

STRATA: Certified Stratified Hubness for Compressed Vector Search

DRAFT v0 (2026-07-24) — not for distribution

Andrew H. Bond
San Jose State University
San Jose, CA, USA
andrew.bond@sjsu.edu

Abstract

Hubness—the emergence of nearest-neighbor “hubs” that appear in hundreds of top- k lists while “anti-hubs” appear in almost none—is a measured property of high-dimensional retrieval, and its damage under index compression lands precisely where aggregate recall cannot see it: on the minority of queries whose true neighbors are anti-hubs. We present STRATA, a system that makes hubness measurement *stratified* and *certified*: the corpus is partitioned by a content-addressed area map (tqp-area-map/1); every instrument reports per-area statistics whose verdicts aggregate as *min-over-strata*; thin strata *ABSTAIN* rather than pass by silence; verdict causes come from a closed registry; and the relational surface—hub census and transit fraction as SQL over a k NN edge table—*refuses* to join relations computed under different area-map digests. The components are deliberately commodity (SQL over vectors, IVF/LOPQ-style partitioning, sixteen years of hubness estimators); the claimed composition is the certification: columns that know their estimator, their minimum-evidence bounds, and their area-map identity, with refusal in the schema. We evaluate STRATA by running its own pre-declared gates and typesetting failures like passes. The mechanism classifier’s confusion-matrix fixtures killed two simpler designs before one survived. A four-rung measure-ladder shows CSLS hubness correction works at the exact layer (count skew -20% , Robin Hood index -30%) but is *refuted* in compressed-domain scoring and refuted *identically* at reconstruction rerank—diagnosing the bottleneck as candidate coverage, then score precision—while reranking stored originals at depth 501 passes the 0.90 bar in every stratum: the codes find, the originals decide. A fragility-driven bit allocator *lowered* the measured floor (fragile-first greedy); a max-min allocator with protected donors raised min-over-strata anti-hub recall from 0.7251 to 0.7697 at an equal 3.0 bits/row budget, and exposed cross-area score interference as a real allocation mechanism. Finally, we show stratified measurement is also a security instrument: per-area centering has an adversarial dual, and a per-area mechanism shift from density to centrality is a poisoning signature.

Keywords

vector search, hubness, quantization, certification, stratified measurement, SQL, threat model

1 Introduction

A compressed vector index that reports recall@10 = 0.98 can be failing one in five of its hardest queries. The mechanism is hubness: in the dimensions embeddings actually occupy, a few corpus rows (*hubs*) appear in hundreds of top- k lists while a large fraction (*anti-hubs*) appear in almost none [16]. Queries whose true nearest neighbors are anti-hubs are the hardest retrieval cases an index has—the signal separating the right answer from the crowd is smallest—and quantization rounds fine distinctions away first, so compression damage lands disproportionately on exactly those queries while the aggregate mean stays green. *Trust the tail, not the mean* is the operating doctrine this system inherits from its parent project’s acceptance-metric work.

This paper adds a second clause: *trust the tail, not the mean—in every area*. A global anti-hub recall number has the same blindness one level up: it can hide a collapsed stratum (a language, a tenant, a cluster) behind healthy neighbors. Global staleness has the dual problem: if any mutation anywhere invalidates every guarantee, guarantees become unaffordable. STRATA (stratified anatomy) partitions measurement, remedies, operating points, and certificates by *area*, so verdicts are taken as **min-over-strata** and mutations stale *bounded regions*, not the world.

The paper’s contributions are one framework and three measured campaigns run under it:

- (1) **The certification framework** (§3). Area maps are content-addressed (tqp-area-map/1: canonical JSON over the full configuration, SHA-256 digest), and an *incomplete* profile matches nothing, including itself. Per-stratum verdicts below minimum evidence are **ABSTAIN**—reported, counted, excluded from aggregation, and *not a pass*. Verdict causes come from a closed registry (an unregistered cause is a `KeyError` at emit time, not a new value quietly entering an artifact). The relational surface exposes hub census and transit fraction as SQL over a k NN edge table, and every relation carries its area-map digest: helpers *refuse*—not warn—to combine relations computed under different maps.
- (2) **The CSLS ladder** (§6). Four pre-declared gates measure where hubness correction actually works under $27.7\times$ compression. The exact layer works (count skew -20% , Robin Hood index -30%). Compressed-domain application is refuted. Reconstruction rerank is refuted with per-stratum numbers *identical to four decimals*, a coincidence that is itself the diagnosis: the bottleneck is candidate coverage, not scoring. A depth sweep then shows coverage saturating at 0.926 while fidelity plateaus at 0.702—past depth ~ 100

DRAFT v0 — 2026-07-24. Working draft; all measurements trace to committed artifacts in docs/results/strata-phase23-gates-2026-07-24/. Do not cite.

the ceiling is score precision—and reranking stored fp32 originals at depth 501 passes the 0.90 bar in every stratum (0.923). The design principle this ladder ends in: *the codes find, the originals decide*.

