates, and produces uncalibrated scores. We present Cammuteur-R1 and the Library-1M architecture. First, a write-once binary index uses SHA-256 deduplication, int8 Matryoshka IVF rooms, CSR postings, and pre-tokenized uint16 shards so warm queries incur zero raw-body reads. Second, equivariant and cross-encoder scoring uses margin-preserving losses, and image queries route to vision-language encoders. Third, two-fold out-of-fold logistic calibration maps raw logits to posterior probabilities for the deterministic ten-branch PickV3 policy. Fourth, 56 Lean 4 core theorems verify ranking geometry, keyed masking, PickV3 safety, stateful warm-lock invariants, and supersession chain ordering without custom axioms. Across 1,000,000 entries (65,839 real cards plus 934,161 synthetic distractors), hybrid search reaches 0.7432 widened Hit@5 at 163.30 ms median latency. On the same million-entry index, the visual slice reaches 0.9167 Recall@5.

Contents

1 Problem Setting and Memory Scale **3**

2 Write-Once Cascade Architecture **3**

2.1 Backbone Selection and Model Currency 3

2.2 Million-Scale Funnel and Body Isolation 4

3 Loss Geometry and Policy Calibration **5**

3.1 Equivariant Scoring and Ranking Objectives 5

3.2 Out-of-Fold Calibration and Decision Records 6

3.3 Deterministic Ten-Branch PickV3 Policy 6

4 Machine-Checked Lean Invariants **6**

5 Dataset Audit and Benchmark Design **8**

5.1 Private Bilingual Note Benchmark Audit 8

5.2 Bilingual Scientific Figure Benchmark 8

5.3 Public Science Decision Benchmark 8

5.4 Million-Scale Synthetic Distractor Protocol 9

6 Empirical Baselines **9**

6.1 Text Reranking and Calibration Baselines 9

6.2 Exit Guard Boundary Audits 9

6.3 Scientific Figure Routing Baselines 10

6.4 Public Science Calibration Baselines 10

7 Million-Scale Library Evaluation **11**

7.1 Write-Once Ingest and Memory Footprint 11

7.2 Million-Scale Text and Funnel Recall 11

7.3 Million-Scale Multimodal Slice 13

8 Claim Boundaries and Reproducibility **13**

1. Problem Setting and Memory Scale

Research agents store mixed records across sessions. In the Cammut memory store inside Arc Harness, one archive holds 65,807 bilingual Russian and English notes beside scientific figures. Growing this store toward one million entries creates three bottlenecks when lookup relies on linear scanning.

First, brute-force scanning over raw text or floating-point vectors exceeds interactive latency budgets on laptops. Second, reading every raw body during candidate search exposes sensitive text past exit boundaries and raises disk I/O. Third, raw cross-encoder logits are poorly calibrated probabilities [1]. Downstream controllers therefore cannot tell confident matches apart from lists needing wider search.

Existing retrieval tools solve individual pieces without linking bilingual scoring and formal routing invariants. Lexical BM25 delivers fast keyword matching [2], yet cross-lingual Russian-English paraphrases give zero lexical overlap. Vector toolkits such as Faiss combine hierarchical navigable small world graphs [3] and product quantization [4] at scale [5]. Late-interaction engines like ColBERTv2 and PLAID compress token vectors 6x to 10x and prune centroids [6, 7]. Matryoshka embeddings similarly nest coarse prefixes inside one vector [8].

None of those components joins bilingual cross-encoding, figure routing, out-of-fold calibration, exit redaction, and verified policy safety. Cammuteur-R1 and Library-1M assemble and verify this full chain. Section 2 specifies the write-once binary library and multimodal funnel. Section 3 formalizes ranking objectives, logistic calibration, and decision packets. Section 4 presents machine-checked Lean 4 invariants. Section 5 details dataset audits and benchmarks. Section 6 reports baseline measurements across text, visual, and public benchmarks. Section 7 evaluates million-scale indexing and funnel recall. Section 8 states claim boundaries.

2. Write-Once Cascade Architecture

2.1. Backbone Selection and Model Currency

Selecting models for an 8 GB laptop GPU balances bilingual coverage, license terms, and architectural currency. We verify every model choice against the Arc Harness registry ledger. Legacy checkpoints serve as historical baselines or cached storage artifacts, and active defaults use current models.

Table 1 separates active models from historical comparators. In text encoding, `ModernBERT` (2024, legacy English precursor superseded by `mmBERT` and `RuModernBERT`) added rotary positions and 8,192-token context [13]. English pretraining misses Russian morphology and mixed-script notes. Cammuteur-R1 therefore uses `deepvk/RuModernBERT-base`, pretrained on two trillion Russian, English, and code tokens [9].