- (3) **Fragility allocation** (§7). Spending bits on measured fragility sounds obviously right and measurably is not: fragile-first greedy allocation upgraded all eight of its targets and still *lowered* the min-over-strata floor by cutting unprotected donors. A max-min allocator with protected donors raised the floor from 0.7251 to 0.7697 at an equal 3.0 bits/row budget—and revealed *cross-area score interference*: strata whose bits never changed improved because downgraded high-headroom areas stopped stealing candidate slots in the global top- k merge.
- (4) **The threat model** (§8). Per-area centering has an adversarial dual—a local centroid only needs to dominate a small neighborhood, which *lowers* the bar for centrality-injection attacks [2]—so per-area centroids are treated as attacker-relevant encoder state that enters the identity profile. Conversely, a per-area mechanism shift from density to centrality without a pipeline change is a poisoning signature: the quality instrument and the security scanner [1] are the same tool.

We are explicit about what is *not* claimed (§9). SQL over vectors is commodity [12, 15]; partitioned indexes with per-cell codebooks are a decade old [4, 9, 11]; hubness measurement and reduction have sixteen years of literature [5, 8, 16, 17] and a standard implementation [6]; and “hubness via SQL” is a five-line GROUP BY once the k NN graph is an edge table. The remaining gap this paper claims is the *composition*: **certified stratified measurement with refusal in the schema**—columns whose values know their estimator, their $n_{\min}$, and their area-map digest, with cross-map joins refused by the relation itself.

Every number in this paper traces to a committed JSON artifact or a CI test fixture in the open-source repository; the evaluation protocol (§4) pre-declares gates and typesets failures like passes. Three of the seven measured gates below exist *because* an earlier gate failed.

2 Background: hubness has a mechanism

For a corpus X , query multiset Q , metric d , and depth k , let $N_k(x)$ be the number of queries whose top- k list contains x . In low dimensions $N_k \approx k$ for everyone; in embedding dimensions the distribution grows a long right tail whose standard scalar summary is its skewness S_k [16]. Two intuitions: distances concentrate, so tiny systematic advantages decide top- k membership; and centrality is such an advantage—a point slightly closer to the data mean is slightly closer to everything on average. Local density is another.

The scalar does not identify the mechanism. The parent project’s anatomy instrument (tqp anatomy) documents two corpora with essentially the same skew whose hubs were completely different objects: one a smooth tail of moderately popular points in denser regions (max $N_{10} = 78$, hubs $\sim 8\%$ below population local scale), the other a handful of centrality super-hubs vacuuming up lists (max $N_{10} = 369$, hubs 34% closer to the mean). The two mechanisms

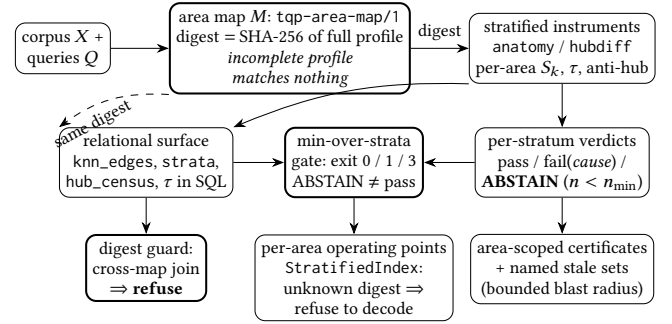


Figure 1: STRATA pipeline. A content-addressed area map stamps every downstream artifact with its digest. Stratified instruments emit per-stratum verdicts from a closed cause registry; thin strata ABSTAIN (reported, counted, excluded—not a pass) and gates aggregate as min-over-strata with exit codes 0/1/3. The same digest guards the relational surface (cross-map joins are refused by the relation, not discouraged by documentation), gates per-area index decoding, and scopes certificates to named stale sets.

take different remedies—centering / localized centering [8] for centrality, CSLS / mutual-proximity rescaling [5, 17] for density—so mechanism attribution is a prescription pad, and getting it wrong wastes the fix. Crucially, the naïve discriminator does not work: $\text{corr}(N_k, -d_k)$ is mechanically rank-coupled to the count in *any* corpus (it exceeds +0.8 on an isotropic Gaussian), which is one of the two classifier designs our confusion matrix killed (§5).

Hubness is a property of the experiment (X, Q, d, k) , never of the corpus alone: the same corpus can look mildly hubby under its own rows as queries and extremely hubby under a real workload. Every STRATA artifact therefore records its query battery, and corpus→corpus readings are not accepted as substitutes for workload readings.

3 The STRATA framework

STRATA’s normative language follows RFC 2119 [3]; its design vocabulary borrows deliberately from OSPF [14] (areas, a backbone, bounded flooding) while documenting exactly which parts of the analogy do *not* transfer (§3.6). The governing rules are inherited from the parent project’s identity discipline: uncertain $\Rightarrow$ no verdict; an incomplete profile matches nothing, including itself; unregistered causes raise; artifacts under different digests must never be compared, merged, or gated together—tools refuse, not warn. Figure 1 shows the pipeline.

3.1 Content-addressed area maps

An **area map** $M = (A, a, a_Q, O)$ consists of area identifiers $A = \{A_1..A_m\}$, a corpus assignment $a : X \rightarrow 2^A \setminus \emptyset$ (multi-assignment permitted), a query assignment a_Q (same rule, or a declared meta-data key such as `-by language`), and a reflexive adjacency/overlap relation $O \subseteq A \times A$ that drives staleness.

Identity is content-addressed. The profile `tqp-area-map/1` is canonical JSON (sorted keys, tight separators, ASCII, NaN forbidden) over {profile, algorithm_id, params, seed, corpus_

fingerprint, assignment_rule, query_assignment, overlap_policy, software_version}, hashed with SHA-256. Two rules give the digest its teeth:

- **Incomplete matches nothing.** Any None field makes the profile incomplete, and an incomplete profile is compatible with nothing—including itself. Unknown is contagious; there is no “probably the same map.”
- **Refuse, not warn.** Two artifacts computed under different digests MUST NOT be compared, merged, or gated together. The predicate is enforced in code (require_same_map raises the registered cause area_map_mismatch); a serialized map whose stored digest does not match its recomputed profile digest is refused at load as tampered.

Boundary rule. Nearest neighbors ignore borders. A conforming configuration must either multi-assign boundary rows or probe $\text{home} \cup O(\text{home})$ at query time; a single-area probe under a partition (non-cover) map is a non-conforming configuration over which certificates must not be issued. This is the part of the OSPF analogy that fails—administrative boundaries do not exist in embedding space—promoted from caveat to normative rule.

3.2 Stratified instruments and area classes

For each row, the instruments split its count into intra and transit components by whether the counting query’s area matches the row’s area, yielding the **transit fraction** $\tau(x) = N_k^{\text{trans}}(x)/\max(N_k(x), 1)$. Per area, the `tpq-strata-report/1` artifact carries count skew S_k , $\max N_k$, mean transit $\bar{\tau}$, the mechanism-attribution correlation fields restricted to the area, and the size-stable companions (Robin Hood index, $\text{frac}(N_k > 2k)$) that survive cross- n comparison. All thresholds are *fields of the report*, not constants; no number may be claimed in prose that is not present in the artifact.

Areas are classified (reported, never hidden) into **backbone** members (high τ and high N_k —transit carriers), **hub** areas, **stub** areas (low counts in and out), and the **not-so-hubby area** (NSHA): high intra-area skew with low transit—local hubs, no global reach. The Phase-1 measured run on a 37-language corpus falsified both of its on-record predictions in the informative direction: per-language S_k varies far less than predicted under BGE-M3 [19] (the encoder homogenizes), and the trained interlingua LaBSE [7] does not concentrate transit on a central core—it *diffuses* it, turning 7 of 13 eligible languages into the first measured backbone-class areas, while emergent BGE-M3 keeps language borders (all-NSHA). The taxonomy’s backbone class, speculative at design time, exists in production encoders and is objective-dependent.

3.3 Verdict semantics: min-over-strata and ABSTAIN

A per-stratum verdict REQUIRES $n_i \geq n_{\min}$ corpus rows and $|Q_i| \geq q_{\min}$ queries (defaults in every artifact below: $n_{\min} = 2000$, $q_{\min} = 500$). Below either bound the verdict is **ABSTAIN**: reported, counted under the registered cause `stratum_insufficient_n`, and excluded from pass/fail aggregation. ABSTAIN is not a pass—skewness on thin strata is noise, and uncertain means no verdict—but it is not a silent fail either: the report counts abstentions separately, and CI can opt in to `-abstain-fails`.

Gates aggregate as **min over eligible (non-ABSTAIN) strata**, with exit codes 0 (pass), 1 (a stratum failed), and 3 (only-ABSTAIN).

Causes and classes are a **closed registry** (`stratum_insufficient_n`, `stratum_anti_hub_gap`, `transit_concentration`, `area_map_mismatch`; classes `backbone`, `hub`, `NSHA`, `stub`, `plain`). Emitting an unregistered cause raises `KeyError`. Schema fields, cause IDs, and class names are API: additions only, with golden report fixtures frozen in CI.

3.4 The relational surface: certified hubness as SQL

Once the k NN graph is an edge table `knn_edges(query_id, neighbor_id, rank)`, the hub census is in-degree—a GROUP BY—and the transit fraction is a join against the strata relation. That much is commodity, and we say so. The house part is that *every* relation the surface registers (`knn_edges`, `strata`, `query_strata`, `strata_report`) carries an `area_map_digest` column, and every helper runs a digest guard first: if the relations it is about to join carry more than one distinct digest, it raises `area_map_mismatch` and returns nothing. The replication predicate is enforced *in the schema*, not in the documentation. The gate itself is a query: `strata_gate` returns the failing rows (ABSTAIN excluded, counted separately), so “rows returned $\Rightarrow$ exit 1” composes with any SQL tooling the operator already has [15]. The scan side streams compressed blocks to Arrow batches, so census queries run without decompressing the corpus into host RAM.

3.5 Per-area operating points and bounded staleness

Per-area bits, PCA dimension, and codebooks acknowledge their LOPQ ancestry [11]; the new element is the allocator input—measured per-stratum fragility (anti-hub recall), not variance (§7)—and the identity gating. Record metadata gains {`area_map_digest`, `area_id`, `area_codec_params`}; a reader encountering an unknown digest MUST refuse to decode (enumerate, don’t decode). Cross-map reads are cache misses, never best-effort; `StratifiedIndex.search` requires the caller’s digest and its refusal path is regression-tested. Areas thinner than the requested PCA dimension clamp their per-area basis and declare it in `area_codec_params`—a legitimate per-area operating point, not a silent fallback.