Cached bank embeddings on disk still use `BGE-M3` (2024, legacy bi-encoder retained strictly to avoid re-embedding stored legacy files) [12]. Earlier distillation runs also read teacher logits from `Qwen3-Reranker` (2025, legacy reranker superseded by `Nemotron-3.5-Rerank-8B` and `Querit-4B`)

Table 1: Model inventory and role assignment across text, multimodal, and decision layers. Every legacy model lists its year and reason.

Model	Status	License	Role and Currency Reason
RuModernBERT-base	Current	Apache-2.0	Primary bilingual RU/EN cross-encoder backbone (150M, 8,192 context) [9]
mmBERT-base	Current	MIT	Multilingual ModernBERT variant covering 1,800 languages (307M total) [10]
Qwen3-VL-Embedding-2B	Current	Apache-2.0	Primary multimodal embedding stage (30+ languages, MRL prefixes) [11]
Qwen3-VL-Reranker-2B	Current	Apache-2.0	Optional top-8 multimodal cross-attention reranker [11]
Claude 5.5 / Gemini 4	Current	Commercial	Cross-family builder and checker models in Harness^2 bench
BGE-M3	Legacy (2024)	MIT	Cached 2024 bank vectors kept only to avoid re-embedding legacy store [12]
ModernBERT-base	Legacy (2024)	Apache-2.0	Superseded English 2024 precursor cited for architecture lineage [13]
Qwen3-Reranker-4B	Legacy (2025)	Apache-2.0	Prior 2025 teacher checkpoint superseded by Nemotron-3.5 and Querit-4B [14]
BGE-VL-base	Legacy (2024)	MIT	English-trained 2024 CLIP baseline measuring cross-lingual figure recall drop [15]
ColPali	Legacy (2024)	Apache-2.0 / Gemma	Founding 2024 late-interaction VLM baseline superseded by Qwen3-VL [16]
Nemotron-ColEmbed-V2	Current	CC-BY-NC-4.0	Top ViDoRe V3 reference excluded from default path by non-commercial license [17]

[14]. Across multimodal baselines, `BGE-VL-base` (2024, legacy English CLIP retriever tested for cross-lingual drop) has no Russian training data [15].

Visual model `ColPali` (2024, legacy model) is superseded by `Nemotron-ColEmbed-V2` and `Qwen3-VL-Embedding` due to multi-vector storage growth [16]. While `Nemotron-ColEmbed-V2` leads ViDoRe V3 at 63.42 nDCG@10 in 2026, its CC-BY-NC-4.0 license blocks default local use [17]. Cammuteur-R1 instead pairs Apache-2.0 `Qwen3-VL-Embedding-2B` with `Qwen3-VL-Reranker-2B` for bilingual visual recall [11].

2.2. Million-Scale Funnel and Body Isolation

Inside `library_1m.py`, `WarmLibrary` and `ColdRawStore` isolate binary shard lookup from raw note text. Once ingest finishes, `warm_locked` raises an assertion error if any query touches `ColdRawStore`. Execution splits into write-once ingest and two warm retrieval stages:

1. **Stage 0 (Write-Once Ingest).** `ingest_real_cards` hashes canonical card content with SHA-256 and skips embedding or tokenization on cached hashes. `finalize_library_disk` writes contiguous Matryoshka-256 `int8` IVF rooms (`emb_q8.bin`, `ivf_rooms.npz`) [8], memory-mapped CSR bilingual BM25 arrays (`bm25_csr`), pre-tokenized `uint16` IDs (`tok_mmap`), and columnar metadata (`meta`). It also links dated contradictions at ingest time through `link_supersessions_matrix`.
2. **Stage 1 (Coarse Funnel to 200).** `stage1_candidates` probes `nprobe=24` IVF rooms (`search_ivf_rooms`) alongside CSR BM25 postings (`search_csr_bm25`) and visual embeddings.

It chases supersession links for unpinned queries and outputs 200 integer indices with `raw_body_reads_warm == 0`.

3. **Stage 2 (Pre-Tokenized Reranking).** `rerank_stage2` narrows 200 candidates to `stage2_k=24` finalists and invokes `score_pretokenized_r1` directly on `tok_mmap` `uint16` tokens. `route_candidates_mode` separates text lists from figure sets, while supersession rules demote stale revisions unless the query pins an old date.
4. **Bounded Widening.** Function `evaluate_296_questions` widens Stage 1 when confidence falls. It sets `nprobe=72`, `pool_k=400`, and `stage2_k=64` while reusing `scored_cache`.

Pre-tokenizing `uint16` surrogate IDs keeps `raw_body_reads_warm == 0` across Stage 1 and Stage 2 without touching raw bodies on disk.