Guarantees then scope naturally: the recall-contract catalog key extends with (`area_map_digest`, `area_id`), and every mutation class has a documented worst-case stale set—a mutation in A_i stales $\text{cert}(A_i)$ plus overlap neighbors; an operating-point change stales exactly $\text{cert}(A_i)$; a map recompute stales everything. A guarantee that can go stale silently is not a guarantee; STRATA’s version has a bounded, *named* blast radius, and recalibration becomes incremental: the flapping area re-runs its measurement, the backbone does not.

3.6 What the OSPF analogy does not license

Hierarchy, border summarization, and bounded blast radius transfer. Administrative boundaries do not (geometry ignores them; hence the mandatory overlap rule), and neither does link-state truth: strata are workload-dependent (a_Q matters, and corpus $\rightarrow$ corpus is not

a substitute for the deployed query distribution). These limits are normative for prose in the project’s documentation, along with a banned list (“eliminates hubness,” “uniform recall guaranteed”) and an approved claim shape: “ Δ measured in this artifact on this corpus.” This paper is written under those rules.

4 Evaluation protocol

STRATA’s phases each carry pre-declared gates; the evaluation below is those gates, run and reported whether they passed or failed. Three of the gates exist only because an earlier gate failed, and the failures carry most of the information.

Corpora. (1) The *Ethics 350k* sample ($n = 350,000$, seed 7; deployed operating point PCA-384 + 3-bit quantization, 27.676× compression, single-pass ADC search; 14 language strata of which 7 are eligible at $n_{\min} = 2000$, $q_{\min} = 500$). (2) The public paired *Gutenberg* rung, BGE-M3 arm ($n = 150,067$ rows, 20 language strata, 14 eligible). Ground truth throughout is exact k NN on the given vectors at $k = 10$; the fidelity instrument is the stratified differential oracle (hubdiff) with anti-hub recall gated at 0.90 over bottom-decile rows of the *global* count distribution (a stratum cannot vote itself out of having anti-hubs). Every report artifact carries its area-map digest, estimator, thresholds, and software version; the artifacts are committed in the repository at docs/results/strata-phase23-gates-2026-07-24/.

5 The mechanism classifier and the confusion matrix that killed two designs

Mechanism attribution (centrality vs. density) selects the remedy, so a misclassification wastes the fix. The classifier is validated by a confusion matrix of four constructed fixture families, one per cell of the prescription pad, each run at three seeds (Table 1). Two earlier designs died against these fixtures before the current one survived:

- (1) **Corpus-wide correlation thresholds.** Classifying on the correlation $\text{corr}(N_k, -d_k)$ fails because it is coupled mechanically to the count in any corpus—it exceeds +0.8 even on an isotropic Gaussian—so it cannot discriminate mechanisms at all.
- (2) **Symmetric percentile thresholds.** Typing each hub by independent percentile cuts on centrality and density fails on the dominance structure: a hub in the deep centrality tail has a small d_k because it is central (apparent density is implied by centrality), so density-typing must be conditioned on *not* being central.

The surviving design types each hub hierarchically by its population percentiles—deep-centrality-tail hubs are central regardless of their density; dense-but-not-central is the healthy real-corpus pattern—and abstains entirely when there is no material hub tail ($\max N_k < 2.5k$): noise hubs get no prescription. One scope limit is declared rather than patched: at global scope, *locally* central hubs read as dense (a point too close to everything in its own region has a small d_k globally), and CSLS is the correct global remedy for both. Splitting them apart is precisely per-area work—the fixture that motivates STRATA.

The differential oracle is validated the same way: a fixture corrupts only queries whose true nearest neighbor is an anti-hub and CI requires aggregate recall to stay above 0.75 while anti-hub recall

Table 1: The classifier’s confusion matrix: four fixture families $\times$ three seeds, all cells required to hold in CI (tests/test_anatomy.py). These fixtures killed the two designs described in §5.

fixture construction	required attribution	required evidence
unit shell + points planted at the center	centrality; prescription: centering	hub_frac_central = 1.0
off-center shells, each with planted local centers	density; prescription: CSLS	dense-noncentral frac ≥ 0.75 ; central frac = 0
density corpus with origin shell also center-planted	mixed	both fractions ≥ 0.25
constant-density (negligible hubness)	unclear: attribution <i>abstains</i>	$\max N_k < 2.5k$; “no material hub tail”

collapses below 0.30—the exact failure shape aggregate metrics cannot see, held as a permanent regression test.

6 The CSLS ladder: where correction actually works

CSLS [5] (and mutual proximity [17]) corrects density-mechanism hubness with one precomputed scalar per record: $r_k(x)$, the mean similarity of x to its k nearest neighbors, subtracted from the raw score at ranking time. One float32 per record makes it a natural optional trailer in the storage format (registered ID 0x10, hubness-scalar/1, carrying its own hash; unknown-optional $\Rightarrow$ skip, so old readers remain conformant). The question the ladder answers: *at which layer of a compressed search stack does the correction survive?* All rungs run on *Ethics 350k* at the deployed 27.676× operating point, gated at min-over-strata anti-hub recall ≥ 0.90 .