3. Loss Geometry and Policy Calibration

3.1. Equivariant Scoring and Ranking Objectives

Typical cross-encoders score each pair (q, d_i) separately through a linear layer. Independent scoring ignores list peers and leaves query logit offsets unaligned. In `train_r1.py`, `EquivariantHead` adapts Deep Sets [18] for list context. For candidate hidden states $\{h_1, \ldots, h_n\} \subset \mathbb{R}^d$, each item joins $[h_i, \bar{h}]$ where $\bar{h} = \frac{1}{n}\sum_{j=1}^{n} h_j$. A shared feed-forward block ($2d \to d \to \text{GELU} \to \text{LayerNorm} \to 1$) emits scores invariant to list order permutations.

Controlled English MS MARCO studies across 27 backbone-objective runs ranked contrastive listwise (`InfoNCE`) and pairwise margin (`MarginMSE`) highest, while binary cross-entropy ranked weakest [19]. Inside `train_r1.py`, four objectives train against clipped teacher log-odds $y \in [-\tau_c, \tau_c]^n$:

$$\mathcal{L}_{\text{mse}}(s, y) = \frac{1}{n}\sum_{i=1}^{n}(s_i - y_i)^2 \tag{1}$$

$$\mathcal{L}_{\text{pair}}(s, y) = \sum_{i=1}^{n}\sum_{j=1}^{n}\big((s_i - s_j) - (y_i - y_j)\big)^2 \tag{2}$$

$$\mathcal{L}_{\text{top-pair}}(s, y) = \sum_{i \neq j} w_{ij}^{\text{RR}}\big((s_i - s_j) - (y_i - y_j)\big)^2 \tag{3}$$

$$\mathcal{L}_{\text{anchor}}(s, y) = \mathcal{L}_{\text{pair}}(s, y) + \lambda_{\text{anc}}\big(\bar{s} - \bar{y}\big)^2 + \mathcal{L}_{\text{lst}}(s, y) \tag{4}$$

Here w_{ij}^{RR} scales pairwise terms by reciprocal teacher rank to block top-5 swaps. The listwise term $\mathcal{L}_{\text{lst}}$ matches softmax probabilities against scaled teacher targets. Section 4 shows that $\mathcal{L}_{\text{pair}}$ ignores uniform shifts $s \mapsto s + c$, while $\mathcal{L}_{\text{anchor}}$ pins mean list level without altering gaps.

3.2. Out-of-Fold Calibration and Decision Records

Neural cross-encoders require post-hoc calibration before threshold decisions [1]. Both `decide.py` and `multimodal_r1.py` fit two-fold out-of-fold logistic calibration (Platt scaling [20]) split by query index parity `int(qid[-3:]) % 2`.

For candidate list x, `feats` extracts four scale features.

$$\phi_{\text{list}}(x) = \big[s_{(1)},\ s_{(1)} - s_{(2)},\ s_{(1)} - s_{(-1)},\ |x|\big]. \tag{5}$$

Similarly, `note_feats` computes $[s_{(r)}, r, s_{(r)} - s_{(r+1)}]$ over the top five positions. Each fold standardizes inputs and fits logistic regression ($C = 1.0$) on the opposite half. That model predicts held-out $\hat{p} = P(\text{hit@5} \mid \phi_{\text{list}}(x))$. On the fitting half, τ_f is the lowest threshold reaching ≥ 0.90 precision (`TARGET = 0.90`). A lower floor `LOW = 0.50` triggers wider search below it.

We measure calibration with Expected Calibration Error (ECE), Brier score [21], and Murphy's Brier Skill Score (BSS) [22].

$$\text{BSS} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(\bar{y}_{\text{fold}} - y_i)^2}. \tag{6}$$

Here $\bar{y}_{\text{fold}}$ denotes the fold hit base rate. Each query writes a `cammut.decision/1` packet holding list score `questions[0].p` and note scores `picks[r].p_right`. The packet logs attention sets (`read_first`, `read_with_care`) and five exit policies without raw text. An automated `guard` masks `secret_like` patterns and public IPv4 strings while blocking external `private_note` transfers.

3.3. Deterministic Ten-Branch PickV3 Policy

From calibrated $\hat{p}$ and eleven state flags, `pick.py` evaluates the ten-branch `PickV3` rule. That tree selects one of nine actions: `Stop`, `Say cannot`, `Ask`, `Search wider`, `Check`, `Answer`, `Plan`, `Act`, or `Learn`.

1. **Halt gates:** `refused_exit` forces `Stop`. Owner refusal or three failed retries forces `Say cannot`.
2. **Consent gates:** Irreversible operations or `high` risk without explicit owner consent trigger `Ask`.
3. **Evidence gates:** Missing facts or `case_match` failure ($\hat{p} < 0.50$) triggers `Search wider` first and `Ask` if already widened.
4. **Action gates:** Uncalibrated lists or scores below τ_f route to `Check`. Calibrated reversible passes select `Answer`, `Plan`, `Act`, or `Learn`.