6.1 Gate A: failed by design error

The first gate scored CSLS-rescored ADC rankings against *uncorrected* exact-cosine ground truth. Anti-hub recall “dropped” by roughly 0.32 in every stratum (min-over-strata 0.437 at $w = 1.0$, 0.571 at $w = 0.5$). The diagnosis: the gate conflated two questions. CSLS is a *different ranking objective*; scoring it against cosine truth measures disagreement, not damage. The gate design, not the remedy, failed first—kept on the scoreboard as the reason Gate A’ exists.

6.2 Gate A’: the two questions separated

A’1 — does CSLS reduce hubness on the exact graph? Yes. Table 2 shows the exact-layer effect at $w = 1.0$: every hub-tail statistic improves. The r_k scalar earns its trailer—measured hub-tail reduction on this corpus at every statistic.

A’2 — does compressed CSLS reproduce exact CSLS? No. Rescoring ADC scores with r_k and judging against exact-CSLS truth: min-over-strata anti-hub recall 0.663 against the 0.90 bar; all 7 eligible strata fail; per-stratum recalls 0.62–0.69—substantially *worse* fidelity than the uncorrected path (0.76–0.84). The reading: CSLS promotes exactly the isolated, fine-distinction rows that 27.7× quantization damages first, so the corrected ranking is *harder* to

Table 2: Gate A'1: exact-layer CSLS vs. cosine, Ethics 350k, $k = 10$, $w = 1.0$ (artifact phase2_efficacy_v2.json).

statistic	cosine graph	CSLS graph	Δ
count skew	3.970	3.177	−20%
count max	287	213	−26%
Robin Hood index	0.372	0.261	−30%
frac($N_k > 2k$)	0.117	0.079	−33%

reproduce from compressed codes than the uncorrected one. Consequence (normative): the hubness-scalar/1 trailer semantics stay provisional, and no unconditional improvement claim is made.

6.3 Gate A'': rerank refuted—identically

The natural rescue is two-stage: fetch ADC candidates (depth 51, the deployed single pass), rerank by cosine on *reconstructed* candidates minus r_k , judge against exact-CSLS truth. Measured: min-over-strata anti-hub recall **0.6627—identical to A'2 to four decimals**, with per-stratum values matching within ± 0.0001 .

The identity is the finding. Moving CSLS from compressed-domain scoring to reconstruction rerank changed *nothing*, so the bottleneck is not the scoring layer—it is **candidate coverage**. Exact-CSLS's true neighbors (isolated, low- r_k rows) frequently do not appear in the plain ADC top-51 at all, and no rerank scheme can recover a candidate that was never fetched. (Reconstruction carries the same information as the codes, so this rung also pre-answers "rerank harder": there is nothing more in the codes to extract.)

6.4 Gate A''': coverage saturates, fidelity does not follow

One ADC sweep at depth 501, every smaller depth a prefix, CSLS rescore at each depth (Figure 2, artifact phase2_depth_curve.json): coverage of the CSLS truth climbs $0.696 \rightarrow 0.788 \rightarrow 0.861 \rightarrow 0.926$ (23 points), while fidelity gains $+0.033$, then $+0.004$, then $+0.002$ —a plateau at 0.702 against coverage 0.926. The diagnosis is therefore *layered*: at depth 51 the ceiling is coverage (A'''s finding); past depth ~ 100 the ceiling becomes **score precision**—the true neighbor is *in* the candidate list and the $27.7\times$ scores are too coarse to place it. Deeper candidates alone are not a remedy: $10\times$ the scan depth buys $+0.039$ fidelity.

6.5 Gate A''': the originals-rerank storage trade passes

The one remaining lever is reranking against stored fp32 originals (4096 B/row on top of the 148 B/row compressed record)—the true CSLS objective evaluated on the candidate list. The pre-declared prediction confirmed exactly: fidelity equals coverage at every depth (0.702/0.792/0.861/0.923 at depths 51/101/201/501 vs. coverage 0.696/0.788/0.861/0.926; artifact phase2_originals_trade.json), and at depth 501 zero strata fail the 0.90 bar.

6.6 The complete menu, and the principle

All three deployment options are now measured, none extrapolated: (1) **exact-layer CSLS**—free, works (A'1); (2) **compressed-path**

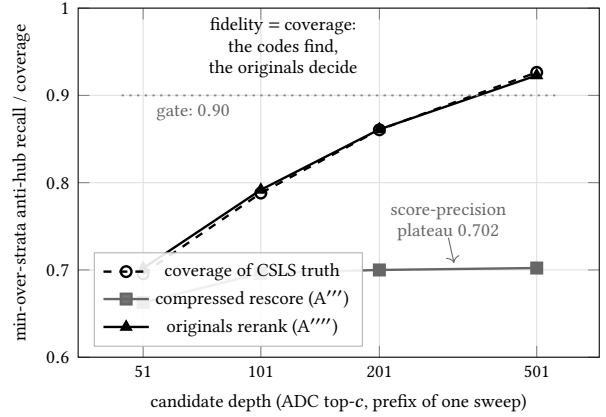


Figure 2: The CSLS ladder's depth sweep (Ethics 350k, $27.676\times$, $w = 1.0$; artifacts phase2_depth_curve.json, phase2_originals_trade.json). Candidate coverage of the exact-CSLS truth saturates (0.926 at depth 501) while compressed-domain rescoring plateaus at 0.702—past depth ~ 100 the binding constraint is score precision, not coverage. Reranking stored fp32 originals tracks coverage at every depth and passes the 0.90 min-over-strata gate at depth 501 (0.923, zero failing strata).

CSLS—refuted three ways (A'2, A'', A''': first coverage-limited, then score-precision-limited); (3) **paid CSLS**—deployed codes + stored originals + depth-501 rerank clears the bar in every stratum. This is the project's tiered rerank architecture doing for the *corrected* ranking what it previously did for the plain one at the billion-row fleet run: **the codes find, the originals decide**. Compressed codes are a candidate generator; verdicts belong to a representation that still contains the fine distinctions the verdict is about.

7 Fragility allocation: raising the floor without digging under it

Per-area operating points make capacity a policy variable: which strata get 4 bits and which get 2, at a fixed corpus-wide budget? STRATA's allocator input is measured fragility—per-stratum anti-hub recall from the differential oracle—rather than variance. Both gates run on the public Gutenberg rung (BGE-M3 arm, 150,067 rows, 20 language strata, 14 eligible) at an equal budget of 3.0 bits/row against the uniform 3-bit baseline, whose min-over-strata anti-hub recall is 0.7251 (spanish).

7.1 Gate B: fragile-first greedy fails, instructively

The obvious allocator upgrades the most fragile strata first while the row-weighted mean stays within budget. It did exactly what it was told: all eight of its 4-bit targets improved ($+0.004$ to $+0.034$). It also cut its unprotected donors below the old floor—english $0.7546 \rightarrow 0.6856$, greek -0.062 , latin -0.053 —dragging min-over-strata down to 0.6856, *below* the uniform baseline it was meant to improve. It flattened the distribution and lowered the minimum. The lesson,

stated as the gate’s epitaph: *min-over-strata is not improved by spending on the weak; it is improved by not digging under the floor.*

7.2 Gate B2: max-min with protected donors passes

The redesigned allocator optimizes the floor directly, under declared rules: recipients are the fragile half (below the median baseline recall), upgraded most-fragile-first; donors need headroom ≥ 0.06 over the baseline *minimum* (the one-bit penalty Gate B measured) and give at most one step; unfundable upgrades revert their donations; ABSTAIN strata hold the uniform width. At the same budget (achieved 2.995 bits/row):

- **min-over-strata anti-hub recall:** 0.7251 $\rightarrow$ 0.7697 (+0.045), no stratum below the old minimum—the pre-declared PASS condition (artifact phase3_trade_v2.json).
- Every recipient-half stratum improved (spanish +0.047, swedish +0.054, italian—the one funded 4-bit upgrade—+0.050); protected latin *improved* (+0.035) instead of being cut; donors paid at most -0.014 (greek); two donors (chinese, esperanto) improved despite their downgrades.

7.3 Cross-area score interference

The unplanned finding is visible in Figure 3: strata whose bit width *never changed* improved by +0.03–0.05. The mechanism is real and the uniform baseline could not have shown it: the global top- k merges ADC scores *across* areas, so a high-headroom area’s coarsely-quantized rows compete for candidate slots against every other area’s true neighbors. Downgrading the high-headroom donors reduced their rows’ ability to steal those slots. **Allocation is not per-area-local:** a bit spent (or saved) in one stratum moves retrieval quality in strata it never touched.

7.4 Limits and allocator defects, on the record

A structural limitation, stated plainly: with this corpus’s donor capacity ($\sim 8k$ eligible donor rows), the largest fragile strata (spanish 10k; finnish, french, german 15k each) were individually unfundable; the allocator funded what it could and the floor still rose. On donor-scarce corpora the safe answer converges toward uniform—that convergence is a feature, not a failure mode.

The gates also found (and CI now pins) three allocator defects: fragile-first greedy lowers the floor (Gate B, replaced by max-min); the first max-min implementation leaked donations when a recipient was unfundable—donors cut, nobody upgraded, strictly worse than uniform—and gave up rather than trying cheaper recipients (donations now revert and iteration continues); and the stratified index crashed on areas thinner than the PCA dimension (per-area dimension now clamps and declares itself in `area_codec_params`).

8 Threat model: measurement as intrusion detection

Two 2026 results moved hubness from a retrieval-quality concern to a security surface, and both validate the mechanism taxonomy of §2. The Black-Hole Attack [2] injects a handful of vectors near the embedding-space centroid—exploiting centrality-mechanism hubness on real BGE embeddings—and hijacks the vast majority of

top-10 results; cluster-wise centroids *amplify* the attack, because a local centroid only needs to dominate a small neighborhood. The Adversarial Hubness Detector [1] ships an open-source hubness-poisoning scanner across FAISS/Pinecone/Qdrant/Weaviate using MAD-based count z-scores and cross-cluster retrieval-spread analysis—the latter a convergent cousin of STRATA’s transit fraction τ .