This policy converts uncertain retrieval into bounded wider search before answering. That progression matches the Propose-Probe-Compose structure of Harness^2 [23].

4. Machine-Checked Lean Invariants

We verified ranking losses, keyed tokenization, `PickV3` safety, and `Library1MSection` invariants inside `CammuteurR1.lean` using Lean 4.34.1 [24]. Following audited proof hygiene [25], all 56

theorems compile against Lean 4 core without `sorry`, `Mathlib`, or custom axioms. Running `lean-gate.mjs` writes `cammuteur-r1-library-lean-gate-2026-11-15.json` with `state="pass"`, `audited_theorems=56`, zero token failures, zero compile errors, and zero axiom failures. A negative control inserting `sorry` at line 967 of a scratch copy exits with code 1 and reports `state="fail_token"`.

Table 2: Audited theorem groups in `CammuteurR1.lean`. Parts 1–14 verify 56 theorems with zero custom axioms (`cammuteur-r1-library-lean-gate-2026-11-15.json`).

Part	Key Theorem Symbols	Verified Mathematical Property
Parts 1–2	`order_of_strictMono`, `dcg_exchange`, `dcg_swap_cost`	Strict monotone maps preserve order; swapping misordered pairs reduces DCG.
Part 3	`zscore_gap`, `zscore_order`, `double_zscore_gap`	Affine normalization preserves order and centers list residuals identically.
Part 4	`logp_gap_squeezed`, `logp_gap_nonneg`	Clipped logit margins stay bounded while preserving non-negative rank gaps.
Part 5	`order_unsafe_pointwise`, `order_safe_pairwise`, `pair_loss_zero_iff_shift`	Pointwise MSE admits order flips; pairwise loss vanishes iff logits shift uniformly.
Part 6	`softmax_shift`, `zscore_is_temperature`	Softmax is shift-invariant; z-score scale matches inverse temperature.
Part 7	`bce_proper`, `bce_strict`, `anchor_pins_level`	Cross-entropy is strictly proper; mean anchoring pins absolute logit level.
Parts 8–9	`cipher_reader_is_blind`, `keyed_tokenizer_exact`, `keyed_tokenizer_keeps_counts`	Keyed masking blinds plaintext; keyed tokenizers keep exact token counts.
Part 10	`list_head_equivariant`	Mean-pooled set heads stay permutation-equivariant under all permutations S_n.
Parts 11–12	`pair_eq_centered_mse`, `top_weighted_pair_safe`	Pairwise MSE equals $2n$ times centered MSE; weighted loss blocks top swaps.
Part 13	`pickV3_stop_wins`, `pickV3_no_act_on_high_without_consent`, `pickV1_broken_no_undo`, `pickV2_broken_stalls_on_yes`	Exhaustive safety over all Boolean inputs for `PickV3` plus counterexamples breaking `PickV1`/`PickV2`.
Part 14	`content_hash_zero_reread`, `reingest_idempotent_state`, `Pin`, `supersession_chain_strict_order`, `cold_read_rejected_when_locked`, `warm_trace_preserves_zero_reads`	Stateful warm-lock read rejection, idempotent digest re-ingest, pin preservation, newest-first chain ordering, and monotone widening recall.

Three formal results guided the implementation. First, `order_unsafe_pointwise` shows that small pointwise MSE errors can invert candidate rankings, whereas `order_safe_pairwise` forbids rank inversions below half the squared margin. Second, `pair_eq_centered_mse` shows pairwise margin MSE equals $2n$ times centered MSE. Unanchored pairwise training can therefore drift globally until `anchor_pins_level` binds absolute logit scale. Third, Part 14 (`Library1MSection`) verifies idempotent deduplication (`content_hash_zero_reread`, `reingest_idempotent_state`), pin preservation (`pin_preserves_original_score`), newest-first chain ordering (`supersession_flips_stale_rank`, `supersession_chain_strict_order`), widening recall monotonicity (`funnel_recall_monotone_widen`), and stateful warm-lock

zero-read invariants (`cold_read_rejected_when_locked`, `warm_trace_preserves_zero_reads`, `naive_scan_reads_every_note`).

5. Dataset Audit and Benchmark Design

We separate retrieval benchmarks into four suites so private bank text never leaks across public reports.

5.1. Private Bilingual Note Benchmark Audit

The private text suite (`s0_dataset_audit_2026-11-15.json`) spans 296 queries over 65,807 real notes. Queries split evenly across Russian (148) and English (148). Five functional classes enter the benchmark: `plain` (90), `mixed` script (60), `two`-fact (60), `dated` (60), and `update` (26). Integrity checks find zero missing identifiers and a mean gold lexical recall of 0.332.