STRATA draws two normative consequences.

Per-area centering has an adversarial dual. Localized centering [8] at region scale is the natural remedy for centrality-mechanism areas—and it *lowers the bar* for centrality-injection attacks, for exactly the reason it works. The framework therefore treats stored per-area centroids as attacker-relevant encoder state: they enter the identity profile (changing them is an epoch event that moves the digest, never a silent tweak), and Phase-2 remedies are never deployed as write-path trust.

Mechanism shift is a poisoning signature. A planted super-hub is a centrality hub, so the anatomy report doubles as a monitoring instrument: a sudden per-area mechanism shift from density to centrality, or a jump in `hub_frac_central`, without a pipeline change is an investigation, not a re-quantization. The monitoring hook is a diff of the mechanism and `hub_frac_central` fields per area across index mutations—which the digest discipline makes meaningful, since both sides of the diff are guaranteed to be measurements of the same experiment. The quality instrument and the security scanner are the same tool; the exact-layer r_k scalar’s measured value today (§6) is A’1’s hub-tail reduction *plus* this monitoring signal, not compressed-path retrieval correction.

9 Related work and honest positioning

Every component is commodity, and the decomposition is stated before a reviewer states it for us. SQL over vector search is shipped by pgvector [12], DuckDB-based stacks [15], and Vector-SQL systems [18]. Partitioned indexes are IVF [9, 10]; per-cell codebooks are LOPQ [11]; boundary duplication is SPANN [4]; emergent hub backbones are HNSW’s upper layers [13]. Hubness emergence is Radovanović et al. [16]; the reduction canon is mutual proximity [17], localized centering [8], and CSLS [5], with scikit-hubness [6] as the standard implementation of measurement and reduction. “Hubness via SQL” is a five-line GROUP BY once the k NN graph is an edge table. The security community is converging from the adversarial side [1, 2], shipping scanners.

What none of these ship, and what this paper claims, is the *composition: a relational surface over certified stratified hubness*—areas that are hubness-diagnosed, fragility-allocated, identity-versioned, and carry per-region guarantees with bounded, named staleness; columns whose values know their estimator, their minimum-evidence bounds, and their area-map digest; and cross-map joins refused by the relation itself rather than discouraged by documentation. The detectors ship scanners; the databases ship `DISTANCE()`; the composition is the claim. No “first” claim is made beyond it, per the project’s positioning rules; a completed literature-check notebook is a precondition for any stronger wording.

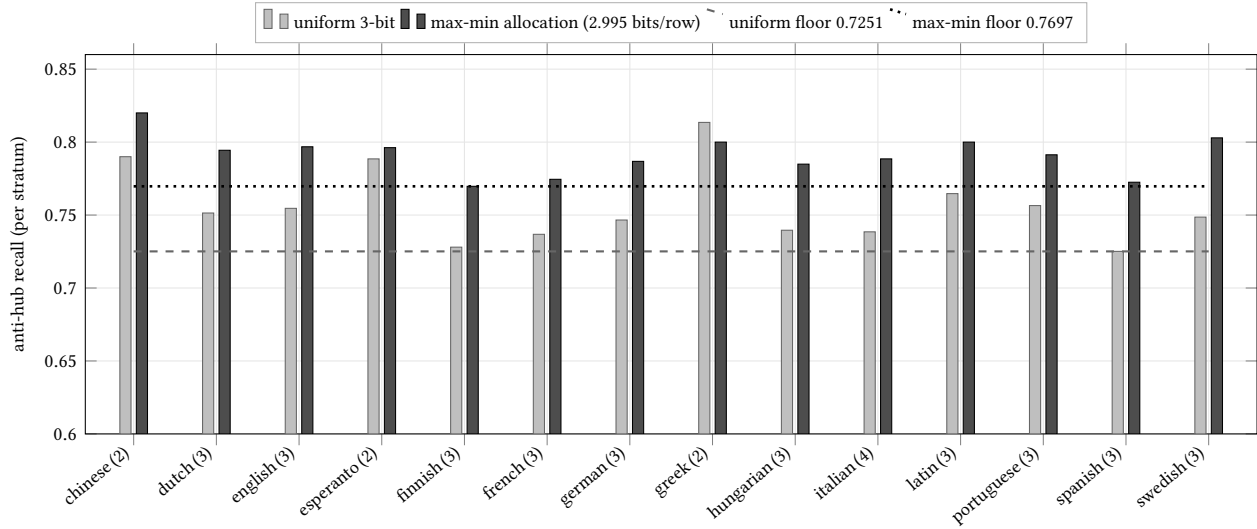


Figure 3: Gate B2: per-stratum anti-hub recall under the uniform 3-bit baseline vs. the max-min allocation with protected donors, Gutenberg rung (BGE-M3 arm, 150,067 rows; 14 eligible strata of 20; artifact phase3_trade_v2. json). Parenthesized tick labels give the allocated bits per stratum (italian is the one funded 4-bit upgrade; chinese, esperanto, and greek are 2-bit donors). The floor rises $0.7251 \rightarrow 0.7697$ at an equal 3.0 bits/row budget with no stratum below the old minimum; donors pay at most -0.014 (greek), and unchanged 3-bit strata improve $+0.03$ – 0.05 —the cross-area score interference of \$7.3.

10 Limitations

The measured campaigns cover two corpora (350k and 150k rows) and one compressed operating-point family; the approved claim shape—“ Δ measured in this artifact on this corpus”—is the only claim shape used, and no cross-corpus generalization is asserted. The originals-rerank pass is a storage trade (4096 B/row of fp32 originals against 148 B/row codes) whose latency cost at deployment scale is characterized in the companion systems work, not here. The Gate-B2 allocator’s response model is one measured bit-step; two-step penalties are unmeasured, and the allocator refuses to extrapolate them by construction. Overlap policy (multi-assignment vs. adjacent-probe) is a recorded open decision: Phase-1/2/3 artifacts here use single-assignment maps, which is a non-conforming configuration for Phase-4 certificates and is labeled as such in every artifact. Certificate catalog wiring beyond the normative core (contract keys, stale sets—implemented and tested) remains future work.

11 Conclusion

STRATA turns hubness measurement from a scalar into a certified, spatial instrument: content-addressed area maps whose incomplete profiles match nothing, min-over-strata verdicts whose thin strata ABSTAIN rather than pass, closed cause registries, and a relational surface that refuses cross-map joins in the schema. Run against its own pre-declared gates, the framework produced findings its unstratified predecessors could not have seen: a confusion matrix that killed two classifier designs; a four-decimal coincidence that located a bottleneck (coverage, then score precision) and ended in a deployable principle—the codes find, the originals decide; an “obviously correct” allocator measurably digging under the floor

it was meant to raise, and its max-min replacement raising that floor $0.7251 \rightarrow 0.7697$ at equal budget while exposing cross-area score interference; and a threat model in which the measurement instrument doubles as the intrusion detector. Trust the tail, not the mean—in every area.

Acknowledgments

DRAFT v0. Artifacts, code, and test fixtures: github.com/ahb-sjsu/turboquant-pro (MIT). All measurements in this paper are committed under `docs/results/strata-phase23-gates-2026-07-24/` with their area-map digests; the framework and gate definitions are `docs/STRATA_RFC.md` and `docs/RESULTS_strata_phase23_gates.md`.

References

- [1] [authors omitted in draft]. 2026. Adversarial Hubness Detector. arXiv:2602.22427. Open-source hubness-poisoning scanner across FAISS/Pinecone/Qdrant/ Weaviate; MAD-based count z-scores and cross-cluster retrieval-spread analysis. TODO-verify author list and exact title before submission.
- [2] [authors omitted in draft]. 2026. Can You Trust the Vectors in Your Vector Database? arXiv:2604.05480. The Black-Hole Attack: centroid-proximate vector injection exploiting centrality-mechanism hubness. TODO-verify author list before submission.
- [3] Scott Bradner. 1997. Key Words for Use in RFCs to Indicate Requirement Levels. IETF RFC 2119.
- [4] Qi Chen, Bing Zhao, Haidong Wang, Mingqin Li, Chuanjie Liu, Zengzhong Li, Mao Yang, and Jingdong Wang. 2021. SPANN: Highly-Efficient Billion-Scale Approximate Nearest Neighbor Search. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [5] Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *International Conference on Learning Representations (ICLR)*.
- [6] Roman Feldbauer and Arthur Flexer. 2020. scikit-hubness: Hubness Reduction and Approximate Neighbor Search. *Journal of Open Source Software* 5, 45 (2020). TODO-verify full author list and page number before submission.

- [7] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-Agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- [8] Kazuo Hara, Ikumi Suzuki, Masashi Shimbo, Kei Kobayashi, Kenji Fukumizu, and Miloš Radovanović. 2015. Localized Centering: Reducing Hubness in Large-Sample Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- [9] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. 2011. Product Quantization for Nearest Neighbor Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 1 (2011), 117–128.
- [10] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2021), 535–547.
- [11] Yannis Kalantidis and Yannis Avrithis. 2014. Locally Optimized Product Quantization for Approximate Nearest Neighbor Search. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2321–2328.
- [12] Andrew Katz. 2023. pgvector: Open-Source Vector Similarity Search for Postgres. <https://github.com/pgvector/pgvector>.
- [13] Yury A. Malkov and Dmitry A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 4 (2020), 824–836.
- [14] John Moy. 1998. OSPF Version 2. IETF RFC 2328.
- [15] Mark Raasveldt and Hannes Mühleisen. 2019. DuckDB: An Embeddable Analytical Database. In *Proceedings of the 2019 International Conference on Management of Data (SIGMOD)*. 1981–1984.
- [16] Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Journal of Machine Learning Research* 11 (2010), 2487–2531.
- [17] Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer. 2012. Local and Global Scaling Reduce Hubs in Space. *Journal of Machine Learning Research* 13 (2012), 2871–2902.
- [18] TODO. 0. TODO: MyScale / Vector SQL reference. SQL-over-vectors commodity claim; find the citable artifact.
- [19] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Liang, and Dian Fang. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*. 2548–2569.