Auditing 85 failure rows sets the measured audit boundaries listed in Table 3. Six questions (`s0-002`, `s0-035`, `s0-075`, `s0-195`, `s0-211`, `s0-296`, 7.1%) have defective gold targets. Fifty-three misses retrieved valid unmarked notes (30 full, 23 partial). We therefore report scores over all 296 queries and over the 239-query clean subset.

Table 3: Private `s0` dataset audit summary across 296 queries and 65,807 real bank notes.

Audit Slice	Count	Measured Audit Boundary
Bank corpus (`bank_notes`)	65,807	Bilingual RU/EN real archive; raw bodies excluded from public output.
Query distribution	296	148 RU, 148 EN; classes: plain 90, mixed 60, two 60, dated 60, update 26.
Integrity checks	0 missing IDs	37 high lexical overlap queries; mean gold token recall 0.332.
Audited failure rows	85 judged	6 defective questions (7.1%); 53 valid unmarked bank hits (30 full, 23 part).
Clean manifest tiers	239 / 51 / 6	239 clean items, 51 multi-valid caution items, 6 flagged items.

5.2. Bilingual Scientific Figure Benchmark

To test multimodal routing without private images, `mm_sci_bench_2026-11-15.json` gathers 32 public scientific PNG figures. That set includes 24 target plates and 8 hard distractors. It adds no synthetic text distractors. Forty-eight bilingual queries (24 EN, 24 RU) run against those figures. Categories cover quantitative plots (20), visual structures (14), and mechanistic workflows (14). A separate 40-item text slice (`s0_clean_40`) tests pure-text routing invariance.

5.3. Public Science Decision Benchmark

We verify out-of-fold calibration and Propose-Probe-Compose routing on public items using `public-science-bench-2026-11-15.json`. That benchmark runs 24 multiple-choice questions evenly split between `EperLuo/BaisBench` and `Yibo-Wen/as-bench`. Eight model arms evaluate those items.

5.4. Million-Scale Synthetic Distractor Protocol

Library-1M tests one-million-vector indexing ($N = 1,000,000$) on laptop hardware. The index combines 65,839 real cards (65,807 real notes and 32 real figures; 62,343 unique SHA-256 digests) with 934,161 bilingual **synthetic distractors**. Total content reaches 996,504 unique SHA-256 hashes. Every million-scale table tags the 934,161-item shard as synthetic distractors, and accuracy is measured on real benchmark queries only.

6. Empirical Baselines

6.1. Text Reranking and Calibration Baselines

Table 4 summarizes text retrieval, INT8 quantization retention, and out-of-fold calibration results across 296 private queries prior to Library-1M indexing.

Table 4: Measured text retrieval, quantization, and calibration baselines across 296 queries (`cammuteur-r1-bench-2026-11-14.json` and `gr1_13.json`).

Model and Precision	Hit@5	Dated	RU / EN	p50 / p95	Gate and Calibration Result
`s0_run_base` (dense)	0.7128	—	—	—	Base score; GD.1/2 `fail_no_signal` (AUC 0.6613)
`s0_run_live` (ONNX)	0.8142	—	—	472 / 950 ms	CPU student; GD.1 AUC 0.8083, GD.2 gain +0.2764
`small_mse` (FP32)	0.7399	—	—	571 / 972 ms	`gr1_13 fail_no_gain` (-0.0743 vs live); AUC 0.8066
`small_mse` (INT8)	0.6858	0.5333	0.608 / 0.764	346 / 812 ms	-0.0541 drop vs FP32 under onednn INT8
`base_mse` (FP32, 512)	0.7973	—	—	3,561 / 5,997 ms	Cand AUC 0.8962, cal ECE 0.0074, BSS 0.1951
`base_mse` (INT8, CPU)	0.7804	0.6833	0.797 / 0.764	1,216 / 2,131 ms	Non-significant -0.0169 shift vs FP32
`base_mse` (GPU BF16)	—	—	—	121 / 232 ms	RTX 4060 Laptop BF16 runtime
`base_three` (FP32, 512)	0.8108	0.7500	0.824 / 0.797	—	Cand AUC 0.8898, cal ECE 0.0073, BSS 0.1891 ($n = 6618$)

Three measured patterns emerge from Table 4. First, Platt calibration reduces logit overconfidence. On `base_three`, candidate ECE drops from 0.5344 to 0.0073 (Brier 0.4614 to 0.0335; BSS 0.1891 across 6,618 pairs). Second, widening candidate lists to 20 notes lifts live Hit@5 from 0.7128 to 0.8142. Both `GD.1` and `GD.2` gates pass on `lab20`. Answered queries reach 0.8889 Hit@5 across 73.0% of requests (+0.2764 gain). Third, `EquivariantHead` keeps permutation variance bounded (`max_perm_diff` $\leq 3.6 \times 10^{-7}$ for sizes 16 to 80) below 0.46 ms p50.

6.2. Exit Guard Boundary Audits

On exit gate `GD.4` (`gd4.json`), the guard blocks 10 of 10 known leak replay cases and refuses 2 of 2 private external transfers. Over 2,000 sampled real notes, regex scanning logs 6 `address` flags and 0 `secret_like` alarms (0.0030 upper-bound false-alarm share). However, 7 of 8 unseen

secret patterns slip past regex checks. Format isolation via `raw_body_reads == 0` therefore remains essential alongside output filtering.

6.3. Scientific Figure Routing Baselines

Table 5 reports multimodal retrieval across 48 bilingual scientific figure queries. Lexical BM25 drops to 0.2374 nDCG@10 on Russian figure queries versus 0.7576 on English. Likewise, `BGE-VL-base` (2024, legacy English CLIP baseline evaluated for cross-lingual drop) reaches only 0.2257 nDCG@10 on Russian figures. By contrast, Apache-2.0 `Qwen3-VL-Embedding-2B` attains 1.0000 nDCG@10 inside 4.37 GB VRAM on local laptop GPUs.

Two-fold Platt fusion (`cammuteur_r1_multimodal`) achieves 0.9923 nDCG@10 (BSS 0.8106, cal ECE 0.0461 across 1,536 pairs). Paired bootstrap comparisons give +0.1299 nDCG@10 over `RuModernBERT-base` ([0.0636, 0.2040]) and +0.5567 over `BGE-VL-base` ([0.4400, 0.6692]). Furthermore, 40 of 40 text-only queries (`s0_clean_40`) route to the text branch, with zero text regression (+0.0122 nDCG@10, [-0.0454, +0.0683]).

Table 5: Measured retrieval and calibration results on 48 figure queries over 32 real PNGs (`gr1_multimodal_2026-11-15.json`).

System	R@1	R@5	nDCG@10	EN / RU	Infer (s)	VRAM and Calibration
`bm25_archive_stub`	0.3542	0.5625	0.4975	0.758 / 0.237	0.003	CPU lexical baseline
`text_cammuteur_r1_stub`	0.7708	0.9583	0.8624	0.893 / 0.832	7.62	720 MB; RuModernBERT
`bge_vl_base` (2024, legacy)	0.2708	0.5208	0.4356	0.646 / 0.226	1.49	394 MB; EN CLIP bias
`qwen3_vl_emb_2b_mm`	1.0000	1.0000	1.0000	1.000 / 1.000	15.06	4,371 MB; Apache-2.0
`qwen3_vl_rerank_2b_top8`	1.0000	1.0000	1.0000	1.000 / 1.000	105.07	4,221 MB; top-8 rerank
`cammuteur_r1_multimodal`	0.9792	1.0000	0.9923	1.000 / 0.985	—	BSS 0.8106, ECE 0.0461

6.4. Public Science Calibration Baselines

On `public-science-bench-2026-11-15.json` ($6.7833 spend across 24 items, 0 blocked), cross-family Propose-Probe-Compose (`hoh-ppc-sonnet-carbon`) reaches top accuracy at 0.9167 (22/24, $3.8772, 150.5 s mean latency). Single-shot `Gemini 4 Argon`, `Carbon`, `Hydrogen`, and `Claude Haiku 5.5` each score 0.8750 (21/24). Single-shot `Claude Sonnet 5.5` scores 0.7917, `Claude Opus 5.5` scores 0.6667, and `Claude Fable 5.2` scores 0.6250.

Single-shot `Claude Fable 5.2` and `Gemini 4 Hydrogen` retain positive two-fold Brier skill scores (0.5475 and 0.2410). Pooled multi-step `hoh-ppc` is overconfident on this small sample (OOF BSS -0.6130, $n = 24$). Confidence cuts from multi-agent consensus therefore need dedicated calibration before use.

7. Million-Scale Library Evaluation

We run `library_1m.py --bench-1m` to build both the persistent real store and the temporary one-million-card stress index (`gr1_library_1m_2026-11-15.json`). Unit checks in `library_1m.py --selftest` pass before measurement.

7.1. Write-Once Ingest and Memory Footprint

On the persistent real collection (`library_catalog_real`, 65,839 real cards), ingest completes in 25.70 s (2,562.2 cards/s) and occupies 40.661 MB across 256 IVF rooms. Deduplication skips 500 duplicate card identifiers and reuses 3,496 matching SHA-256 digests. The store indexes 30,855 BM25 terms (1,871,768 postings), 4,034,903 `uint16` pre-tokenized tokens, and 22,541 supersession links.

Adding 934,161 tagged **synthetic distractors** brings total scale to 1,000,000 cards in 43.66 s (22,904.6 cards/s; 17.96 s for distractor shard generation). The full one-million-card store occupies 461.3 MB across 512 IVF rooms, 965,045 BM25 terms (14,366,172 postings), and 22,718,123 `uint16` tokens. Cold ingest incurs 65,839 raw body reads, while all warm queries log `raw_body_reads_warm == 0`. Temporary million-scale shard files are deleted after benchmarking (`temp_1m_shards_cleaned = true`).

7.2. Million-Scale Text and Funnel Recall

Table 6 summarizes funnel recall, calibrated decision routing, and warm query latency at $N = 1,000,000$ entries. The offline 1M environment ran without local weights for `BGE-M3` (2024, legacy bi-encoder superseded by `Querit-4B` and `Horizon`). Therefore, `hybrid_mrl256_bm25_r1` forms query vectors from unlabeled top-8 quantized Matryoshka-256 neighbor centroids with zero gold access. We pair it with `bm25_live_r1_ablation`, a control that uses pure live BM25 Stage-1 lookup.

Hybrid Matryoshka-256 plus BM25 Stage 1 retains 0.8818 recall at candidate depth 200 across one million entries, compared to 0.6520 for live BM25 alone. Stage 2 `RuModernBERT-base` rescoring reaches 0.7399 narrow Hit@5 and 0.7432 widened Hit@5 at 155.52 ms narrow median latency (213.98 ms p95). Every query produces a valid `cammut.decision/1` packet with zero validation errors. Two-fold out-of-fold Platt scaling achieves 0.264027 Brier skill score (0.141649 OOF Brier against 0.192465 climatology) and 0.076631 ECE.

Triggering `search_wider` on 35 of 296 uncertain hybrid queries (11.82%) lifts Hit@5 on those triggered queries from 0.2857 to 0.3143 (+2.86 percentage points). Under the live-only BM25 ablation, `search_wider` triggers on 99 of 296 queries (33.45%) and raises triggered Hit@5 from 0.0606 to 0.1414 (+8.08 percentage points). Widening raises p95 latency from 213.98 ms to 257.14 ms while expanding Stage-1 candidate pools monotonically.

Table 7 breaks down million-scale Hit@5 across query slices. Mixed-script queries reach 0.9667 Hit@5 in the hybrid arm, while Russian recall stands at 0.7027, against 0.7838 for English after widening. On the 26 contradicting `update` queries, hybrid Hit@5 and `gold_beats_stale` both measure 0.5000, compared to 0.4231 for widened live BM25.

Table 6: Measured Library-1M retrieval, widening, latency, and two-fold OOF calibration across 65,839 real cards plus 934,161 tagged **synthetic distractors** (`gr1_library_1m_2026-11-15.json`).

System Arm (N = 1,000,000)	S1@200	S2@24	Top-1	Top-5	p50 / p95 (ms)	Conditions and Calibration
`hybrid_mrl256_bm25_r1` (narrow)	0.8818	0.8074	0.4392	0.7399	155.52 / 213.98	296q; top-8 MRL-256 centroid + RuModernBERT; BSS 0.2640, ECE 0.0766
`hybrid_mrl256_bm25_r1` (widen)	—	—	0.4426	0.7432	163.30 / 257.14	35/296 widened (0.2857 to 0.3143); 0 packet errors
`bm25_live_r1_ablation` (narrow)	0.6520	0.5574	0.3345	0.5507	89.03 / 142.61	296q live BM25 only; BSS 0.4506, ECE 0.1355
`bm25_live_r1_ablation` (widen)	—	—	0.3581	0.5777	109.83 / 173.64	99/296 widened (0.0606 to 0.1414); 0 packet errors
`visual_slice_mm_sci_bench`	1.0000	0.9375	0.5833	0.9167	70.94 / 88.21	48q / 32 PNGs in 1M; `BGE-VL-base` (2024 legacy snapshot); nDCG@10 0.7684

Table 7: Million-scale Hit@5 slice breakdown across 296 real queries inside the 1,000,000-entry index (65,839 real cards plus 934,161 tagged **synthetic distractors**).

Query Slice (1M)	Count	Hybrid Narrow	Hybrid Widen	BM25 Narrow	BM25 Widen
`plain`	90	0.7778	0.7889	0.4333	0.4778
`dated`	60	0.5167	0.5167	0.3833	0.3833
`two` (two-fact)	60	0.7833	0.7833	0.5833	0.6167
`mixed` (bilingual script)	60	0.9667	0.9667	0.9167	0.9500
`update` (Hit@5)	26	0.5000	0.5000	0.4231	0.4231
`update` (gold beats stale)	26	0.5000	0.5000	0.4231	0.4231
Russian (`ru`)	148	0.7027	0.7027	0.3176	0.3378
English (`en`)	148	0.7770	0.7838	0.7838	0.8176

7.3. Million-Scale Multimodal Slice

The offline 1M environment ran without local weights for `Qwen3-VL-Embedding-2B`. The visual slice therefore evaluates a cached snapshot of `BGE-VL-base` (2024, legacy CLIP fallback; 1,445.72 ms load, 1,676.16 ms infer). Inside the 1,000,000-card index, all 48 figure queries route to the multimodal branch. Stage 1 retains 1.0000 recall at depth 200, and Stage 2 retains 0.9375 at depth 24. Final Recall@5 reaches 0.9167 (0.5833 Recall@1, 0.7130 MRR@10, 0.7684 nDCG@10) at 70.94 ms p50 and 88.21 ms p95 latency. Comparing this 0.7684 nDCG@10 fallback against the 1.0000 standalone score of `Qwen3-VL-Embedding-2B` quantifies the Russian visual gap on local hardware.

8. Claim Boundaries and Reproducibility

The claims carry six limits. First, third-party comparisons rely on verified abstracts, model cards, and source documentation, and external checkpoints are used as released. Second, the scientific figure benchmark covers 48 queries across 32 real PNGs and 8 distractors. Larger visual collections with thousands of similar charts may exhibit lower top-1 separation.

Third, the audit of 85 misses in the 296-query text benchmark finds six defective questions (7.1%) and 53 misses that retrieved valid unmarked bank notes. Reporting both 296-query raw Hit@5 and 239-query clean scores avoids hidden label inflation. Fourth, million-scale indexing benchmarks add 934,161 structured bilingual **synthetic distractors** to 65,839 real cards. Synthetic distractors stress index storage, speed, and quantization noise. Natural topic overlap among one million human notes lies outside this test.

Fifth, the offline 1M environment ran without raw weights for `BGE-M3` (2024, legacy bi-encoder) and `Qwen3-VL-Embedding-2B`. Therefore, `hybrid_mrl256_bm25_r1` pairs unlabeled top-8 Matryoshka centroids with the live `bm25_live_r1_ablation` control, and the 1M visual slice runs the local snapshot of `BGE-VL-base` (2024, legacy CLIP fallback). Sixth, pre-tokenizing `uint16` shards into `tok_mmap` keeps `raw_body_reads_warm == 0` across Stage 1 and Stage 2. Regex exit guards still miss 7 of 8 unseen secret patterns. All local benchmarks, Lean 4 proofs, and public reports reproduce deterministically on laptop hardware without exposing private notes to outside APIs.

References

[1] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

[2] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

[3] Yu. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020. doi: 10.1109/TPAMI.2018.2889473. Preprint arXiv:1603.09320 (2016).

[4] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011. doi: 10.1109/TPAMI.2010.57.

[5] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library, 2024.

[6] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction, 2021.

[7] Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. PLAID: An efficient engine for late interaction retrieval, 2022.

[8] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning, 2022.

[9] DeepVK. RuModernBERT and USER2: Bidirectional encoders for russian, 2025. URL `https://huggingface.co/deepvk/RuModernBERT-base`.

[10] Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. mmBERT: A modern multilingual encoder with annealed language learning, 2025.

[11] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026.

[12] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[13] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024.

[14] 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.

[15] Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. MegaPairs: Massive data synthesis for universal multimodal retrieval, 2024.

[16] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models, 2024.

[17] Gabriel de Souza P. Moreira, Ronay Ak, Mengyao Xu, Oliver Holworthy, Benedikt Schifferer, Zhiding Yu, Yauhen Babakhin, Radek Osmulski, Jiarui Cai, Ryan Chesler, Bo Liu, and Even Oldridge. Nemotron ColEmbed V2: Top-performing late interaction embedding models for visual document retrieval, 2026.

[18] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets, 2017.

[19] Victor Morand, Mathias Vast, Basile Van Cooten, Laure Soulier, Josiane Mothe, and Benjamin Piwowarski. Building better encoder-only cross-encoders: A controlled study of training strategies for neural re-ranking, 2026.

[20] John C. Platt. Probabilities for SV machines. In *Advances in Large-Margin Classifiers*, pages 61–74. The MIT Press, 2000. doi: 10.7551/mitpress/1113.003.0008.

[21] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[22] Allan H. Murphy. A sample skill score for probability forecasts. *Monthly Weather Review*, 102(1):48–55, 1974. doi: 10.1175/1520-0493(1974)102<0048:ASSSFP>2.0.CO;2.

[23] Ruiyao Xu, Yanfei Chen, Zhongying CuiZhu, Bhavana Dalvi Mishra, Yifei Ming, Han Yu, Rujun Han, Chen-Yu Lee, and Tomas Pfister. Harness^2: Recursive agent harnessing for an open world, 2026.

[24] Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*, pages 625–635, 2021. doi: 10.1007/978-3-030-79876-5_37.

[25] Google DeepMind. Scaling verified systems with Lean 4 autoformalization, 2026. URL `https://google-deepmind.github.io/code-verification/`.